Title

Constrained Open Chromatin Regions Reveal Functional Elements Shaping Human Traits

Kohei Tomizuka[1]†, Masaru Koido[1,2]†,*, Akari Suzuki[3]†, Soichiro Yoshino[1], Nao Tanaka[1], Yuki Ishikawa[1], Xiaoxi Liu[1], Satoshi Koyama[4], Kazuyoshi Ishigaki[5], Yasuhiro Murakawa[6], Immune transcript/enhancer consortium (ITEC)[#], Kazuhiko Yamamoto[3,7], Chikashi Terao[1,8,9,*]

Affiliations

[1]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.

[2]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 4-6-1, Shirokane-dai, Minato-ku, Tokyo, 108-8639, Japan.

[3]Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.

[4]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.

[5]Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.

[6]RIKEN-IFOM Joint Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.

[7]Drug Discovery Antibody Platform Unit, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.

[8]Clinical Research Center, Shizuoka General Hospital, 4 Chome-27-1 Kitaando, Aoi Ward, Shizuoka, 420-8527, Japan.

[9]Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, 52-1 Yada, Suruga Ward, Shizuoka, 422-8526, Japan.

[#]Immune transcript/enhancer consortium (ITEC) members

Matteo Guerrini, Hiroaki Hatano, Keiko Hikino, Kazuyoshi Ishigaki, Yuki Ishikawa, Michihiro Kono, Xiaoxi Liu, Yasuhiro Murakawa, Masahiro Nakano, Akiko Oguchi, Raku Son, Akari Suzuki, Kazuhiro Takeuchi, Nao Tanaka, Chikashi Terao, Kohei Tomizuka, Kazuhiko Yamamoto, Soichiro Yoshino

†These authors contributed equally: Kohei Tomizuka, Masaru Koido, Akari Suzuki

*Corresponding authors:

Chikashi Terao (chikashi.terao@riken.jp)

Masaru Koido (masaru.koido@a.riken.jp)

Abstract

Open chromatin regions (OCRs) define cell-type-specific regulatory elements across the genome, yet their functional significance varies, making it challenging to pinpoint biologically essential regions. Here, we introduce CAMBUS (Chromatin Accessibility Mutation Burden Score), a machine-learning framework that identifies active and evolutionarily constrained OCRs by leveraging surrounding DNA sequences. Applying CAMBUS to 29 immune cell types, we identified 66,043 constrained OCRs, which were highly enriched for known enhancers and super-enhancers, independent of known epigenetic markers and annotated

50 regions. These OCRs overlapped with regulatory elements implicated in immune-mediated
51 diseases and experimentally validated functional variants, including rare variants, particularly
52 in leukocyte-related traits. Furthermore, CAMBUS revealed distinct transcriptional regulatory
53 landscapes, linking genetic constraint with gene regulation in immune cells and identifying
54 plausible connections between 1,533 causal variants and 70 complex traits. By defining
55 biologically constrained regulatory elements at high resolution, CAMBUS provides a
56 framework for understanding the selective pressures shaping the non-coding genome and
57 its role in human health and disease.

58

59 **Introduction**

60 Genome sequences define species identities and evolutionary trajectories by
61 shaping cellular, organ, and functional complexities [1,2]. Given that protein-coding regions
62 account for only a small fraction of vertebrate genomes [3], understanding the role of non-
63 coding regions—particularly active regulatory elements—is critical for deciphering the
64 molecular basis of species identity and function [4,5]. Active regulatory regions, such as
65 enhancers, regulate gene expression in a cell-type- and state-specific manner [6-9].
66 Several genome-wide experimental methods, including chromatin
67 immunoprecipitation followed by sequencing (ChIP-seq) for histone modifications, DNase I
68 hypersensitive site sequencing (DNase-seq), and assay for transposase-accessible
69 chromatin using sequencing (ATAC-seq), have been developed to identify candidate
70 regulatory regions. Among them, ATAC-seq has become widely adopted due to its technical
71 simplicity and robust performance [10,11]. Compared with ChIP-seq peaks, ATAC-seq-defined
72 open chromatin regions (OCRs) provide high positional specificity for identifying regulatory
73 elements [12]. Because chromatin accessibility is a prerequisite for gene transcription, mapping

74 and characterizing OCRs—particularly those that are cell-type-specific or conserved across
75 multiple lineages—offers crucial insights into general and context-specific regulatory
76 mechanisms, evolutionary constraint, and complex traits [13,14].

77 Although genome-wide association studies (GWAS) have highlighted the
78 enrichment of disease-associated variants within non-coding regions [15,16], and particularly
79 within regulatory elements [7,17,18], distinguishing active regulatory elements from OCRs
80 remains challenging [15,19,20], as ATAC-seq measures chromatin openness but does not
81 directly assess regulatory activity [21,22]. Therefore, integrating additional layers of information,
82 such as evolutionary constraint and epigenomic signals, is necessary to identify active and
83 functional regulatory regions more precisely.

84 Constrained genomic regions, which show evidence of purifying selection, serve as
85 important markers of functionally active elements [5,23]. Recent large-scale human whole-
86 genome sequencing efforts have enabled the development of constraint metrics such as
87 Gnocchi [24], which leverage local mutability to infer constrained sites. Consortia efforts such
88 as ENCODE Phase 3 have combined chromatin accessibility with histone modification ChIP-
89 seq signals to annotate candidate cis-regulatory elements (cCREs). Still, the systematic
90 identification of constrained, cell-type-specific active regulatory regions across diverse
91 tissues remains an open challenge.

92 Chromatin accessibility quantitative trait loci (caQTLs), which identify genetic
93 variants that alter chromatin openness, provide further evidence that genetic variation at
94 regulatory regions influences complex traits [25,26]. However, caQTL mapping requires large
95 cohorts and is limited in its ability to assess rare variants. Recently, machine learning models
96 trained on DNA sequences have emerged as a powerful alternative, capable of predicting
97 regulatory activity and inferring variant effects without requiring population-level data [27-29].
98 These models have enabled the tissue- and cell-type-specific prediction of gene expression
99 and chromatin features from sequence alone [29,30].

100 Together, these observations highlight the need for methods that integrate
101 evolutionary constraint, chromatin accessibility, and DNA sequence features to prioritize
102 functional regulatory elements. Yet, the constraint landscape of cell-type-specific OCRs
103 remains poorly understood, limiting our ability to interpret their functional relevance.
104 Motivated by this gap, we developed CAMBUS (Chromatin Accessibility Mutation Burden
105 Score), a new framework for prioritizing active and evolutionarily constrained OCRs. By
106 leveraging sequence-embedded constraint information (Fig. 1), CAMBUS provides a
107 principled approach to identify regulatory elements with likely functional importance across
108 diverse cell types. We assessed the characteristics of CAMBUS and their cell-type specificity
109 using OCRs in various cell types in peripheral blood mononuclear cells (PBMCs). By
110 leveraging CAMBUS, we identified 1,533 cell-type-specific regulatory variants with potential
111 functional significance in complex traits across European (EUR) and East Asian (EAS)
112 ancestries (Fig. 1D and 1E). We demonstrated how integrating genetic constraint with
113 chromatin accessibility can refine the biological interpretation of non-coding regulatory
114 variants and their roles in transcriptional regulation in a cell-type-specific manner.
115

116 ## Results

117 ### Cell-type-specific OCR predictions from DNA sequences

118 We sorted PBMCs (peripheral blood mononuclear cells) and Neutrophils from 55 Japanese
119 individuals (Methods). PBMCs were sorted into 27 immune cell types by cell sorter and
120 conducted ATAC-seq for the 27 cell types and PBMC and Neutrophil (hereafter, 29 cell types)
121 in eight subgroups (CD4; $CD4^+$ T-cells, CD8; $CD8^+$ T-cells, B; B cells, Mono; Monocytes, DC;
122 dendritic cells, NK; Natural Killer cell, Neutrophils, and PBMC) (Supplementary Table 1). The
123 number of cell types in this study is comparable to that in recent scATAC-seq studies using
124 PBMC [31,32]. Notably, the 29 cell types include minor cell types, such as double-negative B-
125 cells (~5% in B-cells) [33]. After stringent quality control, we obtained 7,939,714 OCRs that

126 covered 498 Mb (16.1 %) of the human genome (Methods). The median number and size of
127 OCRs for 29 cell types were 252,668 OCRs and 276 bp, respectively. Among the 7,939,714
128 OCRs, 330,367 (4.2%) were observed in only one cell type, and 437,368 (5.5%) were
129 observed in only one out of the eight subgroups. Both hierarchical clustering and dimension
130 reduction using principal component analysis (PCA) and uniform manifold approximation and
131 projection (UMAP) showed that these OCRs could clearly distinguish the eight subgroups
132 but not all 29 immune cell types (Extended Data Fig. 1).

133 We first examined whether machine learning models (CAMBUS ML (Chromatin
134 Accessibility Mutation Burden Score machine learning) models) can predict the OCRs of
135 each of 29 cell types from DNA sequence patterns. For this purpose, we used the deep
136 convolutional neural network to predict whether OCRs overlap within the central 200bp of
137 5kb input DNA sequences (Methods). We observed a notably high mean area under the
138 receiver operating characteristic curve (AUROC) of 0.87, with a minimum of 0.84
139 (neutrophils), to predict OCRs across 29 cell types in the testing dataset (Fig. 2A). We
140 observed that cell-type-specific OCRs were more challenging to predict; however, even for
141 the OCRs observed in only one cell type, the mean AUROC was 0.74 (Fig. 2B), suggesting
142 CAMBUS ML models could effectively capture cell-type-specific genetic architecture
143 underlying OCRs. Notably, this level of prediction accuracy has been enough for successful
144 downstream analyses (such as *in silico* mutagenesis for GWAS variants) in the previous
145 studies: for example, AUROC per cell = 0.762 [34] and mean AUROC for enhancer RNAs =
146 0.69 [29].
147 We also assessed the accuracy of *in silico* mutagenesis on the cell-type-specific OCRs in
148 the 29 cell types. The direction concordance between the *in silico* mutation effects and the
149 estimated effect sizes from caQTL was increased by a higher threshold of the mutation
150 effects, as shown in previous studies [27,29]. Particularly when variants were selected based

151 on a threshold where accessibility probability changed by more than 10%, indicating high
152 confidence of *in silico* mutagenesis, the direction concordance with caQTL effect sizes
153 (Methods) reached 90% for the major six cell types (CD4, CD8, B, Mono, DC, and NK) (Fig.
154 2C; Supplementary Table 2). We also confirmed the good concordance using the
155 independent caQTL study of lymphoblastoid cell lines (LCL; B cell lineage; Fig. 2D; Extended
156 Data Fig. 2A) [25]. As expected, the concordance dropped when we compared *in silico*
157 mutation effects in non-B cell types with the caQTL study (Extended Data Fig. 2B). Thus,
158 CAMBUS ML models incorporating surrounding DNA sequence patterns could accurately
159 predict OCRs within the 29 cell types, including cell-type specific ones, and mutation effects
160 on these OCRs in a cell-type specific manner.

161

162 **Detecting constrained OCRs by CAMBUS**

163 To infer constraint levels in the OCRs, we calculated *in silico* saturated mutation effects,
164 quantifying the predicted changes in accessibility probability for all possible single nucleotide
165 mutations within 2kb of each OCR (whether or not the mutations were reported), inspired by
166 the previous study working on transcript expression levels [27] (Methods). We aggregated and
167 took the mean of mutation effects within 2kb of each OCR and called this value CAMBUS.
168 Hereafter, we defined OCRs with negative CAMBUS as CAMBUS+ (where the '+' denotes
169 positive constraint) based on the local false discovery rate (FDR) method (Fig. 3A, 3B;
170 Method). Besides, we defined a more relaxed CAMBUS+ set by extracting the bottom 5% of
171 the overall CAMBUS distribution as relaxed CAMBUS+ OCRs. The median number of
172 CAMBUS+ OCRs was 1,584, covering 0.056% of the human genome, while that of relaxed
173 CAMBUS+ OCRs was 12,632, covering 0.33%. Notably, read counts from ATAC-seq were
174 not a determinant factor for CAMBUS+ OCRs (Extended Data Fig. 3A), reinforcing that our

175 machine-learning approach incorporated information from DNA sequence patterns into
176 OCRs, not from technical fluctuations or noises.
177 To confirm our key assumption that CAMBUS+ OCRs (or relaxed CAMBUS+ OCRs) are
178 under positive constraint, we first assessed whether CAMBUS+ OCRs overlapped with
179 evolutionarily conserved regions, as defined by phylogenetic P-values (phyloP) scores
180 derived from conservation across 46 vertebrate species [35]. As expected, phyloP scores in
181 CAMBUS+ OCRs were significantly higher than those in control OCRs across 29 cell types
182 (Method; for example, two-sided *t*-test $P < 1.1 \times 10^{-4}$ CAMBUS+ OCRs in NaiveCD4;
183 Supplementary Table 3). Next, we assessed whether CAMBUS+ OCRs are also supported
184 by the genomic non-coding constraint of haploinsufficient variation (Gnocchi), a constraint
185 metric estimated by comparing observed allele counts in diverse human populations with the
186 estimated counts in each 1kb region of the human genome [24]. We found that CAMBUS+
187 OCRs significantly cover the strongly constrained regions compared to control OCRs (for
188 example, two-sided *t*-test $P = 7.5 \times 10^{-59}$ for CAMBUS+ OCRs in NaiveCD4). Although
189 CAMBUS+ OCRs contained a greater proportion of coding regions than control OCRs
190 (Extended Data Fig. 3B), the same trend was observed even when the analysis was
191 restricted to non-coding regions (two-sided *t*-test $P = 2.2 \times 10^{-40}$ for CAMBUS+ OCRs in
192 NaiveCD4; Fig. 3C; Extended Data Fig. 3C). Following the previous study [24], we evaluated
193 the odds ratio (OR) of CAMBUS+ OCRs for the region with Gnocchi ≥ 4, comparing with the
194 OR of the known candidate regulatory elements: ENCODE cCREs (promoter-like, proximal
195 enhancer-like, distal enhancer-like, and CTCF-only cCREs), FANTOM5 enhancers, and
196 super-enhancers. The CAMBUS+ OCRs showed the strongest constraint (OR=11.45 (95%
197 confidence interval (CI) 9.33–14.05), $P=4.7 \times 10^{-68}$ by Fisher's Exact test) among the
198 compared regulatory elements (Fig. 3D; Supplementary Table 4). Meta-analysis
199 demonstrated the consistent tendency across the 29 cell types (OR=8.71 (95% CI 7.85–9.66),
200 $P=3.1 \times 10^{-365}$ from random-effect meta-analysis; Fig. 3E). Notably, the pooled OR was also

201 greater than that of the ENCODE promoter-like regulatory elements, the strongest constraint
202 candidate regulatory elements in the Gnocchi-proposing paper (OR=7.3) [24], indicating that
203 CAMBUS+ could find OCRs under very strong positive constraints. The relaxed CAMBUS+
204 OCRs also showed the same but weaker tendency (Extended Data Fig. 3D).

205 Previous studies [36] demonstrated that enhancers, especially those activated in T-cells,
206 are enriched in archaic genome sequences due to positive selection. Consistently, we
207 observed that CAMBUS+ OCRs showed significant enrichment in the selection signals [37]
208 (OR=1.4 (95% CI 1.3–1.5), P=1.1 × 10^{-26} from random-effect meta-analysis; Extended Data
209 Fig. 3E), the Neanderthal genome sequences observed in the Japanese population [38]
210 (OR=1.4 (95% CI 1.4–1.5), P=4.8 x 10^{-32} from random-effect meta-analysis; Extended Data
211 Fig. 3F), and the Neanderthal and Denisovans' sequences observed in 1000 Genomes
212 populations (Supplementary Table 5) [39]. Besides, we observed the enrichment of CAMBUS+
213 in human ancestor quickly evolved regions (HAQER) [40] (OR=2.35 (95% CI 1.91–2.89), P=8.3
214 × 10^{-16}).

215 In addition, we observed CAMBUS+ OCRs showed significant enrichment in the
216 pathogenic regulatory variants cataloged in the Human Gene Mutation Database (HGMD) [41]
217 (OR=2.70 (95% CI 2.43–2.99), P = 1.7 x 10^{-77}; from random-effect meta-analysis; Extended
218 Data Fig. 3G; Supplementary Table 6).

219 Taken together, CAMBUS+ OCRs can define novel constrained regions, well supported
220 by other constraint metrics.

221 **Enrichment of active regulatory elements by selecting CAMBUS+ OCRs**

222 All OCRs obtained in this study were well overlapped with the known candidate regulatory
223 elements; for example, 73.66% of all OCRs in NaiveCD4 were overlapped with super-
224 enhancers (the median of 29 cell type is 74.70%) and 4.49% with FANTOM5 [6] enhancers
225 (the median of 29 cell type is 5.10%; Supplementary Table 7), both of which are known to be

226 enriched for causal variants [42]. We further explored that CAMBUS+ OCRs could aid in
227 prioritizing functional regulatory elements (Methods). For example, CAMBUS+ OCRs in
228 NaiveCD4 significantly enriched the known regulatory elements, especially for active
229 enhancers like super-enhancers [43] (OR=9.07 (95% CI 6.80–12.11), Fisher's exact test P =
230 2.0×10^{-102}) and FANTOM5 enhancers (OR=6.16 (95% CI 5.39–7.03), Fisher's exact test P
231 = 6.3×10^{-111}) (Supplementary Table 8). Meta-analysis showed a consistent trend across 29
232 cell types (OR=6.63 (95% CI 5.50–7.99), P = 1.1×10^{-87} from random-effect meta-analysis
233 for super-enhancers; Extended Data Fig. 3H, and OR=5.19 (95% CI 4.87–5.54), P = $4.9 \times$
234 10^{-547} from random-effect meta-analysis for FANTOM5 enhancers; Extended Data Fig. 3I).
235 Remarkably, after adjusting with known constraint metrics (Gnocchi and phyloP), we still
236 observed statistically significant enrichment for prioritizing regulatory elements by CAMBUS+
237 OCRs across the 29 cell types (Fig. 3E; Supplementary Table 9). Reflecting the
238 characteristics of the analyzed cell types (blood and immune cell types), a comparison of
239 enrichment across four tissues (blood/hematopoietic, brain, kidney, and lung) revealed the
240 strongest prioritization of blood's regulatory elements by NaiveCD4 CAMBUS+ OCRs (super-
241 enhancers; OR=6.83 (95% CI 6.07–7.71), P=1.1×10^{-217}; FANTOM5 enhancers; OR=7.64
242 (95% CI 6.64–8.78), P=8.7×10^{-179} from logistic regression; Fig. 3F; Supplementary Table
243 10). This pattern was consistently observed across 29 cell types (Supplementary Table 11).
244 We also confirmed that the relaxed CAMBUS+ OCRs exhibited the same tendency
245 (Extended Data Fig. 3J; Supplementary Table 12). Taken together, CAMBUS+ OCRs (and
246 relaxed CAMBUS+ OCRs) effectively capture cell-type-specific active regulatory elements
247 under positive constraints, particularly enhancers, that were not well captured by the previous
248 methods.
249

250 **Heritability enrichment of CAMBUS+ in blood-relevant diseases and traits**

251 We computed heritability enrichments in CAMBUS+ OCRs using stratified LD score
252 regression (S-LDSC) [44] conditioning on known functional annotations and all OCRs from this
253 study (Methods). We used GWAS summary statistics for 33 complex traits, including 14
254 immune-mediated traits (e.g., multiple sclerosis (MS), rheumatoid arthritis (RA), and Atopic
255 dermatitis (AD)), from European (EUR) and East Asian (EAS) populations (Supplementary
256 Table 13). To avoid false positives [45], we used the relaxed CAMBUS+ OCRs (the median
257 proportion of single nucleotide polymorphism (SNPs) across the 29 cell types was 0.5%) for
258 primary analyses instead of the CAMBUS+ OCRs (the median was 0.07%).

259 As a result, the highly significant heritability enrichment was observed in a wide variety of
260 immune-mediated diseases (FDR < 5%; Fig. 4A; Supplementary Table 14). Particularly,
261 plausible pairs of immune traits and cell types were observed to show the strongest
262 enrichment of heritability among the 29 cell types: MS and T_h17 [46], and systemic lupus
263 erythematosus (SLE) and B cells [47,48] (Extended Data Fig. 4A). Besides, other plausible pairs
264 included pairs of RA with CD4 or CD8, and AD with CD4 (particularly Th2 [49]), or CD8.

265 Based on these results, we hypothesized that polygenic risk scores (PRS) prioritized by
266 CAMBUS+ OCRs could enhance their predictive ability. To evaluate this hypothesis, we
267 focused on CAMBUS+ OCRs in double-negative B cells (DN B), which showed significant
268 enrichment for the heritability of SLE in EAS ancestry (Extended Data Fig. 4A) and was also
269 known for its functional associations with SLE [50]. We constructed the PRS based on variants
270 overlapping CAMBUS+ OCRs and evaluated the additional performance of the conventional
271 PRS using five-fold cross-validation using 137,620 individuals of EAS ancestry (1,178 cases
272 and 136,442 controls; Methods) [51]. We performed logistic regression on the five validation
273 datasets to estimate the effect sizes of the PRS derived from the additional variants,
274 conditioning on the conventional PRS. A fixed effect meta-analysis demonstrated that the

275 additional variants significantly improved the predictive performance of PRS ($P = 3.12 \times 10^{-6}$; Supplementary Table 15; Methods).

277

278 **Putative causal variants were enriched in CAMBUS+ OCRs**

279 We assessed whether CAMBUS+ OCRs could prioritize the putative causal variants for
280 complex traits, using the large-scale statistical fine-mapping studies in 94 traits in EUR
281 ancestry [52] and 62 traits in EAS ancestry [53]. We intersected CAMBUS+ OCRs and the
282 putative causal variants and cataloged 1,210 and 323 for trait-related cell-type-specific
283 putative causal variants in both EUR and EAS ancestries, respectively (Methods). The
284 CAMBUS+ OCRs were significantly enriched for all variants in 95% credible sets (median
285 OR:3.85 and 3.94 in EUR and EAS, respectively, Extended Data Fig. 5A). In line with a data
286 source of OCRs, CAMBUS+ OCRs exhibited significantly higher enrichment for the putative
287 causal variants of leukocyte-related traits than non-hematopoietic traits (two-sided *t*-test $P =$
288 1.0×10^{-11} in EUR ancestry and 1.4×10^{-2} in EAS ancestry; Extended Data Fig. 5B). As an
289 example, the enrichment of CAMBUS+ OCRs for type 1 helper T-cells was stronger among
290 variants with high posterior inclusion probability (PIP) (>0.9) for hematopoietic traits in EAS
291 ancestry than those with low PIP (≤ 0.1) (OR=7.62 (95% CI 3.60–16.18), $P = 1.2 \times 10^{-7}$ from
292 logistic regression; Fig. 5A), and this enrichment was further increased when restricted in the
293 leukocyte traits (OR=11.0 (95% CI 4.04–29.96), $P = 2.7 \times 10^{-6}$; Fig. 5A). Meta-analysis
294 showed this trend was consistent across 29 cell types (OR=8.99 (95% CI 7.85–10.29), $P =$
295 8.8×10^{-221} for EUR ancestry; OR=19.24 (95% CI 16.03 – 23.09), $P = 3.9 \times 10^{-221}$ for EAS
296 ancestry; Fig. 5B). The median regional length of CAMBUS+ (845 bp) was much shorter than
297 that of 95% credible sets (3.0M bp for EUR ancestry and 1.1M bp for EAS ancestry). Taken
298 together, CAMBUS+ OCRs can pinpoint or at least narrow down causal variants. Notably,
299 even for rare variants (minor allele frequency (MAF) < 0.01), the putative causal variants of
300 hematopoietic traits were significantly enriched in CAMBUS+ OCRs in at least one cell type

301 (Methods; OR=4.68 (95% CI 2.66–7.99), two-sided Fisher's exact test P = 2.23 × 10^{-6}; Fig.
302 5C). This suggested that CAMBUS+ OCRs can potentially prioritize even rare causal variants
303 from GWAS results which are difficult to prioritize in the conventional statistical fine-mapping
304 [20].
305 Leveraging the overlapped putative causal variants, we found plausible examples of putative
306 causal variants where CAMBUS+ OCRs led to the biological interpretation of GWAS findings.
307 First, rs13122305 was the only one overlapping with CAMBUS+ OCRs in monocytes among
308 seven candidate causal variants in a credible set for monocyte count (Fig. 5D). The nearest
309 gene of this variant is *COQ2*, whose anti-inflammatory functions via regulating gene
310 expression were reported [54,55]. Additionally, we found that the Activity-by-Contact (ABC)
311 model's prediction (ABC enhancer) in monocytes only overlapped with rs13122305 and that
312 *COQ2* was also supported by the ABC model with the second-highest ABC score (0.025)
313 among the eight target genes. Second, among the three candidate causal variants in a
314 credible set in the FCGR region for Albumin/Globulin ratio (AG), only rs2487451 overlapped
315 with non-classical Monocyte CAMBUS+ OCR. Non-classical monocytes, known to display
316 antigen presenting properties [56], were the only cell type among 29 cell types whose OCR
317 overlapped with a candidate causal variant. The nearest gene was *FCGR3B*, encoding a
318 receptor (FcγRIIIB) that binds to the Fc region of IgG (Immunoglobulin) antibodies, which is
319 known to be expressed in non-classical monocytes [57]. Third, rs55646091, a putative causal
320 variant for asthma (PIP = 0.99), overlapped with CAMBUS+ OCR of regulatory T-cells (T-
321 reg) (Fig. 5E) among the 29 cell types. The nearest gene of this variant was *LRRC32*
322 (leucine-rich repeat-containing 32), which regulates TGF-beta. LRRC32 was potentially
323 involved in Asthma [58] and whose encoding protein was known as a surface marker for T-reg.
324 Furthermore, this variant-gene pair was also supported by the ABC model and *LRRC32*
325 displayed the highest ABC score (ABC score = 0.041 that is much higher than the threshold
326 corresponding to 70% recall and 59% precision) [59] among the nine genes connected with the

327 regulatory element overlapping with the OCR and the variant. Fourth, we found that
328 rs8009224 was the only one that overlapped with CAMBUS+ OCRs of plasmablast
329 (Extended Data Fig. 6A) among the seven putative causal variants for hemoglobin A1c levels.
330 This is an intronic variant of *MIDEAS* (Mitotic Deacetylase Associated SANT Domain Protein),
331 which is low frequent in both 1KG EUR (MAF = 3.4%) and EAS (MAF = 1.2%) and also
332 associated with mean corpuscular hemoglobin concentration [60]. The overlapping regulatory
333 element was connected to the 16 genes by the ABC enhancer in B-cell, and the putative
334 target gene with the second highest ABC score is *MIDEAS* (ABC score = 0.097, much higher
335 than the proposed threshold corresponding to 70% recall and 59% precision) [59]. Lastly, we
336 demonstrated that CAMBUS+ in T-reg could prioritize only one putative causal variant
337 (rs9814991) from numerous variants (fifty-seven) in this credible set for eosinophils count
338 (Extended Data Fig. 6B).

339

340 **Discussion**

341 In this study, we developed CAMBUS based on the hypothesis that DNA
342 sequencing patterns surrounding OCRs contain hidden information defining OCR constraints.
343 Although CAMBUS+ OCRs were derived from a fundamentally different approach, they were
344 well supported by conventional constraint metrics as expected, while also capturing high-
345 resolution, cell-type-specific regions that are both active and constrained, beyond existing
346 annotations. An alternative potential approach using ATAC-seq (or DNase-seq) could
347 prioritize regions with high accessibility (i.e., regions with many read counts). However,
348 previous studies have suggested no relationship between accessibility and functionality or
349 constraints [61-63]. Consistently, accessibility was not a determinant factor in defining
350 CAMBUS+ OCRs. Thus, leveraging surrounding DNA sequence patterns through machine
351 learning offers a unique and effective strategy for identifying biologically important OCRs.

352 CAMBUS+ OCRs can mark active and functional regulatory elements, particularly
353 tissue-specific active enhancers, even after being adjusted for known constraint metrics
354 (Gnocchi and phyloP). Furthermore, CAMBUS+ OCRs captured a significant proportion of
355 the heritability for many complex traits, highlighting their biological and genetic relevance.
356 CAMBUS+ OCRs effectively prioritized regulatory causal variants for immune-related traits,
357 even rare causal variants. By defining biologically constrained regulatory elements,
358 CAMBUS+ OCRs will provide a foundation for understanding the regulatory architecture of
359 complex traits and their biological insights into genetic variations especially in non-coding
360 regions.

361 We demonstrate that CAMBUS+ OCRs can effectively prioritize causal rare variants
362 for complex traits. A previous study showed that rare variants tend to be enriched in the
363 constraint genome [64], consistent with our study as CAMBUS+ OCRs can mark such
364 constrained regions. In GWAS using SNP arrays, low statistical power and low imputation
365 accuracy for rare variants make it difficult to detect rare variant associations and identify
366 causal rare variants through statistical fine-mapping. Despite these challenges, CAMBUS+
367 OCRs can facilitate the rare variant prioritization that might otherwise be missed.

368 Eosinophilic asthma is a subtype characterized by elevated levels of eosinophils [65].
369 In this analysis, we identified a variant (rs55646091) associated with asthma linked to
370 increased eosinophil counts (PIP = 0.03). This variant exhibited positive effect sizes in both
371 traits (0.20 for asthma, 0.072 for eosinophil counts), suggesting that carriers of this variant
372 demonstrate elevated eosinophil levels, which contribute to inflammation and the onset of
373 asthma [66]. Notably, this variant overlaps specifically with the OCR in T-reg cells among 29
374 blood cell types analyzed. The nearest gene of this variant is *LRRC32*, which regulates TGF-
375 beta, a known surface marker of T-reg cells, which has been previously implicated in immune
376 regulation [67]. The findings from this analysis provide additional evidence supporting a
377 biological mechanism in which the variant, through its effects on T-reg cells, leads to

378 increased eosinophil counts, thereby contributing to the development of asthma. This insight
379 deepens our understanding of the role of T-reg cells and *LRRC32* in asthma pathogenesis,
380 particularly in eosinophilic inflammation.

381 In this study, we focused on various immune cell types; however, our approach is
382 broadly applicable to other tissues and cell types. As GWAS results continue to accumulate,
383 integrating CAMBUS+ OCRs with these expanding resources has the potential to uncover
384 further insights into the regulatory mechanisms underlying complex traits. Moreover, this
385 method can be applicable to OCRs identified at the single-cell level, further extending its
386 applicability. Taken together, CAMBUS+ OCRs not only capture constrained and functional
387 regulatory elements but also facilitate the interpretation of non-coding genetic variants by
388 prioritizing causal variants in a cell-type-specific manner without relying on LD information
389 and allele frequencies.

390

391 ## Method

392 ### Sample and library preparation

393 We collected peripheral blood (PB) cells from 55 healthy Japanese volunteers and sorted
394 them into 29 cell types. PBMCs were isolated by density gradient separation with Ficoll-
395 Paque (GE healthcare) immediately after blood drawing. Erythrocytes were lysed with a
396 potassium ammonium chloride buffer, and non-specific binding was blocked with Fc-gamma
397 receptor antibodies. PBMC is stained with cell surface specific antibodies (Supplementary
398 Table 22). 27 immune cell populations were sorted by FACS (BDAria III) using the gating
399 strategy (Supplementary Table 1; Extended Data Fig. 7). In addition, neutrophils were
400 isolated by MACSxpress Neutrophil Isolation Kits human (Miltenyi Biotec) from 4 × 106
401 PBMC cells. Sorted cells, isolated PBMC, and Neutrophils were stored at −80°C in CytoStor-
402 10 (Charles River Laboratories Cell Solutions). Libraries for ATAC-seq were prepared using

403 a modified protocol of Omni-ATAC [68]. Genomic DNA from whole blood was isolated using
404 QIAamp DNA Blood Mini Kits, respectively. ATAC-seq was performed using 150bp paired-
405 end on Illumina NovaSeq6000. In total, we obtained a median of 13.34M (+/- 7.56M) reads
406 per sample. This study was approved by the Ethics Committees of RIKEN. Written informed
407 consent was obtained from each volunteer.

408 **Mapping ATAC-seq data**

409 ATAC-seq data were processed following the ENCODE ATAC-seq pipeline (version 1.8.0)
410 as follows: Raw reads were trimmed using cutadapt (version 2.5) and aligned to human
411 genome assembly GRCh38 analysis set using bowtie2 (version 2.3.4.3) for each sample.
412 After alignment, duplicated reads were removed using MarkDuplicates (Picard version
413 2.20.7-SNAPSHOT), retaining uniquely mapped reads and improperly paired reads using
414 samtools (version 1.9), and removed reads that mapped to the mitochondrial genome.

415 In total, we obtained a median of 5.39M (s.d. 3.28M) QCed reads per sample.

416 **ATAC-seq peak calling**

417 For each 29 immune cell type, we used the samtools merge command to merge the QCed
418 reads for each cell type and used MACS2 (version 2.1.1) to identify the ATAC-seq peaks
419 (OCRs) following a recent study [26] with parameters (-p 0.01 --shift -75 –extsize 150 --
420 nomodel). We obtained a median of 253,553 (±66,738) OCRs with a median peak width of
421 271bp (±18bp) (FDR<0.05). Using the alignment file for each cell type, we calculated the
422 read counts for each OCR with bedtools (version 2.30.0)'s coverage function.

423

424 **Quality control for ATAC-seq dataset**

425 We first assessed the quality of ATAC-seq data using Pearson correlations of chromatin
426 accessibility among the samples for each cell type, using 614,516 peaks that were identified
427 from across all samples using the same parameters described above (these peaks were only
428 used for the quality control purpose). We filtered OCRs out when the Pearson correlations
429 were not greater than 0.65 (determined by manual inspection of the distribution of these
430 correlations). Additionally, we excluded outlier samples with low Pearson correlations by
431 visual inspection (Supplementary Table 16). Samples with a mapped rate of less than 0.5
432 were further excluded. Finally, we obtained 1,386 samples (the median number of donors
433 per cell type is 48; Supplementary Table 1).

434 **Hierarchical clustering and dimension reduction**

435 We first chose the top 5,000 accessible OCRs (defined by TPM) from each cell type and
436 merged them across 29 cell types, resulting in a total of 145,000 unique regions. We then
437 created a binary matrix (0, 1) to indicate whether each 300bp bin across the genome was
438 overlapped by the corresponding OCRs. Hierarchical clustering was performed using Ward's
439 method of Euclidean distance using the binary matrix. PCA for the binary matrix was
440 performed using the binaryPCA function from the scBFA package (version 1.16.0) [69] in R
441 with default parameters. UMAP was performed with the top five PCs using the umap function
442 from the scikit-learn package (version 0.18.1) in Python with default parameters.

443 **Strategy for the CAMBUS ML model**

444 The CAMBUS ML model was built upon the DeepSEA architecture, a deep CNN model [30].
445 We used the Selene library (v0.4.7) to train the CAMBUS ML model. A 5kb reference
446 sequence, randomly chosen with Selene, is used as input features, five times longer than

447 the sequence used in the original DeepSEA model to capture broader genomic contexts. The
448 model outputs a sigmoid function-derived score, which can be interpreted as a proxy for the
449 likelihood that the OCRs overlap with the central 200bp of the input sequence (see
450 Supplementary Table 17 for the configuration details). We referred to this predicted
451 probability for chromatin accessibility as the accessibility probability.

452 We combined 7,939,714 OCRs across 29 cell types using the `bedtools merge` function and
453 obtained 1,349,934 non-overlapping unique input sequences. According to Selene's default
454 settings, we used chromosomes 8 and 9 as testing, chromosomes 6 and 7 as validation, and
455 the other autosomal chromosomes as training data. The details of Selene's parameters are
456 described in Supplementary Table 17.

457 ***In silico* saturated mutagenesis using the trained CAMBUS ML model**

458 For all bases within 1kb from the center of each OCR, we used the trained CAMBUS ML
459 model to calculate the predictions for all possible single-based mutations. This
460 comprehensive approach provides an unbiased analysis of mutational effects across the
461 entire sequence, rather than focusing on only specific known polymorphisms. To quantify the
462 mutation effect, we calculated the difference in log fold changes of predictions between the
463 reference and alternative alleles for each site and averaged them across all bases. In total,
464 we obtained 505 million mutation effects (median across the 29 cell types).

465 **Validation of CAMBUS ML model with *in silico* mutagenesis**

466 We compared the *in silico* mutation effects predicted by the CAMBUS ML model with effect
467 sizes from the caQTL study. The samples for caQTL (n=74, including additional independent
468 samples to increase statistical power) were genotyped using Infinium ASA-24+v1.0 kit, and
469 whole genome imputation was conducted after stringent quality control (Supplementary

470 Table 18). The caQTLs were mapped using RASQUAL (v1.1) with default parameters. The
471 permutation to empirically obtain the null distribution for each region was performed with
472 default parameters, resulting in a median of 654 lead caQTLs with FDR < 0.1, which we
473 compared against the *in silico* mutation effects. Additionally, we calculated the *in silico*
474 mutation effects for 297,308 autosomal lead SNVs in the previous study's caQTL. The
475 positions of those lead SNVs were mapped to the hg38/GRCh38 coordinate using the liftOver
476 tool with the hg19ToHg38.over.chain file to be consistent with our study. To further refine the
477 comparison, we filtered the target caQTL peaks>50% overlapping with our OCRs and
478 selected the variants inside the peak. In total, we used 156,039 caQTLs peak pairs to
479 compare with the mutation effect.

480 **Identification of putative constraint open chromatin region**

481 We modeled the distribution of the mean mutation effect (*in silico* saturated mutagenesis)
482 within 2kb of each OCR (CAMBUS) as a mixture of a Gaussian null distribution. We
483 hypothesized that the OCRs with negative CAMBUS are likely positive constraint regions
484 and OCRs with neither negative nor positive CAMBUS as control OCRs. We calculated local
485 FDR with the locfdr function in the locfdr R package (version 1.1-8) for each cell type. This
486 analysis identified a median of 1,584 putative positive constraint OCRs across 29 cell types
487 with FDR < 0.5, referred to as CAMBUS+ OCRs, and a median of 248,749 control OCRs
488 across 29 cell types [27]. We also defined relaxed CAMBUS+ OCRs by extracting the bottom
489 5% of the CAMBUS distribution for each cell type (Supplementary Table 19).

490

491 **Enrichment of CAMBUS+ OCRs in constraint regions**

492 We evaluated whether CAMBUS+ OCRs are enriched in constraint regions of the genome
493 using Gnocchi. We used the 1,984,900 QCed Gnocchi (see URLs) and calculated the OR
494 following the methodology described in the previous paper [24] for 1,843,559 non-coding 1-kb
495 windows. In brief, we counted the number of overlapping 1kb windows within the following
496 bins: (∞, -4), [−4,−3), [−3,−2), [−2,−1), [−1,−0), [0,1), [1,2), [2,3), [3,4), [4,∞) with each
497 CAMBUS+ OCR. ORs were calculated by comparing these overlaps with all 1kb windows.
498 Fisher's exact test was conducted to estimate the statistical significance (p-value). The OR
499 of the known regulatory elements for each bin was obtained from the gnomAD download site
500 (see URLs) and used for the comparison.

501 To calculate the enrichment of the phyloP score in CAMBUS+ OCRs, we used the scores for
502 all 46 species (see URLs) and mapped the positions of those datasets to the hg38/GRCh38
503 coordinate using liftOver as described previously. The overlapping regions were considered
504 to be overlapped if they had more than 50% overlap with each other.

505 **Enrichment of CAMBUS+ OCRs in the positive selection signals**

506 We evaluated whether CAMBUS+ OCRs are enriched in the positive selection signals
507 observed in the previous study [37]. For each CAMBUS+ OCR, we counted overlapping
508 annotations for the positive selection signals observed in the EUR population and estimated
509 the OR compared with all OCRs using a logistic regression model. To perform enrichment
510 analysis of HAQERs [40], we first created merged OCRs by combining CAMBUS+ OCRs from
511 29 different cell types (merged CAMBUS+ OCRs). The number of overlapped merged
512 CAMBUS+ OCRs in the HAQERs was counted, and OR was estimated compared with all
513 OCRs using a logistic regression model.

514 Also, we evaluated whether CAMBUS+ OCRs are enriched in the archaic genome. For each
515 CAMBUS+ OCR, we counted overlapping annotations for the Neanderthal genome
516 sequences observed in the Japanese population [38] and the Neanderthal and Denisovans'
517 sequences observed in 1000 Genomes populations [39], and estimated the OR compared with
518 all OCRs using a logistic regression model.

519 **Correlation between CAMBUS and regulatory elements**

520 We evaluated the functional enrichment of CAMBUS+ OCRs in known regulatory elements,
521 including ENCODE cCREs, FANTOM5 enhancers, and super-enhancers (see URLs). For
522 each CAMBUS+ OCR, we counted overlapping annotations for each known regulatory
523 element and, calculated the OR compared with all OCRs, and tested it using Fisher's exact
524 test, as described in the previous section. For tissue-expressed FANTOM5 enhancers, we
525 obtained count per million (CPM) data for 249,599 enhancers across 347 tissue/cell types
526 with hg19 coordinates (see URLs) and selected enhancers expressed in hematopoietic_cell
527 (N = 163), brain (N = 85), kidney (N = 16), and lung (N = 8) based on a threshold of CPM >=
528 0.1 in at least 10% of individuals in each enhancer. As described previously, the positions
529 were mapped to the hg38/GRCh38 coordinate using liftOver. For tissue-expressed super-
530 enhancers, we obtained Blood (N = 115), brain (N = 203), kidney (N = 34), and lung (N =
531 119) enhancers with hg38 coordinates (see URLs) and merged these enhancers by tissue
532 with `bedtools merge` function. The adjusted OR with Gnocchi and phyloP was estimated
533 using a logistic regression model, adjusting for the average Gnocchi and phyloP of
534 overlapping CAMBUS+ OCRs and control CAMBUS OCRs.

535 For all analyses, regions were considered to be overlapped if they had more than 50%
536 overlap with each other.

537

538 **Heritability enrichment analysis**

539 To estimate heritability enrichment in each OCR, we conducted stratified-linkage
540 disequilibrium score regression (S-LDSC) with LDSC (ver.1.0.1) using the baseline model
541 (version 1.2; 53 categories). The p-value of the coefficient for CAMBUS+ OCR (or relaxed
542 CAMBUS+ OCR) was obtained by adjusting for the 53 categories in the baseline model and
543 all OCRs in each cell type. Multiple testing correction across all trait-cell type pairs was
544 performed to estimate FDR using the Storey method [70]. The summary statistics were
545 obtained from Alkes Price Lab (See URLs and our previous study [9]), and five GWAS results
546 in EAS ancestry: systemic lupus erythematosus [51], rheumatoid arthritis [71], Systemic sclerosis
547 [72], Atopic dermatitis [73], and Takayasu arteritis [74]. For this analysis, CAMBUS OCRs' positions
548 were mapped to the hg19 coordinate using the UCSC liftOver tool with
549 hg38ToHg19.over.chain file to be consistent with summary statistics. After mapping to hg19,
550 regions mapped to different chromosomes were excluded from the analysis.

551 **PRS analysis**

552 The PRS was constructed using PLINK v1.9, following a systematic approach with p-value
553 thresholds using SLE GWAS in EAS ancestry. To select relevant variants for the PRS, we
554 selected SNPs that met 20 types of p-value thresholds: 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2,
555 0.1, 0.05, 0.01, 0.005, 5.0×10^{-4}, 5.0×10^{-5}, 1.0×10^{-5}, 5.0×10^{-6}, 1.0×10^{-7}, 5.0×10^{-7} and
556 5.0×10^{-8}. After selecting SNPs based on these thresholds, we applied LD clumping to
557 remove SNPs in high linkage disequilibrium (LD) using nine types of r^2 thresholds (0.1, 0.2,
558 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) and a 250kb physical distance. This process yielded 180 PRS
559 models. To determine the optimal PRS model, we applied a 5-fold cross-validation strategy.
560 The dataset was divided into five folds, with four folds used for training and one for testing in
561 each iteration. We computed PRSs for all 180 models using variants from SLE GWAS in

562 EAS ancestry [51] for each training set. We calculated the AUC with pROC (v1.18.4) to select
563 the combination of p-value and r^2 parameters that yielded the best AUC for each training set.
564 We then took the two variant sets; (a) variants filtered with the best parameters, and (b) all
565 variants overlapping with DN B CAMBUS+ OCRs as prioritized variants. To mitigate
566 overfitting due to LD structures, we removed the variants with LD>0.7 with variants in set (b).
567 We then calculated the PRS for both sets (a) and (b) for each of the five test sets, we
568 estimated the effect size of PRS from set (b), controlling for the PRS from set (a) using a
569 logistic regression model. Finally, we performed a fixed-effects meta-analysis to evaluate the
570 improvement in predictive power with the prioritized variants.

571 **Fine-mapping enrichment analysis**

572 We downloaded the UKB fine-mapped variants of 94 complex diseases and traits from the
573 Finucane lab (see URLs) and obtained BBJ fine-mapped variants of 62 quantitative traits
574 from a recent study [53]. Regarding the UKB fine-mapped variants, we used PIP from the
575 SuSiE. The positions of those datasets were mapped to the hg38/GRCh38 coordinate using
576 liftOver as described previously. In total, we obtained 830,654 trait-variant pairs for UKB and
577 544,745 trait-variant pairs for BBJ. The number of overlapped variants for each trait in 29 cell
578 types of CAMBUS+ OCRs was counted, and the OR was estimated using all OCRs in each
579 29-cell type as the background and calculated P-value from the logistic regression model.
580 After Bonferroni correction (P=0.05 divided by the number of tested phenotypes), the median
581 of OR is calculated from the significant enrichment in UKB and BBJ (Extended Data Fig. 5A,
582 B; Supplementary Table 20).

583 We categorized each trait into hematopoietic and non-hematopoietic. Furthermore,
584 hematopoietic traits are categorized into leukocyte and non-leukocyte traits (Supplementary

585 Table 21). The number of traits and variants in each category are described in Supplementary
586 Table 21.

587 Then, we counted the overlap of variants for each leukocyte and non-leukocyte and non-
588 hematopoietic for each 29 cell-type CAMBUS+ OCRs. We calculated the OR using the
589 method described above.

590 To perform enrichment analysis of rare variants, we first created merged OCRs by combining
591 CAMBUS+ OCRs from 29 different cell types (merged CAMBUS+ OCRs). The number of
592 overlapped variants with MAF < 0.01 and PIP > 0.1 in merged CAMBUS+ OCRs was counted
593 for all traits and only hematopoietic traits, and the OR was calculated using the variants with
594 MAF < 0.01 and PIP <= 0.1 as the background and calculated one-sided Fisher's exact P-
595 value.

596 **Catalog cell-type-specific putative causal variants**

597 We chose variants as putative causal variants from each of the credible sets overlapping
598 only one variant with CAMBUS+ OCRs. Then, we filtered the variants that overlapped with
599 only one cell type of CAMBUS+ OCRs. In total, we obtained 186 and 62 cell-type-specific
600 putative causal variants. Furthermore, we chose the variants that overlapped only one
601 subgroup (one of the CD4, CD8, B, Mono, DC, NK, Neutrophil; Supplementary Table 1).
602 Finally, we obtained 1,210 and 323 for trait-related cell-type-specific putative causal variants
603 from both EUR and EAS ancestries, respectively.

604 **Statistical analysis**

605 For p-value calculation with R, in case of insufficient precision with standard floating-point
606 arithmetic, the Rmpfr package (ver 0.9.2) was used to handle high-precision floating-point

arithmetic to calculate p-value with z-score. For the meta-analysis, the pooled OR was calculated using a random effect model with metafor R package (version 2.0-0).

Data availability

The CAMBUS will be publicly available online soon after acceptance.

Code availability

The CAMBUS and its calculation scripts will be publicly available online soon after acceptance.

URLs

Gnocchi is at https://gnomad.broadinstitute.org/downloads#v3-genomic-constraint. Enrichment for Gnocchi with the known cCREs is at gs://gnomad-nc-constraint-v31-paper/fig_tables/constraint_z_genome_1kb.annot.txt. ENCODE cCREs V3 is at https://screen.encodeproject.org/. FANTOM5 enhancer is at https://zenodo.org/record/556775/. Super enhancer is at https://bio.liclab.net/sedb/download.php. Phylop is at https://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/primates. GWAS summary statistics are at https://alkesgroup.broadinstitute.org/sumstats_formatted, and https://alkesgroup.broadinstitute.org/UKBB/. ABC predictions are at https://mitra.stanford.edu/engreitz/oak/public/Nasser2021/AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz. UKB fine-mapped variants are at https://www.finucanelab.org/data. FANTOM5 tissue-expressed data is at https://zenodo.org/records/5348471.

Acknowledgment

We would like to thank all the participants in the study. We deeply thank all the members of the Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences for the technical support. We acknowledge the use of Biorender for aspects of the figures. This work was supported by Japan Agency for Medical Research and Development (AMED) grants 21ek0109555, 21tm0424220, 21ck0106642, 23ek0410114, and 23tm0424225. Japan Society for the Promotion of Science (JSPS) KAKENHI grant JP20H00462 and JP23K14164, Takeda Hosho Grants for Research in Medicine.

Author contributions

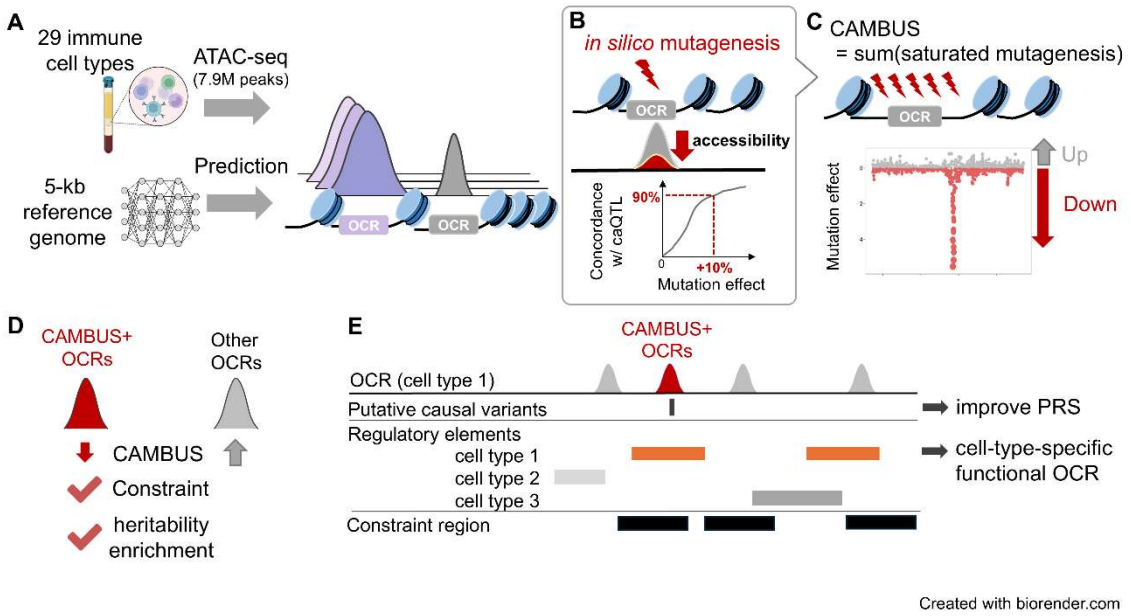
K.T., M.K., and C.T. conceived the study design. A.S. performed ATAC-sequencing and DNA genotyping. K.T. and N.T. performed whole genome imputation. K.T. analyzed this study with the help of S.Y., N.T., Y.I., X.L., S.K., and M.K. M.K. and C.T. supervised this study. S.Y., N.T., S.K., K.I., Y.M., Immune transcript/enhancer consortium, and K.Y. provided helpful comments and discussion. K.T., M.K., A.S., and C.T. wrote the manuscript.

Ethics declarations

Competing interests

The authors declare no competing financial interests.

646 Figure legends



647

648 **Fig. 1 | Study design.** Overview of the study design. **A,** Dataset comprising 29 cell types
649 was used to validate our approach, and a machine learning model was developed to predict
650 OCRs. **B,C,** Identification of putative constraint OCRs (CAMBUS+). **D,** Characteristics of
651 CAMBUS+ OCRs based on CAMBUS. **E,** Characterization of CAMBUS+ OCRs and
652 prioritization for putative causal variants.

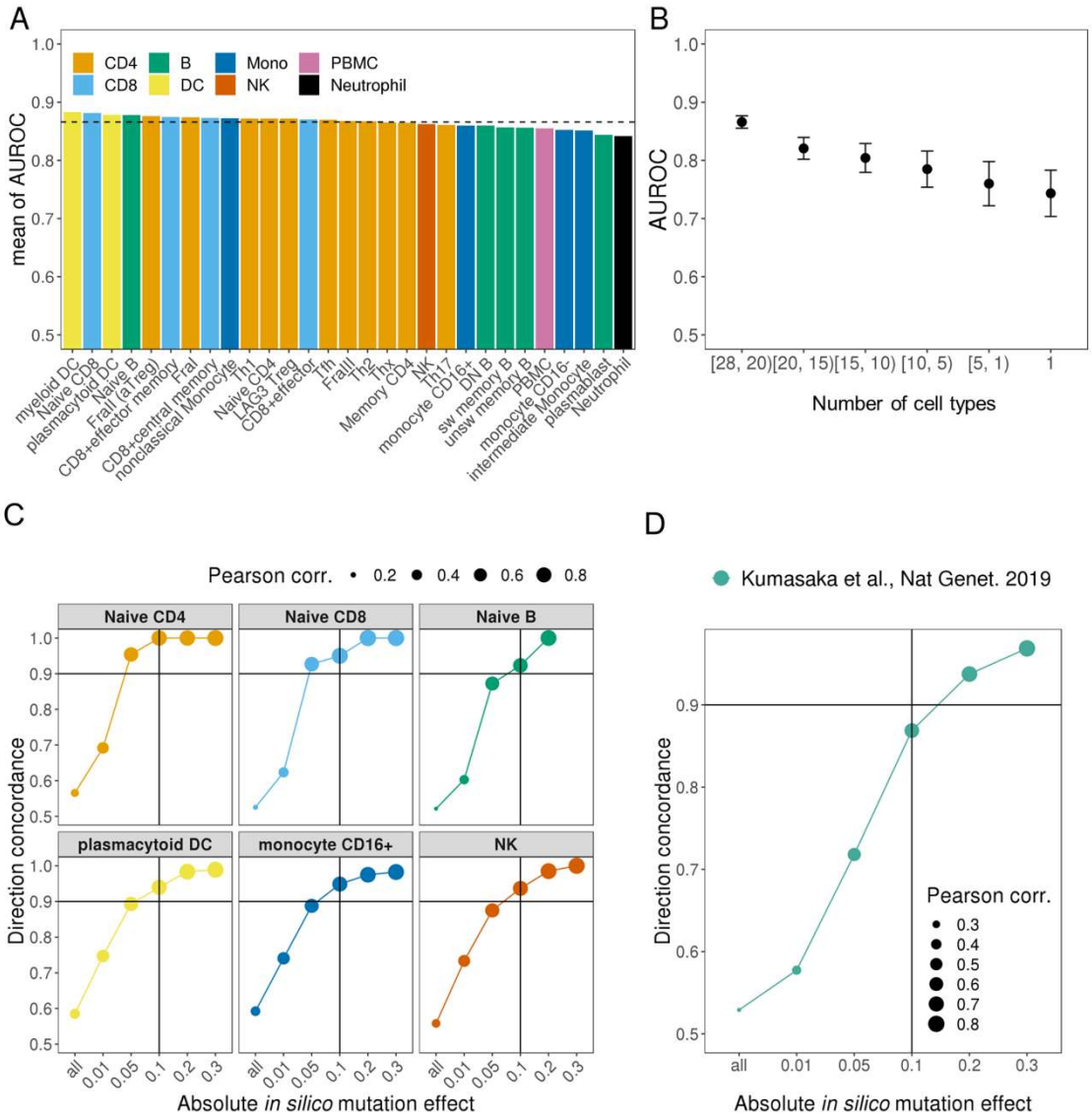
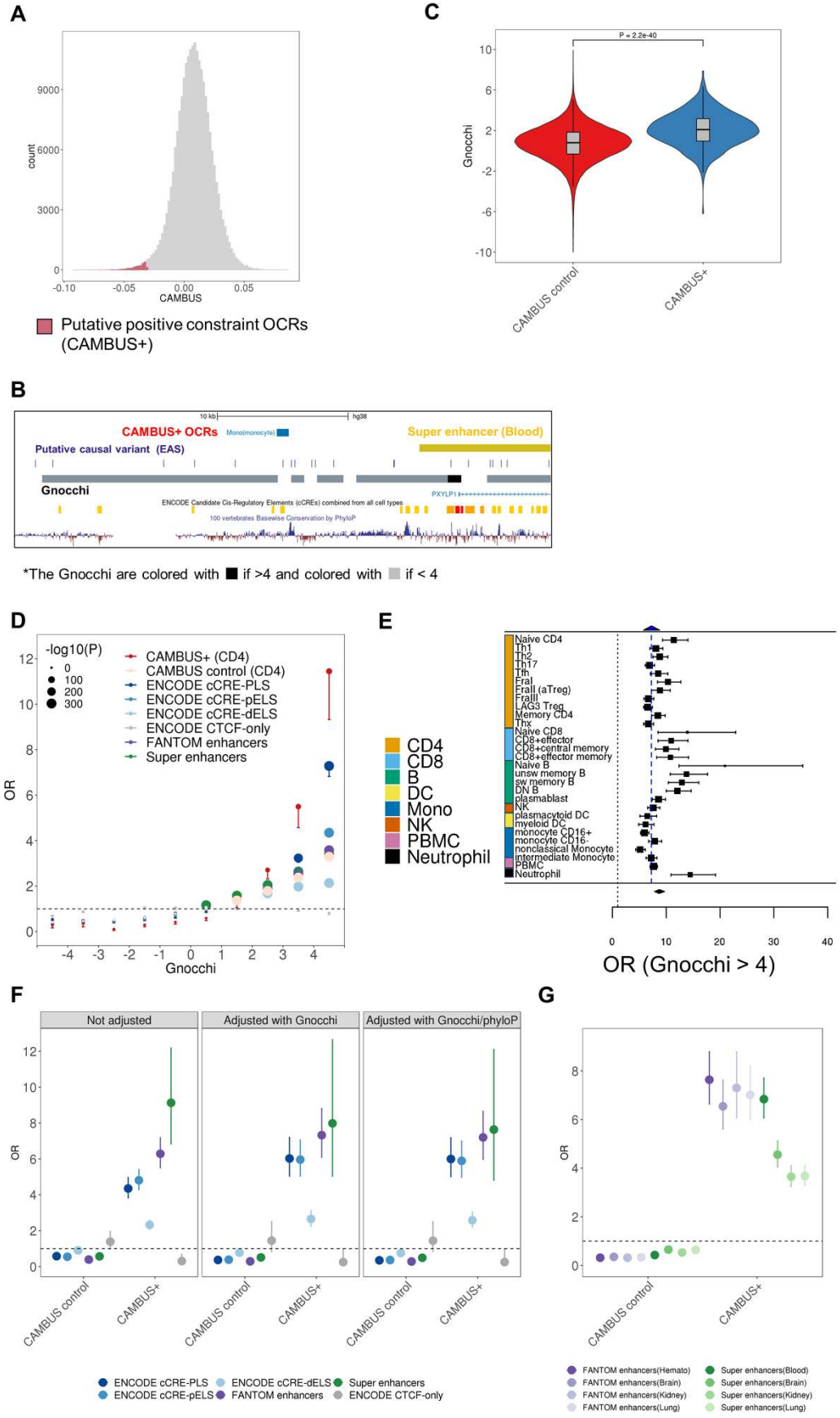


Fig. 2 | Accuracy of CAMBUS ML model and its mutation effect. A, Predictive accuracy of the CAMBUS ML models across 29 cell types. The dashed line represents the mean AUROC across the 29 cell types (0.87). **B,** Specificity of CAMBUS ML models with respect to OCR sharing. The x-axis shows the number of cell types sharing OCRs. For example, the leftmost category represents OCRs shared among 20 to 28 cell types, and the rightmost category indicates OCRs unique to a single cell type. **C,** Direction concordance between *in*

660 *silico* mutation effect from CAMBUS ML and effect size from our caQTL study. **D,** Direction
661 concordance between *in silico* mutation effect from the CAMBUS ML for Naïve B and effect
662 size from independent LCL caQTL study. The Pearson correlation coefficient (Pearson corr.)
663 was calculated to assess the relationship between *in silico* mutation effects and eQTL effect
664 sizes.

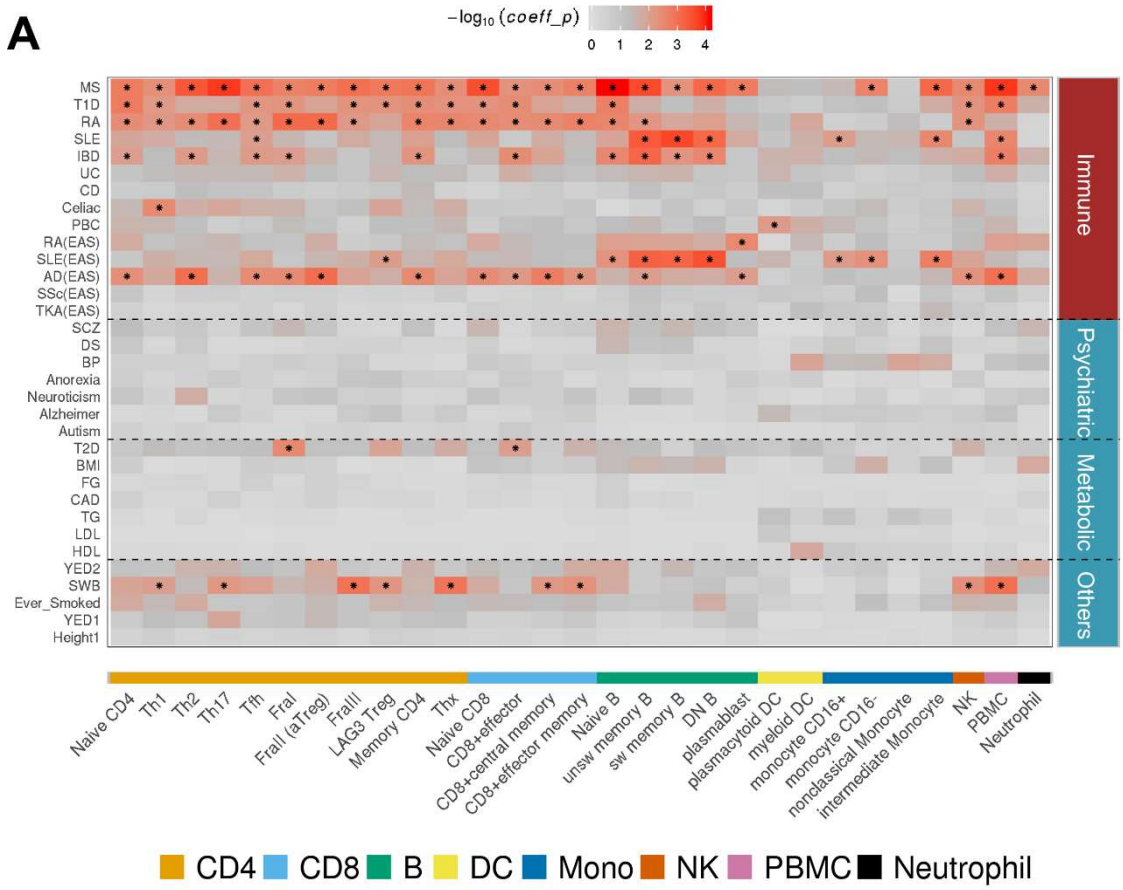
665


A
count
CAMBUS
Putative positive constraint OCRs (CAMBUS+)
C
P = 2.2e-40
Gnocchi
CAMBUS control
CAMBUS+
B
CAMBUS+ OCRs
Super enhancer (Blood)
Putative causal variant (EAS)
Gnocchi
*The Gnocchi are colored with ■ if >4 and colored with ■ if < 4
D
-log10(P)
CAMBUS+ (CD4)
CAMBUS control (CD4)
ENCODE cCRE-PLS
ENCODE cCRE-pELS
ENCODE cCRE-dELS
ENCODE CTCF-only
FANTOM enhancers
Super enhancers
OR
Gnocchi
E
CD4
CD8
B
DC
Mono
NK
PBMC
Neutrophil
OR (Gnocchi > 4)
F
Not adjusted
Adjusted with Gnocchi
Adjusted with Gnocchi/phyloP
OR
CAMBUS control
CAMBUS+
ENCODE cCRE-PLS
ENCODE cCRE-pELS
ENCODE cCRE-dELS
FANTOM enhancers
Super enhancers
ENCODE CTCF-only
G
OR
CAMBUS control
CAMBUS+
FANTOM enhancers(Hemato)
FANTOM enhancers(Brain)
FANTOM enhancers(Kidney)
FANTOM enhancers(Lung)
Super enhancers(Blood)
Super enhancers(Brain)
Super enhancers(Kidney)
Super enhancers(Lung)

666

667 **Fig. 3 | CAMBUS+ OCRs can mark constraint regions and active enhancers. A,**
668 Distribution of CAMBUS. The red part represents putative constraint OCRs. **B,** Genome
669 browser view at representative constraint region around CAMBUS+ in monocytes
670 (chr3:141,217,899-141,218,844). The black box represents Gnocchi with Z > 4, and the grey
671 box represents Gnocchi with Z < 4. **C,** Distribution of non-coding Gnocchi in CAMBUS+
672 OCRs and CAMBUS control in NaiveCD4. The t-tests were performed between CAMBUS+
673 OCRs and CAMBUS+ control. **D,E,** Enrichment (odds ratio; OR) of CAMBUS+ in non-coding
674 Gnocchi (n = 1,843,599), comparing with the known regulatory elements (using the previous
675 study's results, see URLs) in non-coding Gnocchi (n = 1,843,599). We calculated the OR of
676 CAMBUS+ (n = 1,266) and CAMBUS control (n = 251,807) for NaiveCD4 using various
677 Gnocchi' thresholds in (**D**) and OR for each CAMBUS+ OCR in 29 cell types and their pooled
678 OR through random effects meta-analysis using Gnocchi > 4 in (**E**). The blue dashed line
679 represents the OR of ENCODE cCRE-PLS (**E**). **F**, Enrichment of CAMBUS+ OCRs for
680 NaiveCD4 (n = 1,266) in the known cCREs (ENCODE cCRE-PLS; n = 40,891, ENCODE
681 cCRE-pELS; n = 172,027, ENCODE cCRE-dELS; n = 789,200, ENCODE CTCF-only; n =
682 35,839, FANTOM5 enhancers; n = 63,280, Super enhancers; n = 16,472), adjusted for
683 Gnocchi and phyloP. **G,** Enrichment of CAMBUS+ for NaiveCD4 (n = 1,266) in the tissue
684 expressed FANTOM5 enhancers (hematopoietic cells; n = 64,225, Brain; n = 62,978, Kidney;
685 n = 35,177, Lung; n = 26,630), and Super enhancers (Blood; n = 8,055, Brain; n = 9,372,
686 Kidney; n = 4,712, Lung; n = 9,878). The dashed line represents no enrichment (OR=1.0; **D,**
687 **E, F, G**). CAMBUS control in NaiveCD4 (n = 251,807) is used as background in (**F, G**). The
688 error bars represent 95% confidence intervals, with only lower bounds shown in C (See
689 Supplementary Table 4 for higher bounds).

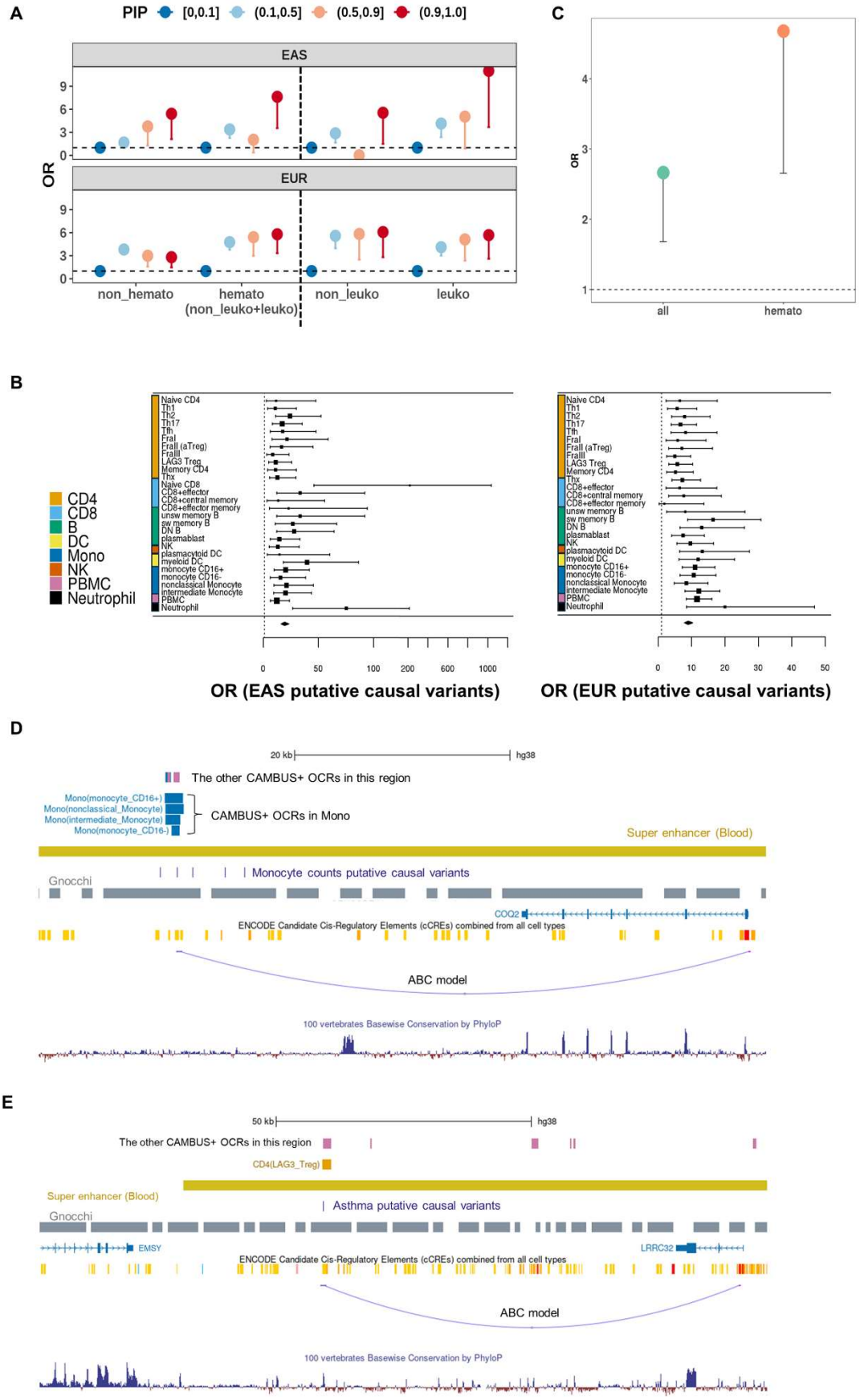
690



691

692 **Fig. 4 | Heritability enrichment for immune-mediated diseases. A,** The heritability
693 enrichment of various GWAS traits (in total, 33 GWAS for EUR and EAS ancestries;
694 Supplementary Table 13) in CAMBUS+ OCRs for each of 29 cell types. The heatmap colors
695 indicate -log10-scale p-value of the coefficient for CAMBUS+ annotation in S-LDSC
696 (coefficient P). '*' indicates significance (qvalue < 0.05) after FDR correction.

697

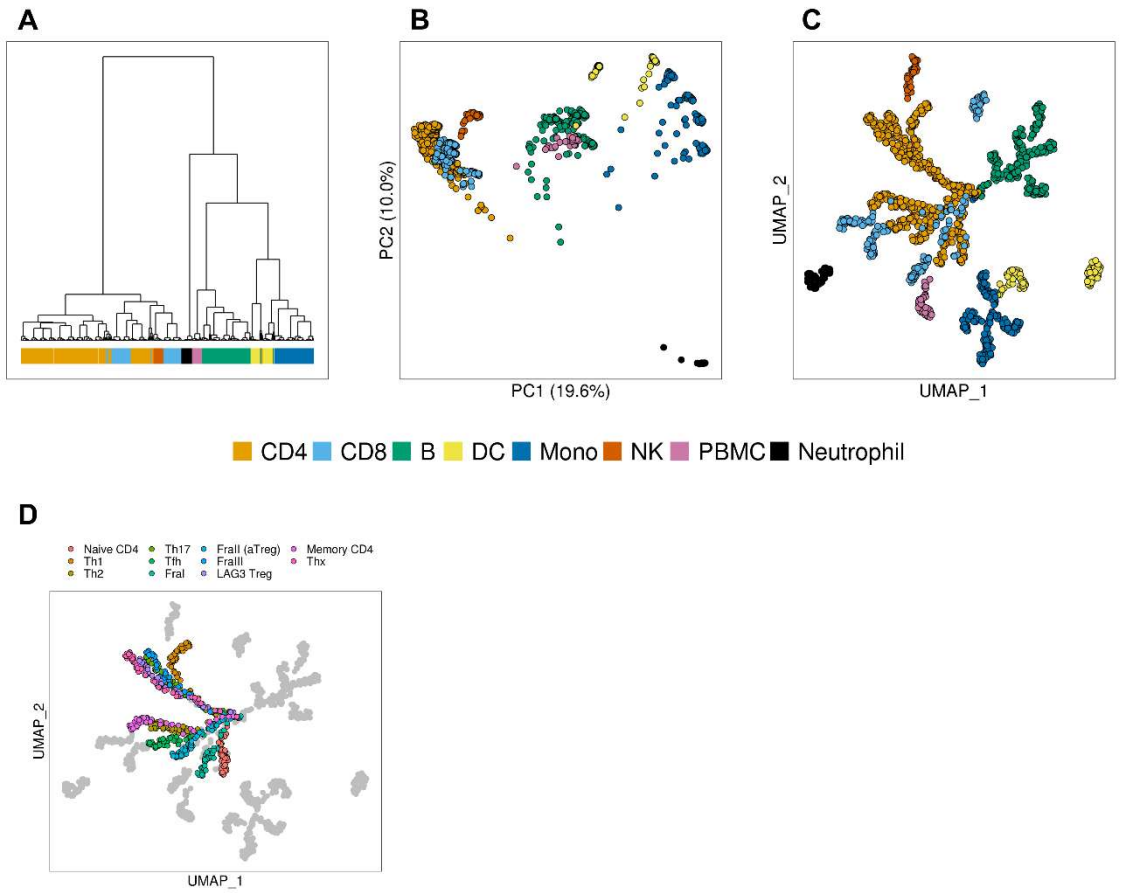

A
PIP [0,0.1] (0.1,0.5] (0.5,0.9] (0.9,1.0]
EAS
EUR
OR
non_hemato
hemato (non_leuko+leuko)
non_leuko
leuko
C
OR
all
hemato
B
CD4
CD8
B
DC
Mono
NK
PBMC
Neutrophil
OR (EAS putative causal variants)
OR (EUR putative causal variants)
D
20 kb
hg38
The other CAMBUS+ OCRs in this region
Mono(monocyte_CD16+)
Mono(nonclassical_Monocyte)
Mono(intermediate_Monocyte)
Mono(monocyte_CD16-)
CAMBUS+ OCRs in Mono
Super enhancer (Blood)
Monocyte counts putative causal variants
Gnocchi
COQ2
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ABC model
100 vertebrates Basewise Conservation by PhyloP
E
50 kb
hg38
The other CAMBUS+ OCRs in this region
CD4(LAG3_Treg)
Super enhancer (Blood)
Asthma putative causal variants
Gnocchi
EMSY
LRRC32
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ABC model
100 vertebrates Basewise Conservation by PhyloP

698

699 **Fig. 5 | Prioritization of putative causal variants using CAMBUS+ OCRs. A,** Enrichment
700 of putative causal variants in leukocyte (leuko), non-leukocytes (non_leuko), hemato (leuko
701 + non_leuko), and non-hematopoietic (non_hemato) traits in CAMBUS+ OCRs for Th1.
702 Putative causal variants were obtained from statistical fine-mapping analysis of many
703 complex traits' GWAS results for EAS and EUR populations [52,53]. **B,** Forest plot showing the
704 enrichment of CAMBUS+ OCRs across 29 cell types and the pooled OR in highly confident
705 putative causal variants (PIP > 0.9). Variants with PIP <= 0.1 were used as the background
706 for OR calculation. **C,** Enrichment of CAMBUS+ OCRs for any of 29 cell types in putative
707 causal rare variants (maf < 0.01 in EAS), evaluated by comparing the proportion of
708 CAMBUS+ OCRs overlapping with putative causal rare variants with PIP > 0.1. The left dot
709 represents the OR calculated using all the putative causal variants, while the right dot
710 represents the OR based only on the putative causal variants for hematopoietic phenotypes.
711 The dashed line represents no enrichment (OR=1.0). **D,E,** Representative examples of
712 variants associated with monocyte counts prioritized with CAMBUS+ in Mono (**D**) and asthma
713 prioritized with CAMBUS+ in regulatory T-cells (**E**) in putative causal variants in EUR
714 ancestry and surrounding genomic annotations. The navy boxes represent putative causal
715 variants in this credible set. The blue dashed line represents an enhancer-gene connection
716 to the nearest protein-coding gene predicted by the ABC model for $CD14^+$ monocytes (**D**)
717 and HUVEC (**E**). The gene with the highest ABC score is also shown in (**D**). All the other
718 CAMBUS+ OCRs around putative causal variants (rs13122305 (**D**), rs55646091 (**E**)) are also
719 shown in the figures (The purple box represents PBMC).

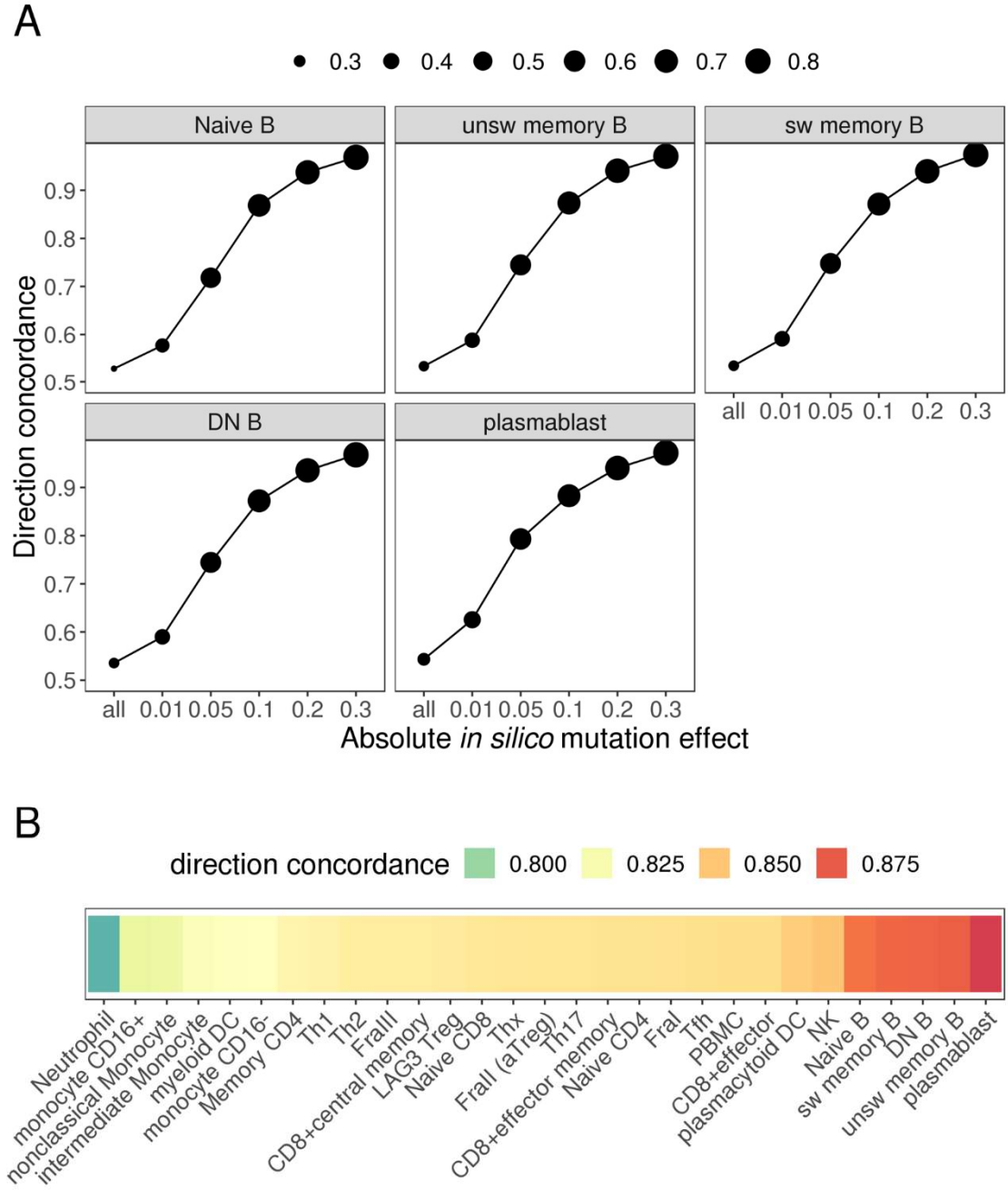
720

721 Extended Data Figure



722

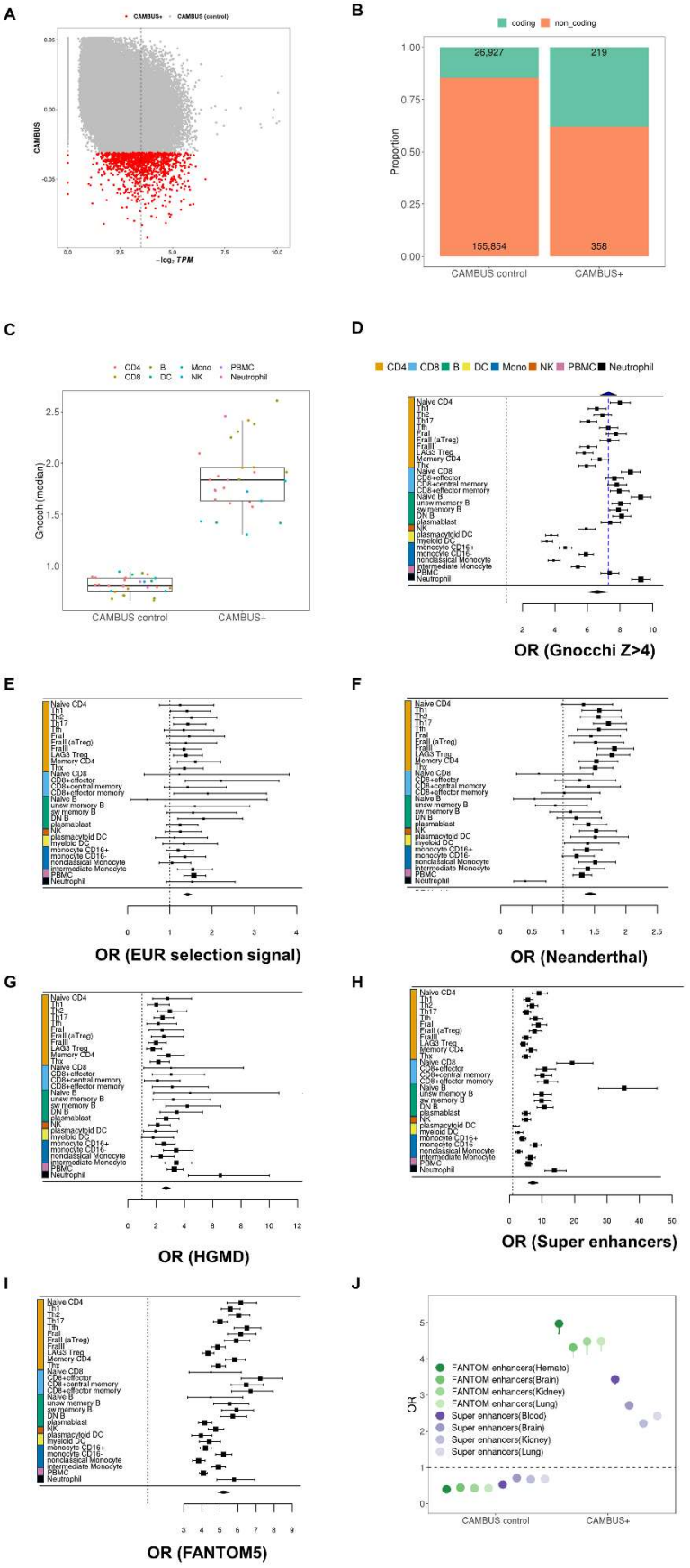
723 **Extended Data Fig. 1 | Clustering analysis and dimension reduction using PCA and**
724 **UMAP. A-C,** (**A**) Hierarchical clustering, (**B**) PCA, and (**C**) UMAP (**C**) of 1,386 samples. The
725 top 5,000 most accessible OCRs were used for these analyses. The colors indicate the
726 classification of samples into eight subgroups across 29 cell types. **D,** UMAP of 1,386
727 samples (**C**) colored by eleven types of CD4; other subgroups are colored with grey.



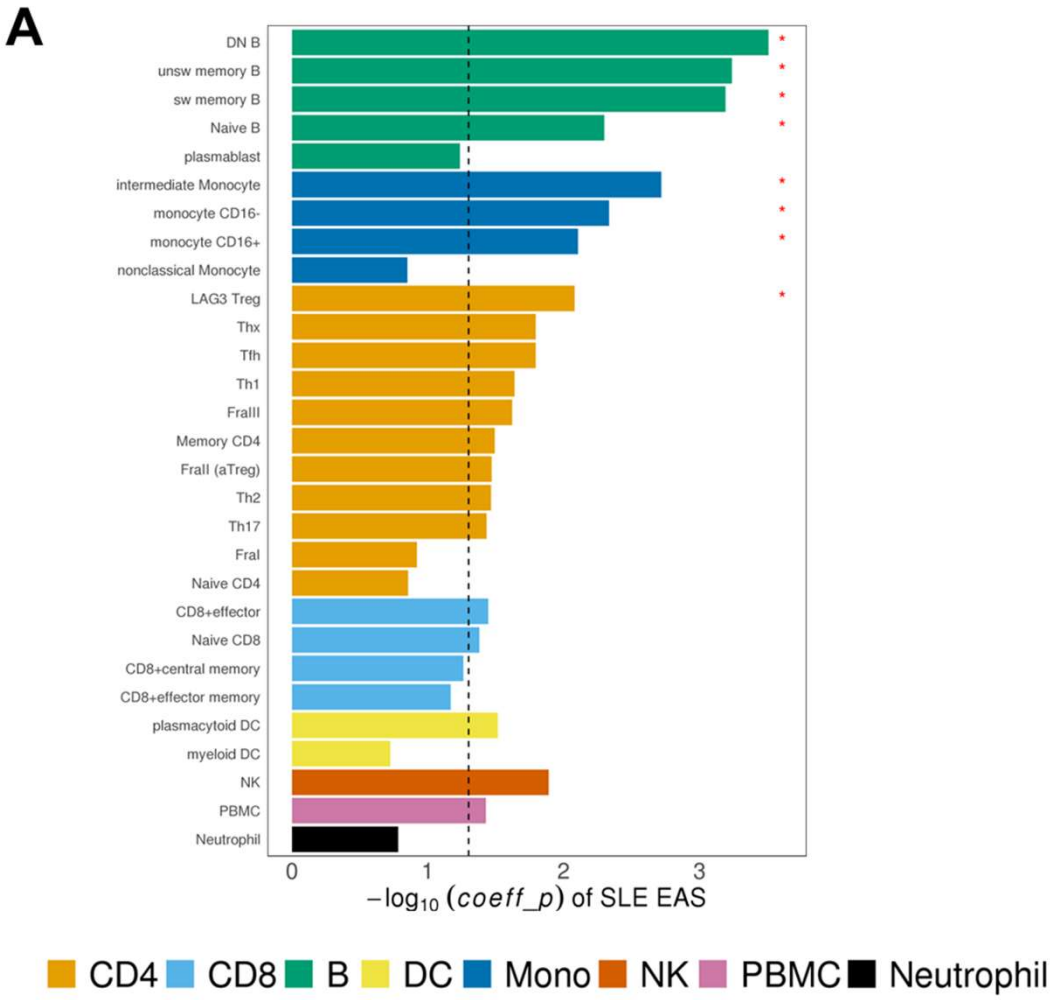
728

729 **Extended Data Fig. 2 | Accuracy of *in silico* mutation effect. A,** Comparison of direction
730 concordance between *in silico* mutation effects predicted by the CAMBUS ML models
731 (across five B-cell types) and effect sizes from an independent LCL caQTL study. The

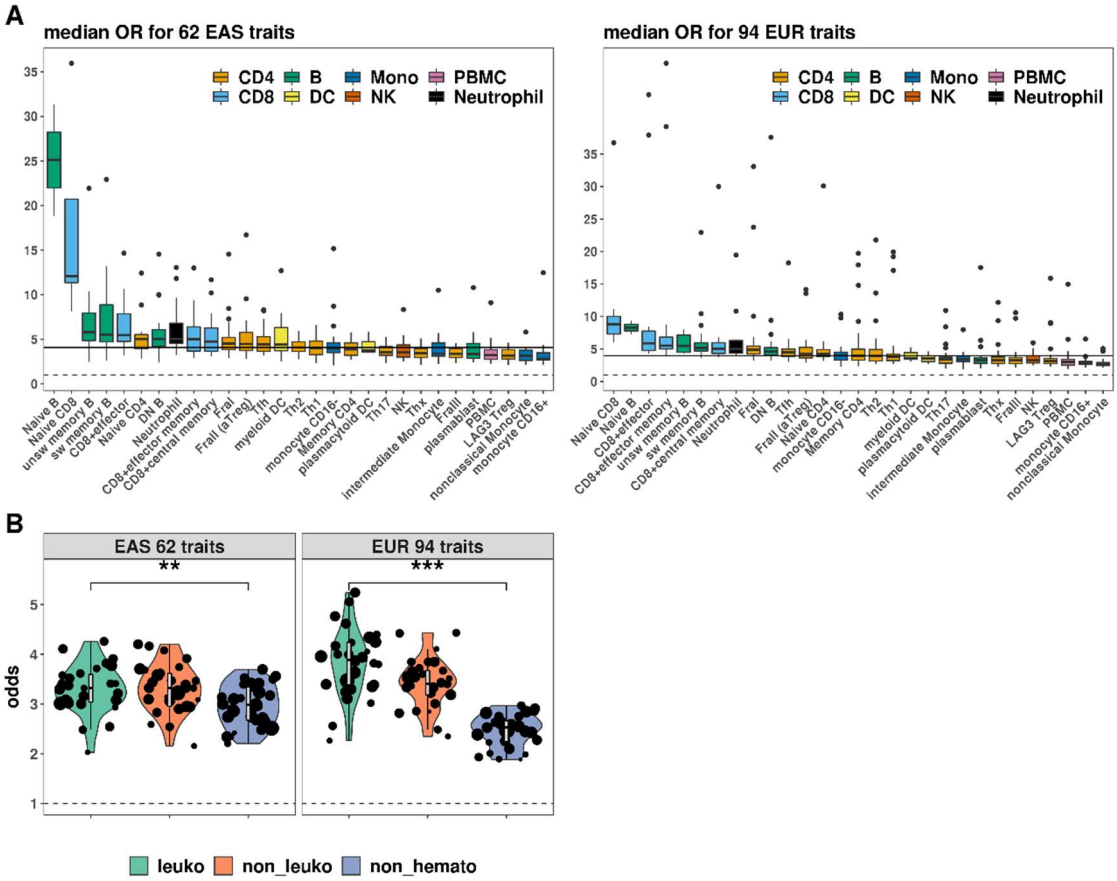
732 Pearson correlation coefficient (Pearson corr.) was calculated to assess the relationship
733 between *in silico* mutation effects and eQTL effect sizes. The x-axis represents a threshold
734 of *in silico* mutation effects to calculate direction concordance, where 'all' indicates no
735 threshold was applied. **B,** Comparison of direction concordance between *in silico* mutation
736 effects from the CAMBUS ML models and effect size from the LCL caQTL study. Absolute *in*
737 *silico* mutation effects greater than > 0.1 are used (see Fig. 2C).


A
CAMBUS+
CAMBUS (control)
CAMBUS
−log2 TPM
B
coding
non_coding
Proportion
26,927
219
155,854
358
CAMBUS control
CAMBUS+
C
CD4
CD8
B
DC
Mono
NK
PBMC
Neutrophil
Gnocchi(median)
CAMBUS control
CAMBUS+
D
CD4 CD8 B DC Mono NK PBMC Neutrophil
OR (Gnocchi Z>4)
E
OR (EUR selection signal)
F
OR (Neanderthal)
G
OR (HGMD)
H
OR (Super enhancers)
I
OR (FANTOM5)
J
OR
FANTOM enhancers(Hemato)
FANTOM enhancers(Brain)
FANTOM enhancers(Kidney)
FANTOM enhancers(Lung)
Super enhancers(Blood)
Super enhancers(Brain)
Super enhancers(Kidney)
Super enhancers(Lung)
CAMBUS control
CAMBUS+

739 **Extended Data Fig. 3 | CAMBUS+ can mark constraint regions and active enhancers.**
740 **A,** Scatter plot of median TPM and CAMBUS in Naïve CD4. The vertical dashed line
741 represents the top 5% of the median TPM. **B,** The number of OCRs in coding and non-coding
742 regions for each CAMBUS+ OCRs and CAMBUS control. The region of coding and non-
743 coding are referred to previous study [24]. **C,** The distribution of non-coding Gnocchi (median)
744 for each CAMBUS+ OCRs and CAMBUS control in each 29 cell type. **D,** OR of relaxed
745 CAMBUS+ OCRs across 29 cell types in Gnocchi> 4 and pooled OR estimated through
746 random effect meta-analysis. The blue dashed line represents the OR of ENCODE promoter-
747 like elements. **E-I,** OR and the pooled OR of the CAMBUS+ OCRs in UKB selection signals
748 (**E**), Neanderthal genome sequences derived from samples of EAS ancestry (**F**), pathogenic
749 regulatory variants (**G**), super-enhancers (**H**) and FANTOM5 enhancers (**I**). **J,** ORs of relaxed
750 CAMBUS+ (NaiveCD4; n = 12,736) in cell-type-specific regulatory elements: the tissue
751 expressed FANTOM5 enhancers (hematopoietic cells; n = 64,225, Brain; n = 62,978, Kidney;
752 n = 35,177, Lung; n = 26,630), and super-enhancers (Blood; n = 8,055, Brain; n = 9,372,
753 Kidney; n = 4,712, Lung; n = 9,878). The dashed line represents no enrichment (OR=1.0; **D-**
754 **J**)

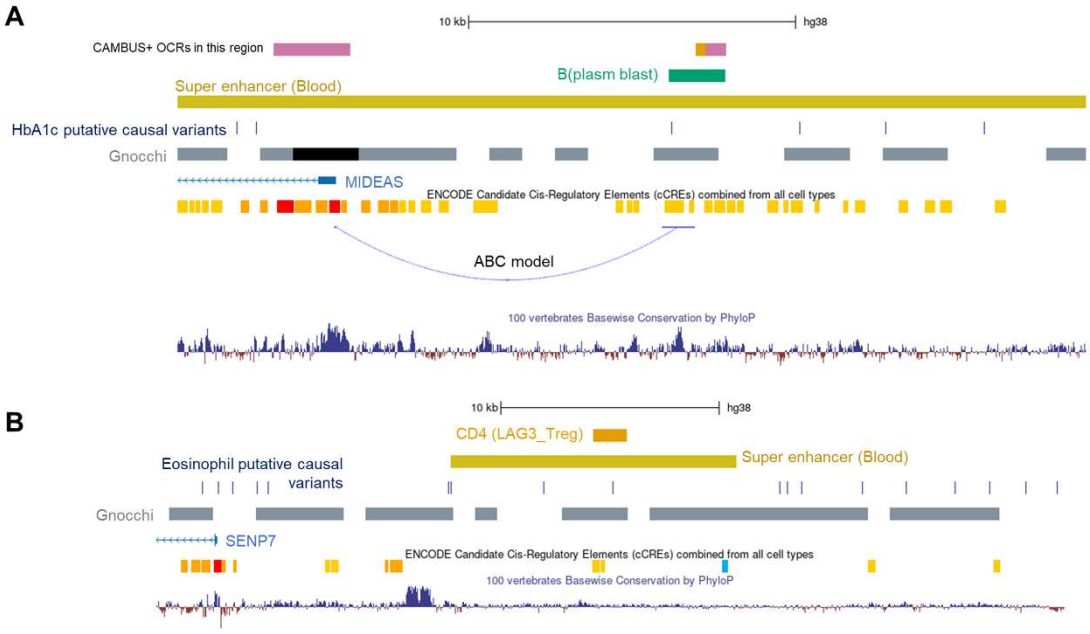


Extended Data Fig. 4 | Heritability enrichment for immune-mediated diseases. A, The -log10-scale coefficient P of CAMBUS+ OCRs in 29 cell types for SLE GWAS in EAS ancestry. The dashed line indicates a coefficient P = 0.05. '*' indicates significance (qvalue < 0.05) after FDR correction.



Extended Data Fig. 5 | Prioritization of putative causal variants using CAMBUS+ OCRs. **A,** Median OR of all variants in 95% credible sets in CAMBUS+ OCRs across 29 cell types. We used putative causal variants from statistical fine-mapping studies in EUR and EAS ancestries [52,53]. The solid line represents the median OR across 29 cell types. **B,** Enrichment of the putative causal variants in leukocyte (leuko), non-leukocytes (non_leuko), and non-hematopoietic (non_hemato) traits in CAMBUS+ OCRs across 29 cell types. Only cell types with P < 0.05 are shown (27 cell types in BBJ leuko; 28 in UKB non_hemato; 29 in BBJ non_leuko, BBJ non_hemato, UKB leuko, and UKB non_leuko), and dots indicate -log10-scale p-value. t-tests were performed between leuko and non_hemato traits, with significance levels denoted as P <0.05 (‘*’) and P < 0.001 (‘***’).

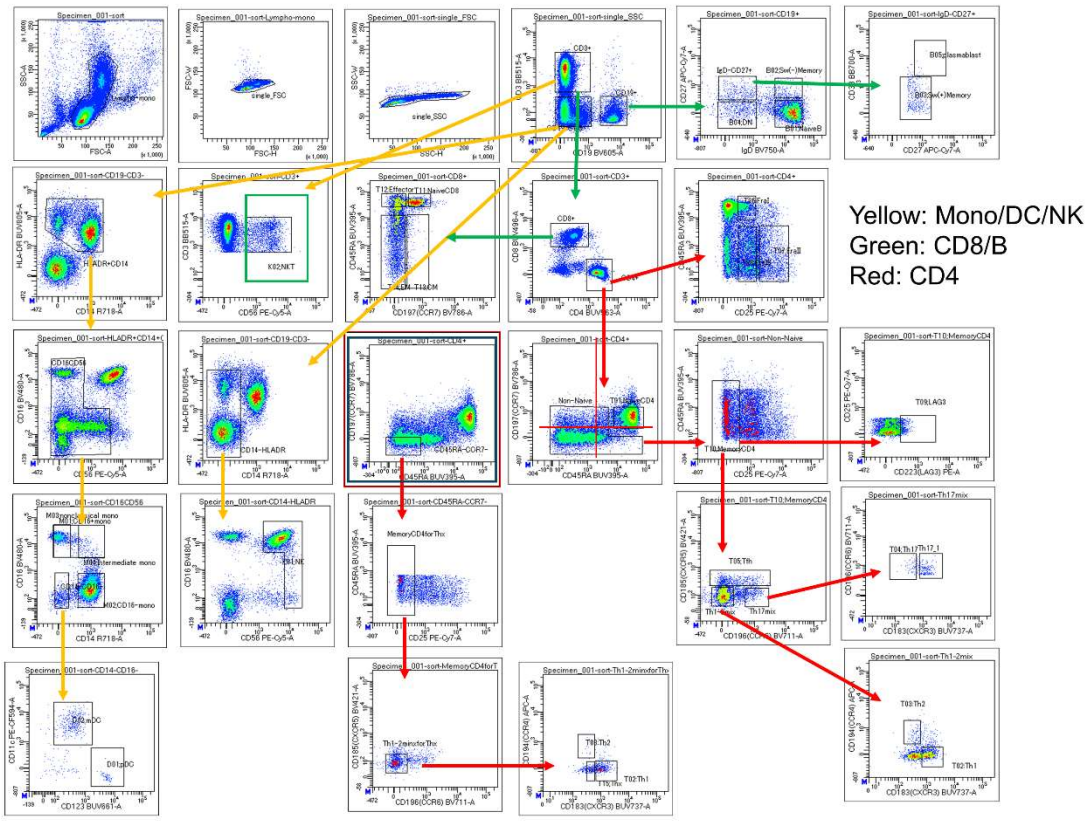
772



773

774 **Extended Data Fig. 6 | Prioritization of putative causal variants using CAMBUS+ OCRs.**
775 **A,B,** representative example of variants associated with Hemoglobin A1c level prioritized
776 with CAMBUS+ in plasma blast in putative causal variants in EUR ancestry (**A**) and
777 eosinophils prioritized with CAMBUS+ in regulatory T-cell (**B**) and their surrounding genomic
778 annotations. The navy boxes represent putative causal variants in the credible set. The blue
779 dashed line represents an enhancer-gene connection to the nearest protein-coding gene
780 predicted by the ABC model in B-cells (**A**). All the other CAMBUS+ OCRs around putative
781 causal variants (rs8009224 (**A**)) are also shown in the figures (The purple box represents
782 PBMC).

A



Extended Data Fig. 7 | Sorting strategy for immune cell subsets. A, Representative FACS plots describing gating strategies for 27 cell types. See Supplementary Table 1 for flow cytometry antibodies used in this study.

References

1 Pollen, A. A., Kilik, U., Lowe, C. B. & Camp, J. G. Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. *Nat Rev Genet* **24**, 687-711 (2023). https://doi.org:10.1038/s41576-022-00568-4

2 O'Bleness, M., Searles, V. B., Varki, A., Gagneux, P. & Sikela, J. M. Evolution of genetic and genomic features unique to the human lineage. *Nat Rev Genet* **13**, 853-866 (2012). https://doi.org:10.1038/nrg3336

3 Leypold, N. A. & Speicher, M. R. Evolutionary conservation in noncoding genomic regions. *Trends Genet* **37**, 903-918 (2021). https://doi.org:10.1016/j.tig.2021.06.007

4 Franchini, L. F. & Pollard, K. S. Human evolution: the non-coding revolution. *BMC Biol* **15**, 89 (2017). https://doi.org:10.1186/s12915-017-0428-9

5 Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675-1678 (2012). https://doi.org:10.1126/science.1225057

6 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014). https://doi.org:10.1038/nature12787

7 Soskic, B. *et al.* Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nat Genet* **51**, 1486-1493 (2019). https://doi.org:10.1038/s41588-019-0493-9

8 Gupta, A. *et al.* Dynamic regulatory elements in single-cell multimodal data implicate key immune cell states enriched for autoimmune disease heritability. *Nat Genet* **55**, 2200-2210 (2023). https://doi.org:10.1038/s41588-023-01577-7

9 Oguchi, A. *et al.* An atlas of transcribed enhancers across helper T cell diversity for decoding human diseases. *Science* **385**, eadd8394 (2024). https://doi.org:10.1126/science.add8394

10 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013). https://doi.org:10.1038/nmeth.2688

11 Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* **21**, 22 (2020). https://doi.org:10.1186/s13059-020-1929-3

12 Consortium, E. P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020). https://doi.org:10.1038/s41586-020-2493-4

13 Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001 e5919 (2021). https://doi.org:10.1016/j.cell.2021.10.024

14 Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322 (2008). https://doi.org:10.1016/j.cell.2007.12.014

15 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012). https://doi.org:10.1126/science.1222794

16 Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179-189 (2020). https://doi.org:10.1038/s41586-019-1879-7

17 Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* **45**, 124-130 (2013). https://doi.org:10.1038/ng.2504

18 Zeng, B. *et al.* Genetic regulation of cell type-specific chromatin accessibility shapes brain disease etiology. *Science* **384**, eadh4265 (2024). https://doi.org:10.1126/science.adh4265

19 Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017). https://doi.org:10.1016/j.ajhg.2017.06.005

20 Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018). https://doi.org:10.1038/s41576-018-0016-z

21 Thibodeau, A., Uyar, A., Khetan, S., Stitzel, M. L. & Ucar, D. A neural network based model effectively predicts enhancers from clinical ATAC-seq samples. *Sci Rep* **8**, 16048 (2018). https://doi.org:10.1038/s41598-018-34420-9

22 Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207-220 (2019). https://doi.org:10.1038/s41576-018-0089-8

23 Radke, D. W. *et al.* Purifying selection on noncoding deletions of human regulatory loci detected using their cellular pleiotropy. *Genome Res* **31**, 935-946 (2021). https://doi.org:10.1101/gr.275263.121

24 Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92-100 (2024). https://doi.org:10.1038/s41586-023-06045-0

25 Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* **48**, 206-213 (2016). https://doi.org:10.1038/ng.3467

26 Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat Genet* **50**, 1140-1150 (2018). https://doi.org:10.1038/s41588-018-0156-2

27 Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**, 1171-1179 (2018). https://doi.org:10.1038/s41588-018-0160-6

28 Eraslan, G., Avsec, Z., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20**, 389-403 (2019). https://doi.org:10.1038/s41576-019-0122-6

29 Koido, M. *et al.* Prediction of the cell-type-specific transcription of non-coding RNAs from genome sequences via machine learning. *Nat Biomed Eng* **7**, 830-844 (2023). https://doi.org:10.1038/s41551-022-00961-8

30 Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-934 (2015). https://doi.org:10.1038/nmeth.3547

31 Zhang, J. Y. *et al.* Single-cell landscape of immunological responses in patients with COVID-19. *Nat Immunol* **21**, 1107-1118 (2020). https://doi.org:10.1038/s41590-020-0762-x

32 Benaglio, P. *et al.* Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex immune trait variants using single nucleus ATAC-seq in peripheral blood. *PLoS Genet* **19**, e1010759 (2023). https://doi.org:10.1371/journal.pgen.1010759

33 Li, Y., Li, Z. & Hu, F. Double-negative (DN) B cells: an under-recognized effector memory B cell subset in autoimmunity. *Clin Exp Immunol* **205**, 119-127 (2021). https://doi.org:10.1111/cei.13615

34 Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat Methods* **19**, 1088-1096 (2022). https://doi.org:10.1038/s41592-022-01562-8

35 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121 (2010). https://doi.org:10.1101/gr.097857.109

36 Silvert, M., Quintana-Murci, L. & Rotival, M. Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. *Am J Hum Genet* **104**, 1241-1250 (2019). https://doi.org:10.1016/j.ajhg.2019.04.016

37 Nait Saada, J. *et al.* Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat Commun* **11**, 6130 (2020). https://doi.org:10.1038/s41467-020-19588-x

38 Liu, X. *et al.* Decoding triancestral origins, archaic introgression, and natural selection in the Japanese population by whole-genome sequencing. *Sci Adv* **10**, eadi8419 (2024). https://doi.org:10.1126/sciadv.adi8419

39 Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53-61 e59 (2018). https://doi.org:10.1016/j.cell.2018.02.031

40 Mangan, R. J. *et al.* Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**, 4587-4603 e4523 (2022). https://doi.org:10.1016/j.cell.2022.10.016

41 Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577-581 (2003). https://doi.org:10.1002/humu.10212

42 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013). https://doi.org:10.1016/j.cell.2013.09.053

43 Jiang, Y. *et al.* SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res* **47**, D235-D243 (2019). https://doi.org:10.1093/nar/gky1025

44 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-1235 (2015). https://doi.org:10.1038/ng.3404

45 Tashman, K. C., Cui, R., O'Connor, L. J., Neale, B. M. & Finucane, H. K. Significance testing for small annotations in stratified LD-Score regression. *medRxiv*, 2021.2003.2013.21249938 (2021). https://doi.org:10.1101/2021.03.13.21249938

46 Moser, T., Akgun, K., Proschmann, U., Sellner, J. & Ziemssen, T. The role of TH17 cells in multiple sclerosis: Therapeutic implications. *Autoimmun Rev* **19**, 102647 (2020). https://doi.org:10.1016/j.autrev.2020.102647

47 Jarvinen, T. M. *et al.* Replication of GWAS-identified systemic lupus erythematosus susceptibility genes affirms B-cell receptor pathway signalling and strengthens the role of IRF5 in disease susceptibility in a Northern European population. *Rheumatology (Oxford)* **51**, 87-92 (2012). https://doi.org:10.1093/rheumatology/ker263

48 Fillatreau, S., Manfroi, B. & Dorner, T. Toll-like receptor signalling in B cells during systemic lupus erythematosus. *Nat Rev Rheumatol* **17**, 98-108 (2021). https://doi.org:10.1038/s41584-020-00544-4

49 Honda, T. & Kabashima, K. Reconciling innate and acquired immunity in atopic dermatitis. *J Allergy Clin Immunol* **145**, 1136-1137 (2020). https://doi.org:10.1016/j.jaci.2020.02.008

50 Chung, M. K. Y. *et al.* Functions of double-negative B cells in autoimmune diseases, infections, and cancers. *EMBO Mol Med* **15**, e17341 (2023). https://doi.org:10.15252/emmm.202217341

51 Yin, X. *et al.* Meta-analysis of 208370 East Asians identifies 113 susceptibility loci for systemic lupus erythematosus. *Ann Rheum Dis* **80**, 632-640 (2021). https://doi.org:10.1136/annrheumdis-2020-219209

52 Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv*, 2021.2009.2003.21262975 (2021). https://doi.org:10.1101/2021.09.03.21262975

53 Koyama, S. *et al.* Population-specific putative causal variants shape quantitative traits. *Nature Genetics* (2024). https://doi.org:10.1038/s41588-024-01913-5

54 Armanfar, M., Jafari, A., Dehghan, G. R. & Abdizadeh, L. Effect of coenzyme Q10 supplementation on exercise-induced response of inflammatory indicators and blood lactate in male runners. *Med J Islam Repub Iran* **29**, 202 (2015).

55 Mantle, D., Heaton, R. A. & Hargreaves, I. P. Coenzyme Q10 and Immune Function: An Overview. *Antioxidants (Basel)* **10** (2021). https://doi.org:10.3390/antiox10050759

56 Mukherjee, R. *et al.* Non-Classical monocytes display inflammatory features: Validation in Sepsis and Systemic Lupus Erythematous. *Sci Rep* **5**, 13886 (2015). https://doi.org:10.1038/srep13886

57 Nassir, N. *et al.* Single-cell transcriptome identifies FCGR3B upregulated subtype of alveolar macrophages in patients with critical COVID-19. *iScience* **24**, 103030 (2021). https://doi.org:10.1016/j.isci.2021.103030

58 Li, Y. K. *et al.* Single-nucleotide polymorphism rs6592645 confers asthma risk through regulating LRRC32 expression. *Exp Ther Med* **26**, 451 (2023). https://doi.org:10.3892/etm.2023.12150

59 Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664-1669 (2019). https://doi.org:10.1038/s41588-019-0538-0

60 Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P. R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet* **53**, 1260-1269 (2021). https://doi.org:10.1038/s41588-021-00892-1

61 Kuderna, L. F. K. *et al.* Identification of constrained sequence elements across 239 primate genomes. *Nature* **625**, 735-742 (2024). https://doi.org:10.1038/s41586-023-06798-8

62 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015). https://doi.org:10.1038/nature14248

63 Martens, L. D., Fischer, D. S., Yepez, V. A., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat Methods* **21**, 28-31 (2024). https://doi.org:10.1038/s41592-023-02112-6

64 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020). https://doi.org:10.1038/s41586-020-2308-7

65 Nelson, R. K., Bush, A., Stokes, J., Nair, P. & Akuthota, P. Eosinophilic Asthma. *J Allergy Clin Immunol Pract* **8**, 465-473 (2020). https://doi.org:10.1016/j.jaip.2019.11.024

66 Harb, H. & Chatila, T. A. Regulatory T-cells in asthma. *Curr Opin Allergy Clin Immunol* **23**, 151-157 (2023). https://doi.org:10.1097/ACI.0000000000000887

67 Tran, D. Q. *et al.* GARP (LRRC32) is essential for the surface expression of latent TGF-beta on platelets and activated FOXP3+ regulatory T cells. *Proc Natl Acad Sci U S A* **106**, 13445-13450 (2009). https://doi.org:10.1073/pnas.0901944106

68 Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**, 959-962 (2017). https://doi.org:10.1038/nmeth.4396

69 Li, R. & Quon, G. scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biol* **20**, 193 (2019). https://doi.org:10.1186/s13059-019-1806-0

70 Storey, J. D. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**, 479-498 (2002). https://doi.org:10.1111/1467-9868.00346

71 Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat Genet* **54**, 1640-1651 (2022). https://doi.org:10.1038/s41588-022-01213-w

72 Ishikawa, Y. *et al.* GWAS for systemic sclerosis identifies six novel susceptibility loci including one in the Fcgamma receptor region. *Nat Commun* **15**, 319 (2024). https://doi.org:10.1038/s41467-023-44541-z

73 Tanaka, N. *et al.* Eight novel susceptibility loci and putative causal variants in atopic dermatitis. *J Allergy Clin Immunol* **148**, 1293-1306 (2021). https://doi.org:10.1016/j.jaci.2021.04.019

74 Terao, C. *et al.* Genetic determinants and an epistasis of LILRA3 and HLA-B*52 in Takayasu arteritis. *Proc Natl Acad Sci U S A* **115**, 13045-13050 (2018). https://doi.org:10.1073/pnas.1808850115