IMPACT: Genomic annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors

Tiffany Amariuta[1,2,3,4,5], Yang Luo[1,2,3], Steven Gazal[3,6], Emma E. Davenport[1,2,3], Bryce van de Geijn[3,6], Harm-Jan Westra[1,2,3,7], Nikola Teslovich[2], Yukinori Okada[8,9], Kazuhiko Yamamoto[10], RACI consortium[†], GARNET consortium[†], Alkes Price[3,6,11], Soumya Raychaudhuri[1,2,3,4,5,12]

[1]Center for Data Sciences, Harvard Medical School, Boston, Massachusetts, USA.
[2]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA.
[5]Graduate School of Arts and Sciences, Harvard University, Boston, Massachusetts, USA.
[6]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
[7]Faculty of Medical Sciences, University of Groningen, Groningen, Netherlands.
[8]Division of Medicine, Osaka University, Osaka, Japan.
[9]Osaka University Graduate School of Medicine, Osaka, Japan.
[10]Immunology Frontier Research Center (WPI-IFReC), Osaka University, Osaka, Japan.
[11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
[12]Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK.

*Correspondence to:
Alkes Price
665 Huntington Ave, Harvard T.H. Chan School of Public Health, Building 2, Room 211
Boston, MA 02115, USA.
aprice@hsph.harvard.edu; 617-432-2262 (tel); 617-432-1722 (fax)

Soumya Raychaudhuri
77 Avenue Louis Pasteur, Harvard New Research Building, Suite 250D
Boston, MA 02115, USA.
soumya@broadinstitute.org; 617-525-4484 (tel); 617-525-4488 (fax)

[†]Author information listed in Supplementary Note

Despite significant progress in annotating the genome with experimental methods, much of the regulatory noncoding genome remains poorly defined. Here we assert that regulatory elements may be characterized by leveraging local epigenomic signatures at sites where specific transcription factors (TFs) are bound. To link these two identifying features, we introduce IMPACT, a genome annotation strategy which identifies regulatory elements defined by cell-state-specific TF binding profiles, learned from 515 chromatin and sequence annotations. We validate IMPACT using multiple compelling applications. First, IMPACT predicts TF motif binding with high accuracy (average AUC 0.92, s.e. 0.03; across 8 TFs), a significant improvement (all p<6.9e-15) over intersecting motifs with open chromatin (average AUC 0.66, s.e. 0.11). Second, an IMPACT annotation trained on RNA polymerase II is more enriched for peripheral blood cis-eQTL variation (N=3,754) than sequence based annotations, such as promoters and regions around the TSS, (permutation p<1e-3, 25% average increase in enrichment). Third, integration with rheumatoid arthritis (RA) summary statistics from European (N=38,242) and East Asian (N=22,515) populations revealed that the top 5% of CD4+ Treg IMPACT regulatory elements capture 85.7% (s.e. 19.4%) of RA h2 (p<1.6e-5) and that the top 9.8% of Treg IMPACT regulatory elements, consisting of all SNPs with a non-zero annotation value, capture 97.3% (s.e. 18.2%) of RA h2 (p<7.6e-7), the most comprehensive explanation for RA h2 to date. In comparison, the average RA h2 captured by compared CD4+ T histone marks is 42.3% and by CD4+ T specifically expressed gene sets is 36.4%. Finally, integration with RA fine-mapping data (N=27,345) revealed a significant enrichment (2.87, p<8.6e-3) of putatively causal variants across 20 RA associated loci in the top 1% of CD4+ Treg IMPACT regulatory regions. Overall, we find that IMPACT generalizes well to other cell types in identifying complex trait associated regulatory elements.

Transcriptional regulation is the foundation for many complex biological phenotypes, from gene expression to disease susceptibility. However, the complexity of gene regulation, controlled by more than 1,600 human transcription factors (TFs)[1] influencing some 20,000 protein coding gene promoters, has made functional annotation of the

79 regulome difficult. Tens of thousands of genomic annotations have been experimentally
80 generated, enabling the success of unsupervised methods such as chromHMM[2] and
81 Segway[3] to identify global chromatin patterns that better characterize genomic function.
82 However, linking specific regulatory processes to these identified patterns is
83 challenging. Furthermore, although genome-wide association studies (GWAS) have
84 identified ~10,000 trait associated variants across hundreds of polygenic traits[4], most
85 variants lie in noncoding regulatory regions with uncertain function.
86
87 With continually increasing numbers of genomic annotations generated from high-
88 throughput experimental assays, *in-silico* functional characterization of variants has
89 growing potential. These assays include genome-wide open chromatin, histone mark,
90 and RNA expression profiling, each separately possible at the single cell level. Initially
91 contributed by genomic consortia, such as ENCODE[5] and Roadmap[6], these assays
92 have become more common place as easy-to-implement protocols have been
93 developed, thereby contributing to the growing rate of genomic annotation generation.
94
95 Recently, integration of datasets, particularly those indicating regulatory elements, with
96 GWAS data has successfully led to the identification of categories of disease-driving
97 variants enriched for genetic heritability (h2)[7–9]. Such regulatory annotations identify
98 active promoters and enhancers through open chromatin or histone mark occupancy
99 assays in a cell type of interest[7,8,10–13]. However, these annotations include both cell-
100 type-specific and nonspecific elements, the latter of which may affect a wide range of
101 cellular functions that are not necessarily intrinsic to disease-driving cell-states.

102 Therefore, we hypothesized that the identification of regulatory elements specifically
103 driving functional states would help us to not only better characterize regulatory
104 elements genome-wide, but also better capture polygenic h2 of complex traits and
105 diseases. Once the most enriched classes of regulatory elements are recognized, then
106 it may become possible to generate biologically-founded mechanistic hypotheses.
107
108 Here, we introduce IMPACT (Inference and Modeling of Phenotype-related ACtive
109 Transcription), a diversely applicable genome annotation strategy to predict cell-state-
110 specific regulatory elements. We take a two-step approach to define IMPACT regulatory
111 elements. First, we choose a single key TF, known to regulate a cell-state-specific
112 process, and then identify binding motifs genome-wide, distinguishing between those
113 that are bound and unbound using genomic occupancy identified by ChIP-seq in the
114 corresponding cell-state. Second, IMPACT predicts TF occupancy at binding motifs by
115 aggregating and performing feature selection on 503 cell-type-specific epigenomic
116 features and 12 sequence features in an elastic net logistic regression model (**Methods***:*
117 *IMPACT Model*). Epigenomic features include histone mark ChIP-seq, ATAC-seq,
118 DNase-seq and HiChIP (**Table S1**) assayed in hematopoietic, adrenal, brain,
119 cardiovascular, gastrointestinal, skeletal, and other cell types, while sequence features
120 include coding, intergenic, etc. The IMPACT model framework can easily be expanded
121 to accommodate thousands of epigenomic feature annotations and is amenable to the
122 increasing rates of data generation. From this regression we learn a TF binding
123 chromatin profile, which IMPACT uses to probabilistically annotate the genome at

124 nucleotide-resolution. We refer to high scoring regions as cell-state-specific regulatory
125 elements (**Figure 1A**).
126
127 The IMPACT model assumes that binding sites of a given TF may be characterized by a
128 consensus epigenomic signature, which may also characterize binding sites of other
129 related TFs. If this signature is robust, each TF binding site genome-wide should share
130 some of its characteristics. Therefore, IMPACT should be able to predict with high
131 accuracy genome-wide TF occupancy of a motif, which has been challenging[14]. To test
132 this model assumption, we used IMPACT to predict regulatory elements based on
133 experimental binding identified via ChIP-seq of eight TFs assayed in eight different cell-
134 states: T-BET, GATA3, STAT3, FOXP3, REST, HNF4A, TCF7L2, and RNA Polymerase
135 (Pol) II in CD4+ Th(T helper)1, CD4+ Th2, CD4+ Th17, CD4+ Treg (T regulatory), fetal
136 brain, liver, pancreatic, and lymphocytic cells, respectively (**Table S2**)[5,15–20] (**Methods***:*
137 *IMPACT Model*). IMPACT predicts TF occupancy at binding motifs with high accuracy
138 (average AUC across 8 TFs 0.92 (s.e. 0.03), 10-fold cross validation on 80% of data,
139 AUC evaluated on the withheld 20%, **Figure 1B**, **Figure S1, Table S3**). In comparison,
140 IMPACT significantly outperforms an approach that does not use data aggregation or a
141 machine learning framework to predict TF binding (all p<6.9e-15); this approach
142 predicts binding on motifs that overlap cell-type-specific active promoters (H3K4me3
143 ChIP-seq) or cell-type-specific open chromatin (DNase-seq) (average AUC 0.66 (s.e.
144 0.11), **Figure 1B**). We anecdotally observe that IMPACT predictions at canonical CD4+
145 T cell TF-specific targets (**Figure 1C**, **Figures S2-4**) substantially vary within a single

146 ChIP-seq peak, suggesting that chromatin signatures exist on a much smaller scale and
147 that IMPACT may be more sensitive as a regulatory annotation than TF ChIP-seq.
148
149 We developed IMPACT to model regulation specific to a functional cell-state, the most
150 general of which may be active cellular transcription. Expression quantitative trait loci
151 (eQTLs) are genetic variations that modulate transcription[21]. Most cis eQTLs map to
152 TSS (transcription start site) and promoter annotations, and more rarely to the 5' UTR
153 (untranslated region)[22]. We hypothesized that an IMPACT annotation tracking active
154 transcription, trained on RNA Polymerase (Pol) II binding sites, would capture cis eQTL
155 causal variation better than the most strongly enriched canonical annotations.
156
157 We obtained SNP-level summary statistics from a large and previously published eQTL
158 analysis on 3,754 peripheral blood samples[23]. We then used IMPACT to annotate SNPs
159 tested in the eQTL analysis with RNA Pol II specific regulatory element probabilities,
160 learned from Pol II binding profiles in T cells and B cells, the predominant cell
161 populations of peripheral blood. Next, we computed a genome-wide enrichment
162 (**Methods**: *eQTL enrichment*) of chi-squared cis eQTL association statistics, averaged
163 over all genes, across Pol II IMPACT and several sequenced-based annotations, such
164 as TSS windows, promoters, and enhancers, separately, revealing that Pol II IMPACT
165 has the strongest chi-squared enrichment (1.7, permutation $p<1e-3$, 25% average
166 increase in enrichment) (**Figure 2**). These results argue that Pol II IMPACT regions
167 better localize active promoter and proximal regulatory regions driving eQTLs than the
168 compared canonical genomic annotations, which may be less specific due to their larger

169 sizes and restrictive binary characterization. This suggests that IMPACT may be more
170 effective at prioritizing causal SNP variation when fine-mapping eQTLs. These results
171 also argue that the biological basis of eQTLs are related to Pol II regions, which is a
172 refinement over previous observations that eQTL causal variation is concentrated near
173 and around TSS and promoter regions.
174
175 We previously hypothesized that IMPACT annotations of pathogenic cell-states would
176 more precisely capture polygenic trait h2, compared to regulatory annotations that don't
177 resolve cell-states. Testing this hypothesis requires a polygenic trait with a well-studied
178 disease-driving cell type. Genetic studies of rheumatoid arthritis (RA), an autoimmune
179 disease that attacks synovial joint tissue leading to permanent joint damage and
180 disability[24], have suggested a critical role by CD4+ T cells[7,8,11,12,25–29]. However, CD4+ T
181 cells are extremely heterogeneous: naive CD4+ T cells may differentiate into memory T
182 cells, and then into effector T cells including Th1, Th2, and Th17 and T regulatory cells,
183 requiring the action of a limited number of key transcription factors (TFs): T-BET or
184 STAT4, GATA3 or STAT6, STAT3 or RORγt, FOXP3 or STAT5, respectively[30]. As
185 these CD4+ T effector cell-states contribute to RA risk[7,11,28], we hypothesized that CD4+
186 T cell-state-specific IMPACT regulatory element annotations would better capture RA
187 h2 than annotations that generalize CD4+ T cells and ignore the differential functionality
188 of effector cell-states. Therefore, we built IMPACT annotations in four CD4+ T cell-
189 states, Th1, Th2, Th17, and Treg. We then integrated S-LDSC (stratified linkage
190 disequilibrium (LD) score regression)[7] with publicly available European (EUR, N =
191 38,242)[7,31] and East Asian (EAS, N = 22,515)[32] RA GWAS summary statistics to

192 partition the common SNP h2 of RA. While partitioning h2 with IMPACT annotations, we
193 use a set of 69 baseline annotations, a subset of the 75 annotations referred to as the
194 baseline-LD model[9], specifically excluding T cell-related annotations, to control for cell-
195 type-nonspecific functional, LD-related, and minor allele frequency (MAF) associations.
196 We use two metrics to assess the contribution of our IMPACT annotation to RA h2:
197 enrichment and per-annotation standardized effect size, τ^*. Enrichment is defined as
198 the proportion of h2 divided by the genome-wide proportion of SNPs in the annotation,
199 and τ^* is defined as the proportionate change in per-SNP h2 associated with a one
200 standard deviation increase in the value of the annotation[9]. τ^* captures the unique
201 contribution of an annotation to the h2 model, conditional on other present annotations.
202 The τ^* of an annotation that successfully captures trait h2 will be significantly greater
203 than zero; the greater the τ^*, the more h2 that the annotation captures.
204
205 We found that each CD4+ T cell-state-specific IMPACT annotation is significantly
206 enriched with RA h2 in both European and East Asian populations (average enrichment
207 = 20.05, all *p*<1.9e-04, **Figure 3A**, **Table S4**). We estimated total genome-wide
208 polygenic RA h2 to be about 18% for EUR and 21% for EAS (**Methods**: *S-LDSC*). As
209 RA is primarily driven by CD4+ T cells, we expectedly find that τ^* is significantly positive
210 for all CD4+ T IMPACT annotations separately conditioned on the cell-type-nonspecific
211 baselineLD annotations (all *p*<2.1e-03, **Figure 3B, Figure S5**). We then selected the
212 top 5% of regulatory SNPs according to each CD4+ T IMPACT annotation and find that
213 the Treg annotation explains the greatest proportion of RA h2, 85.7% (s.e. 19.4%,
214 enrichment p<1.6e-5) meta-analyzed between both EUR and EAS populations (**Figure**

215 **3C**). Other CD4+ T cell-state annotations explain similarly large proportions of RA h2.
216 Furthermore, we observe that the top 9.8% of CD4+ Treg IMPACT regulatory elements,
217 consisting of all SNPs with a non-zero annotation value, capture 97.3% (s.e. 18.2%,
218 enrichment p<7.6e-7) of RA h2 in EUR. This powerful result is the most comprehensive
219 explanation for RA h2, to our knowledge, to date.
220
221 We then compared the enrichments and τ^* of each CD4+ T IMPACT annotation to that
222 of various T cell functional annotations, using the EUR RA summary statistics: TF
223 binding motifs, genome-wide TF occupancy (ChIP-seq), an annotation that assigns
224 each SNP a value proportional to the number of IMPACT epigenomic features it
225 overlaps (Averaged Tracks), the five largest τ^* CD4+ T cell-specific histone mark
226 annotations[7], the five largest τ^* CD4+ T cell-specifically expressed gene sets and their
227 regulatory elements[8], and CD4+ T cell super enhancers[33]. While the CD4+ Treg
228 IMPACT annotation has similar enrichment to CD4+ T histone marks (22.9x compared
229 to 23.4x on average, respectively), we observe that the Treg IMPACT annotation is
230 substantially more enriched than CD4+ T specifically expressed gene sets (22.9x
231 compared to 2.9x on average, respectively) (**Figure 4A, Figure S6**). We note that the
232 average RA h2 captured by these CD4+ T histone marks, ranging in size from 1-3% of
233 SNPs, is 42.3%. The average RA h2 captured by these CD4+ T specifically expressed
234 gene sets, ranging in size from 11-13% of SNPs, is 36.4%. This is a large contrast to
235 the 85.7% of RA h2 captured by the top 5% of SNPs in the Treg IMPACT annotation.
236 Then, we computed the annotation τ^* while pairwise conditioned on each other and on
237 the baseline-LD. We found that the CD4+ Treg and Th2 IMPACT τ^* are significantly

238 positive (all Treg τ^* > 1.9, all Th2 τ^* > 1.7; all Treg τ^* p<5.0e-3, all Th2 τ^* p<0.01) and
239 more significant, thereby capturing RA h2 better, than all other T cell annotations,
240 except H3K27ac in Th2 cells (**Figure 4B, Figure S6**). Previous work has reported that
241 the threshold for impactful values of $|\tau^*|$ is approximately 0.24[9]. Overall, these results
242 suggest that IMPACT is identifying areas of concentrated RA h2 that other T cell
243 regulatory annotations have not.
244
245 We next applied our CD4+ T IMPACT annotations to 41 additional polygenic traits[9,34,35]
246 and observed consistently significantly positive τ^* for immune-mediated traits, such as
247 Crohn's, "all autoimmune disease", respiratory ear/nose/throat, and "allergy and
248 eczema" (mean τ^* = 3.2; all p<5.9e-4, p<1.9e-5, p<3.6e-3, p<1.7e-3, respectively), and
249 several blood traits, eosinophil and white blood cell counts (mean τ^* = 2.5; all p<1.6e-
250 11, p<0.02, respectively), but not for non-immune-mediated traits (**Figure 5, Table S5**).
251 For comparison, we created several other IMPACT annotations in different cell types
252 targeting h2 in a range of traits and highlight a few examples. For a liver IMPACT
253 annotation, trained on HNF4A[5] (hepatocyte nuclear factor 4A), τ^* is positive for LDL and
254 HDL (mean τ^* = 2.0; p<0.02, p<1.2e-3, respectively), liver-associated traits[20,36]. For a
255 macrophage IMPACT annotation, trained on IRF5[37], τ^* is positive for some immune-
256 mediated and blood traits (mean τ^* = 2.8, all p<8.2e-3) and intriguingly also for
257 schizophrenia (τ^* = 0.9, p<4.9e-5), as the literature supports a putative MHC
258 association[38]. Finally, for a CD4+ Treg IMPACT annotation, trained on STAT5[19], an
259 alternative key TF for Tregs, the values of τ^* across all traits resemble that of FOXP3.
260 This argues that the ability of IMPACT to capture polygenic h2 does not depend on the

261 choice of key regulator. Overall, this suggests that IMPACT is a promising strategy to
262 identify complex trait associated regulatory elements across a range of cell-states.
263
264 As we have demonstrated the ability of IMPACT to annotate the noncoding genome,
265 validated by captured polygenic h2, we hypothesized that this improved genomic
266 annotation might inform functional variant fine-mapping. Using a GWAS of 11,475
267 European RA cases and 15,870 controls[39], an independent study from the European
268 summary statistics used in our h2 analyses, our group recently fine-mapped a subset of
269 20 RA risk loci, each with a manageable number of putatively causal variants, and
270 created 90% credible sets of these SNPs[40]. We computed the enrichment of fine-
271 mapped causal probabilities across these 20 loci in the top 1% of our CD4+ T cell-state-
272 specific IMPACT annotations (**Methods**: *Posterior Probability Enrichment*). We found
273 that the Treg annotation is significantly enriched (2.87, permutation p<8.6e-03, **Figure**
274 **6A, Table S6**), while other annotations are not. The Treg IMPACT annotation may thus
275 be useful to prune putatively causal RA variants. Furthermore, we observe uniquely
276 high Treg enrichment in the *BACH2* and *IRF5* loci (16.2 and 8.1, respectively),
277 suggesting putatively causal SNPs in these loci may function in a Treg-specific context.
278
279 In the same study, our group observed both differential binding of CD4+ T nuclear
280 extract via EMSA and differential enhancer activity via luciferase assays at two credible
281 set SNPs, narrowing down the list of putatively causal variants in the *CD28/CTLA4* and
282 *TNFAIP3* loci[40]. We observed that both variants with functional activity were located at
283 predicted IMPACT regulatory elements, suggesting that IMPACT may be used to

284 narrow down credible sets to reduce the amount of experimental follow up. First, at the
285 *CD28/CTLA4* locus, IMPACT predicts high probability regulatory elements across the
286 four CD4+ T cell-states at the functional SNP rs117701653 and lower probability
287 regulatory elements at other credible set SNPs rs55686954 and rs3087243 (**Figure**
288 **6B**). Second, at the *TNFAIP3* locus, we observe high probability regulatory elements at
289 the functional SNP rs35926684 and other credible set SNP rs6927172 (**Figure 6C**) and
290 do not predict regulatory elements at the other 7 credible set SNPs. The CD4+ Th1
291 specific regulatory element at rs35926684 suggests that this SNP may alter gene
292 regulation specifically in Th1 cells and hence, we suggest any functional follow-up be
293 done in this cell-state. Fewer than 11% of the credible set SNPs in the other 18 fine-
294 mapped loci have high IMPACT cell-state-specific regulatory element probabilities
295 (**Figure S7**). We note that disease-relevant IMPACT functional annotations may be
296 integrated with existing functional fine mapping methods, like PAINTOR[41] or
297 CAVIARBF[42], to assign causal posterior probabilities to variants.
298
299 In summary, IMPACT predicts cell-state-specific regulatory elements based on
300 epigenomic and sequence profiles of experimental cell-state-specific TF binding. First,
301 we observed that the robust epigenomic footprint of TF binding sites allows for accurate
302 motif binding prediction. Second, we show the versatility of IMPACT to model functional
303 cell-state regulatory annotations to more precisely capture causal variation in gene
304 expression and disease susceptibility. Lastly, we demonstrated that IMPACT may
305 identify functional variants a priori. We recognize several important limitations to our
306 work: 1) we have not experimentally validated the activity of any of our predicted

307 regulatory elements, 2) predicted regulatory elements are limited to genomic regions
308 that have been epigenetically assayed, 3) IMPACT, as presented, is limited to cell-
309 states in which ChIP-seq of a key TF has been performed. In light of these limitations,
310 IMPACT is an emerging strategy for identifying trait associated regulatory elements and
311 generating hypotheses about the cell-states in which variants may be functional,
312 motivating the need to develop therapeutics that target specific disease-driving cell-
313 states.
314

315 ## Methods

316

317 ### Data

318

319 *Genome-wide Annotation Data.* We obtained publicly available genome-wide
320 annotations in a broad range of cell types for the GRCh37 (hg19) assembly. The
321 accession numbers and or file names for features downloaded from NCBI, Blueprint,
322 Roadmap, and chromHMM are listed in **Table S1**. Features from Finucane et al 2015[7]
323 are labeled as they were in supplemental tables of this study. Cell-type-specific
324 annotation types include ATAC-seq, DNase-seq, FAIRE-seq, HiChIP, polymerase and
325 elongation factor ChIP-seq, and histone modification ChIP-seq. Sequence annotations
326 were downloaded from UCSC's publicly available bedfiles and include conservation,
327 exons, introns, intergenic regions, 3'UTR, 5'UTR, promoter-TSS, TTS, and CpG
328 islands). All genome-wide feature data, except conservation, is represented in standard

329 6-column bedfile format. Conservation is represented in bedgraph format, in which the
330 average score is reported for each 1024 bp window genome-wide.
331
332 *TF ChIP-seq data.* We determined genome-wide TF occupancy from publicly available
333 ChIP-seq (**Table S2**) of 14 key regulators (T-BET[15,16], GATA3[17], STAT3[18], FOXP3[19],
334 STAT5[19], IRF5[37], IRF1[43], CEBPB[44], PAX5[45], REST[5], RXRA[5], HNF4A[20], TCF7L2[5], RNA
335 Pol II[5,16,46]) assayed in their respective primary cell-states Th1, Th2, Th17, Tregs, Tregs,
336 monocytes, monocytes, monocytes, B cells, fetal brain cells, brain cells, liver cells,
337 pancreatic cells, and lymphocytes. ChIP-seq peaks were called by macs[47] [v1.4.2
338 20120305] (all p<1e-5).
339
340 **IMPACT Model**
341
342 We build a model that predicts TF binding on a motif by learning the epigenomic profiles
343 of the TF binding sites. We use logistic regression to model the log odds of TF binding
344 based on a linear combination of the effects β_j of the *j* epigenomic features, where β_0 is
345 an intercept.
346
347 $$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j$$
348
349 From the log odds, which ranges from negative to positive infinity, we compute the
350 probability of TF binding, ranging from 0 to 1.
351 $$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j)}}$$

We use an elastic net logistic regression framework implemented by the *cv.glmnet* R [v1.0.143] package[48], in which optimal β are fit according to the following objective function,

$$\underset{\beta}{argmin}(\|Y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|)$$

We use a regularization strategy to constrain the values of $\beta$and help prevent overfitting. Elastic net regularization is a compromise between the lasso (L1) and ridge (L2) penalties. Both penalties have advantageous effects on the model fit of *j* features. Lasso performs feature selection, allowing only a subset to be used in the final model, pushing some β_j to 0, which is useful for large feature sets such as IMPACT and avoids overfitting. When used alone, lasso will arbitrarily select one of several correlated features to be included in the final model; incorporating the ridge term limits this effect of lasso. Furthermore, ridge provides a quadratic term, making the optimization problem convex with a single optimal solution. The elastic net logistic regression has the following parameters: *alpha*, *lambda*, and *type*. Alpha is the mix term between the lasso and ridge penalties in the objective function, which controls the sparsity of betas. We set *alpha* to 0.5, such that the L1 and L2 terms equally contribute, correlated features may be retained to some degree, and not too many betas are pushed to 0. We set *lambda* to *lambda.min*, the value of λ that yields minimum mean cross-validated error, and *type* set to response, such that our predictions are in probability space rather than log odds space. We choose to not employ a deep learning approach in order to retain interpretability of our model β. More precisely, knowledge of which genomic annotations

375 are most informative for predicting transcriptional regulation will be useful for deciding
376 which high-throughput experimental assays should be preferentially performed in an
377 effort to learn more about the regulome.
378
379 *Training IMPACT.* We train IMPACT on gold standard regulatory and non-regulatory
380 elements of a particular TF, meaning that there is one IMPACT model per TF/cell-state
381 pair. To build the regulatory class, we scanned the TF ChIP-seq peaks, mentioned
382 above, for matches to the TF-specific binding motif, using HOMER[49] [v4.8.3]. Each
383 match must receive a sequence similarity score greater than or equal to the threshold
384 provided by the PWMs (position weight matrices) or by HOMER (**Table S2**). We only
385 scan for the TF motif of the corresponding TF ChIP-seq dataset, e.g. we only looked for
386 T-BET motifs in the T-BET ChIP-seq data. We retained the highest scoring motif match
387 for each ChIP-seq peak and used the genomic coordinates of the center two
388 nucleotides to create a *GenomicRanges* object in R. For every run of IMPACT, 1,000
389 regulatory ranges are randomly selected, labeled with a 1, to train the model.
390 Controlling the number of ranges used will standardize the logistic regression output
391 such that predictions and model fits will be more comparable between cell-state models.
392
393 To build the non-regulatory class, we scanned the entire genome for TF motif matches,
394 again using HOMER, and selected motif matches with no overlap with that TF's ChIP-
395 seq peaks, e.g. we scan for T-BET motifs, and only retain regions not overlapping T-
396 BET ChIP-seq peaks. We do not check for overlap with other TF (i.e. GATA3, STAT3,
397 FOXP3) ChIP-seq peaks. Similarly to the regulatory set, we used the genomic

coordinates of the center two nucleotides of retained motifs to create the non-regulatory set, and label them as 0 in the regression. The motif matching process in both classes serves as a modest control for sequence content, as motifs are conserved regions of DNA. For every run of IMPACT, 10,000 regions of the non-regulatory set are randomly selected to train the model. This value is one order of magnitude larger than the regulatory set to reflect that genome-wide, we expect far more non-regulatory than regulatory elements. We justify setting this value to 10,000 if we assume that ~10% of the genome is regulatory, then for every positive region, we need nine negative regions. This would require 9*1,000 regulatory elements = 9,000 non-regulatory elements, a conservative estimate of the number of non-regulatory elements we actually use.

The sets of regulatory and non-regulatory elements are first partitioned into a random sampling of 80% each, to be used for 10-fold cross validation (CV), in which these sets are further partitioned into 90%/10% train/test. The remaining 20% to be used as a validation set (data completely unseen by the CV).

IMPACT is trained on standard 6-column bedfiles, of regions that are 2 base pairs wide, i.e.

```
chr1    2500   2501   region_1     0     +
```

Epigenomic and sequence features are represented twice in the model, first with respect to local regions, and secondly with respect to distal regions. In the local case,

421 for each region we annotate, we iterate through all genome-wide features, asking if
422 there is overlap. In the distal case, we look for feature overlap of a more distal
423 nucleotide (i.e. 1,000 base pairs up *or* downstream, such that overlap at either distal
424 position will count; we do not distinguish between the upstream or downstream overlap
425 in the feature matrix). Our rationale is that although a nucleotide may not intersect a
426 particular feature, it may be informative to know that there is one nearby. The upstream
427 and downstream distal coordinates (parameter set to 1,000 bp) are computed by
428 subtracting or adding, respectively, the parameter value to these coordinates. If the
429 computed distal coordinate is negative, the value is replaced with 1. If the computed
430 distal coordinate is larger than the length of the chromosome, the value is replaced with
431 the length of the chromosome. We set our distal parameter to 1,000 bp (**Table S3**) and
432 do not check for overlap of any additional nucleotides within this distance. The feature
433 matrix, on which the model is trained, may look like the following with dimensions
434 (11,000 rows by 2+2X columns):
435

Training region	Regulatory Class (1 or 0)	Conservation	Feat. 1 (local)	Feat. X (local)	Feat. 1 (distal)	Feat. X (distal)
Regulatory region N_r=1	1	43.25	1	0	1	1
. . .	. . .					

. . . . .	. . . . .					
Regulatory Region N_r=1,000	1	68.75	1	1	0	1
Non-regulatory Region N_{nr}=1	0	13.20	0	0	0	0
. . . .	. . . .					
Non-regulatory Region N_{nr}=10,000	0	14.34	0	0	1	0

436

437 We applied IMPACT genome-wide to assign nucleotide-resolution regulatory
438 probabilities, using the model β learned from the elastic net logistic regression CV. We
439 show the largest magnitude β for each cell-state IMPACT model in **Figure S2**.
440
441 **Stratified-LD Score Regression (S-LDSC)**
442
443 *Genome-wide association data.* We collected RA GWAS summary statistics[31] for
444 38,242 European individuals, combined cases and controls, and 22,515 East Asian
445 individuals, comprised of 4,873 RA cases and 17,642 controls[32]. Reference SNPs, used
446 to estimate European LD scores, were the set of 9,997,231 SNPs with minor allele
447 count greater or equal than five in a set of 659 European samples from phase 3 of 1000
448 Genomes Projects[50]. The regression coefficients were estimated using 1,125,060
449 HapMap3 SNPs and heritability was partitioned for the 5,961,159 reference SNPs with
450 MAF ≥ 0.05. Reference SNPs, used to estimate East Asian LD scores, were the set of
451 8,768,561 SNPs with minor allele count greater or equal than five in a set of 105 East
452 Asian samples from phase 3 of 1000 Genomes Projects[50]. The regression coefficients
453 were estimated using 1,026,051 HapMap3 SNPs and heritability was partitioned for the
454 5,469,053 reference SNPs with MAF ≥ 0.05. Frequency and weight files (1000G EUR
455 phase3, 1000G EAS phase3) are publicly available and may be found in our URLs.
456
457 *Methodology.* We apply S-LDSC[7] [v1.0.0], a method developed to partition polygenic
458 trait heritability by one or more functional annotations, to quantify the contribution of
459 IMPACT cell-state-specific regulatory annotations to RA and other autoimmune disease

460 heritability. We annotate common SNPs (MAF ≥ 0.05) with multiple cell-state-specific
461 IMPACT models, assigning a regulatory element score to each variant. Then S-LDSC
462 was run on the annotated SNPs to compute LD scores. Here, the two statistics we use
463 to evaluate each annotation's contribution to disease heritability are enrichment and
464 standardized effect size (τ^*).
465
466 If a_{cj} is the value of annotation *c* for SNP *j*, we assume the variance of the effect size of
467 SNP *j* depends linearly on the contribution of each annotation *c*:

468 $$Var(\beta_j) = \sum_c a_{cj}\tau_c$$

469 where τ_c is the per-SNP contribution from one unit of the annotation a_c to heritability. To
470 estimate τ_c, S-LDSC estimates the marginal effect size of SNP *j* in the sample from the
471 chi-squared GWAS statistic X_j^2:

472 $$X_j^2 = N\,\hat{\beta}_j^2$$

473 Considering the expectation of X_j^2 and following the derivation from Gazal et al 2017[9],
474

475 $$[\mathrm{X}_j^2] = N \sum_c (\tau_c \sum_k a_c(k)\, r_{jk}^2) + 1$$

476

477 $$E[\mathrm{X}_j^2] = N \sum_c \tau_c\, l(j,c) + 1$$

478

479 where *N* is the sample size of the GWAS, $l(j,c)$ is the LD score of SNP *j* with respect to
480 annotation *c*, and r_{jk}^2 is the true, e.g. population-wide, genetic correlation of SNPs *j* and

481 *k*. Since τ_c is not comparable between annotations or traits, Gazal et al 2017[9] defines τ_c^*
482 as the per-annotation standardized effect size, a function of the standard deviation of
483 the annotation *c*, $sd(c)$, the trait-specific SNP-heritability estimated by LDSC h_g^2, and the
484 total number of reference common SNPs used to compute h_g^2, *M* = 5,961,159 in EUR
485 and 5,469,053 in EAS:

486 $$\tau_c^* = \frac{sd(c)\tau_c}{h_g^2/M}$$

487 We define enrichment of an annotation as the proportion of heritability explained by the
488 annotation divided by the average value of the annotation across the *M* common (MAF
489 ≤ 0.05) SNPs. Enrichment may be computed for binary or continuous annotations
490 according to the equation below, where $h_g^2\,(c)$ is the h2 explained by SNPs in annotation *c*.

491 $$Enrichment = \frac{h_g^2\,(c)/h_g^2}{\frac{\sum_j a_c(j)}{M}} = \frac{\sum_j a_c(j)\hat{\tau}_c / \sum_j \sum_c a_c(j)\hat{\tau}_c}{\frac{\sum_j a_c(j)}{M}}$$

492 Enrichment does not quantify effects that are unique to a given annotation, whereas τ^*
493 does.
494
495 Each S-LDSC analysis involves conditioning IMPACT annotations on 69 baseline
496 annotations, a subset of the 75 annotations referred to as the baseline-LD model. Our
497 69 annotations consist of 53 cell-type-nonspecific annotations[7], which include histone
498 marks and open chromatin, 10 MAF bins, and 6 LD-related annotations[9] to assess if
499 functional enrichment is cell-type-specific and to control for the effect of MAF and LD
500 architecture. Consistent inclusion of MAF and LD associated annotations in the baseline
501 model is a standard recommended practice of S-LDSC. When conditionally comparing

502 two annotations, say A and B, in a single S-LDSC model, the two annotations may have
503 similar enrichments if they are highly correlated. However, the τ^* for the annotation with
504 greater true causal variant membership will be larger and more statistically significant
505 (e.g. > 0). Specifically, a τ^* of 0, means that the annotation does not change per-SNP
506 h2. A strongly negative $\tau *$ means that membership to the categorical annotation
507 decreases per-SNP h2, while a strongly positive τ^* means that membership to the
508 annotation increases per-SNP h2. The significance of τ^* is computed based on a test of
509 how different from 0 the $\tau *$ is.
510
511 *MHC exclusion.* We note that S-LDSC excludes the MHC (major histocompatibility
512 complex) due to its extremely high gene density and outlier LD structure, which is
513 thought to be the strongest contributor to RA disease h2[51]. However, our work supports
514 the notion that there is an undeniably large amount of RA h2 located outside of the
515 MHC.
516
517 **eQTL enrichment**
518 We acquired SNP-level summary statistics for 3,754 peripheral blood samples[23]. We
519 computed a genome-wide enrichment of cis eQTL causal association across various
520 functional annotations. To this end, we gathered summary statistics for each gene in the
521 dataset, choosing only one probe when there was more than one tested on the array,
522 and retained only the association statistics for the SNPs in a cis window of 1 Mb
523 upstream and downstream of the gene TSS. For each gene in the context of each
524 annotation, an enrichment was calculated explicitly as:

525 $$Enrichment_{g,a} = \frac{(\sum_{i=1}^{N} \chi^2{}_g)/N}{(\sum_{j=1}^{M} \chi^2{}_g)/M}$$

526 where g is the gene, a is the annotation, N is the number of variants within annotation a,
527 M is the number of variants outside annotation a, i is the ith variant, j is the jth variant,
528 and χ^2 is the chi-squared statistic of the association between gene g and SNP i or j. We then
529 computed genome-wide standard errors by block jackknifing the genome into 200
530 adjacent bins and computed a distribution of enrichment values when leaving one bin
531 out at a time. This strategy is designed to prevent the genes of any one region of the
532 genome from dominating the enrichment statistic. Furthermore, we used a permutation
533 strategy to establish a null distribution. To this end, we randomly permuted the chi-
534 squared associations in the cis-window of each gene 1000 times, while matching on 50
535 LD bins across the cis-window, and recomputed the enrichment with each of the
536 functional annotations.
537
538 **Posterior probability enrichment**
539
540 Previous work from our group aimed to define the most likely causal RA variant for each
541 locus harboring a genome-wide significant variant[40]. To this end, posterior probabilities
542 were computed with the approximate Bayesian factor (ABF), assuming one causal
543 variant per locus. The posteriors were defined as:

544 $$P_i = \frac{ABF_i}{\sum_{k=0}^{n} ABF_k}$$

545 where *i* is the ith variant, and *n* is the total number of variants in the locus. As such, the
546 ABF over all variants in a locus sum to 1. Then, for each of the defined 20 RA-

547 associated loci[40], we computed the enrichment of high posterior probabilities in the top
548 1% of cell-state-specific IMPACT regulatory elements (**Table S5**). For each RA-
549 associated locus *l,*

550 $$enrichment = \frac{\sum_{i}^{M_l} P_c(i)}{\sum_{j}^{M_l} \frac{1}{M_l}}$$

551 where $P_c(i)$ is the posterior causal probability of SNP *i*, such that *i* belongs to the top 1%
552 of the cell-state-specific IMPACT annotation *c*, M_l is the number of SNPs in locus *l* for
553 which we have a computed posterior probability. The denominator formulates the null
554 hypothesis that each SNP in a locus is equally causal. We computed the average of
555 these enrichment values over the 20 RA-associated loci. A permutation distribution was
556 calculated by computing an average enrichment value over these 20 loci, in 1,000 trials,
557 in which random posterior probabilities (of the same quantity M_l) were selected. The
558 permutation *p*-value was calculated by comparing the quantile value of the IMPACT
559 enrichment to the assumed-normal permutation distribution defined by its mean and
560 standard deviation, using the *pnorm* function in R.
561
562 **Caveats**
563
564 This approach might be easily applied to a wide-range of other diseases, recognizing
565 several important caveats. First, it relies on choosing a key regulator TF with a known
566 consensus binding motif. Second, while for CD4+ T cells there is extensive literature of
567 cell-states and their key regulatory drivers, this may not be readily available for all cell
568 types. Third, it requires that primary cell ChIP-seq data, and therefore a specific

antibody, are available for the desired TF in the disease-driving cell type. We prefer primary cell ChIP-seq data as opposed to cell line data in order to more closely approximate physiological regulatory element activity and chromatin dynamics. Fourth, as most immune populations are difficult to sort ex-vivo, collected cell-state-specific ChIP-seq data may not be 100% homogeneous. We expect that IMPACT predicted regulatory elements should be robust to experimental binding data of a heterogeneous population of cells. For instance, in a population of T cells that are 25% Tregs, and 75% other T lymphocytes, we expected more FOXP3 binding in the Tregs, as it is a specific regulator, relative to the other cells. Therefore, our TF occupancy profile should predominantly reflect that of Tregs. As there is especially poor availability of TF occupancy data in non-immune primary cells, we encourage the generation of more cell-state-specific data. Furthermore, due to IMPACT's skew toward immune cell type features (~50%), we recommend supplementing relevant cell-type-specific features to avoid overfitting to immune cell types.

Data Availability

All code for this paper has been made publicly available in the following GitHub repository:
https://github.com/immunogenomics/IMPACT

URLs

1. S-LDSC tutorial and instructions: github.com/bulik/ldsc

2. 1000G: www.1000genomes.org
3. RA EUR summary statistics: http://plaza.umin.ac.jp/yokada/datasource/software.htm
4. RA EAS summary statistics: http://jenger.riken.jp/en/result
5. 1000G Phase 3 LD scores, CD4+ T cell specifically expressed genes (binary functional annotations): http://data.broadinstitute.org/alkesgroup/LDSCORE/
6. Immgen.tsv: https://gist.github.com/nachocab/3d9f374e0ade031c475a

Acknowledgements

We would like to thank Eric Lander and Jesse Engreitz for helpful discussions. This work is supported in part by funding from the National Institutes of Health (NHGRI T32 HG002295, UH2AR067677, 1U01HG009088, U01 HG009379 (Price/Raychaudhuri U01), and 1R01AR063759) and the Doris Duke Charitable Foundation Grant #2013097.

Author Contributions

Developed IMPACT: T.A., Y.L., E.D.

Heritability analysis: T.A., Y.L., S.G., B.v.d.G.

ATAC-seq: H-J.W., N.T.

RA genetic data: Y.O., K.Y., S.R.

RA disease/genetic interpretation: Y.O., K.Y., S.R.

Statistical analysis: T.A., Y.L., S.G., B.v.d.G.

T.A. wrote the initial draft; all authors contributed to the final manuscript.

Work was conceived by T.A., S.R., A.P. and supervised by S.R. and A.P.

615

616 Competing Financial Interests

617 The authors declare no competing financial interests.

618

619 References

620 1. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172,** 650–665 (2018).

621 2. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and

622 characterization. *Nat. Methods* **9,** 215–6 (2012).

623 3. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin

624 structure through genomic segmentation. *Nat. Methods* **9,** 473–476 (2012).

625 4. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and

626 Translation. (2017). doi:10.1016/j.ajhg.2017.06.005

627 5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the

628 human genome. *Nature* **489,** 57–74 (2012).

629 6. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes.

630 *Nature* **518,** 317–330 (2015).

631 7. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using

632 genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

633 8. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes

634 identifies disease-relevant tissues and cell types. *Nat. Genet.* **50,** 621–629 (2018).

635 9. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex

636 traits shows action of negative selection. *Nat. Genet.* **49,** 1421–1427 (2017).

637 10. Maurano, M. T. *et al.* Systematic localization of common disease-associated

638 variation in regulatory DNA. *Science* **337,** 1190–5 (2012).
639 11. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping
640 complex trait variants. *Nat. Genet.* **45,** 124–130 (2013).
641 12. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune
642 disease variants. *Nature* **518,** 337–343 (2014).
643 13. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide
644 Association Studies of 18 Human Traits. *Am. J. Hum. Genet.* **94,** 559–573 (2014).
645 14. ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge -
646 syn6131484. Available at:
647 https://www.synapse.org/#!Synapse:syn6131484/wiki/402026. (Accessed: 16th
648 April 2017)
649 15. Soderquest, K. *et al.* Genetic variants alter T-bet binding and gene expression in
650 mucosal inflammatory disease. *PLOS Genet.* **13,** e1006587 (2017).
651 16. Hertweck, A. *et al.* T-bet Activates Th1 Genes through Mediator and the Super
652 Elongation Complex. *Cell Rep.* **15,** 2756–2770 (2016).
653 17. Gustafsson, M. *et al.* A validated gene regulatory network and GWAS identifies
654 early regulators of T cell–associated diseases. *Sci. Transl. Med.* **7,** 313ra178-
655 313ra178 (2015).
656 18. Tripathi, S. K. *et al.* Genome-wide Analysis of STAT3-Mediated Transcription
657 during Early Human Th17 Cell Differentiation. *Cell Rep.* **19,** 1888–1901 (2017).
658 19. Schmidl, C. *et al.* The enhancer and promoter landscape of human regulatory and
659 conventional T-cell subpopulations. *Blood* **123,** e68–e78 (2014).
660 20. Gertz, J. *et al.* Distinct Properties of Cell-Type-Specific and Shared Transcription

661 Factor Binding Sites. *Mol. Cell* **52,** 25–36 (2013).
662 21. Davenport, E. E. *et al.* Identification of drug eQTL interactions from repeat
663 transcriptional and environmental measurements in a lupus clinical trial. *bioRxiv*
664 118703 (2017). doi:10.1101/118703
665 22. Liu, X. *et al.* Functional Architectures of Local and Distal Regulation of Gene
666 Expression in Multiple Human Tissues. *Am. J. Hum. Genet.* **100,** 605–616 (2017).
667 23. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral
668 blood. *Nat. Genet.* **46,** 430–437 (2014).
669 24. Firestein, G. S. Evolving concepts of rheumatoid arthritis. *Nature* **423,** 356–361
670 (2003).
671 25. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the
672 association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44,**
673 291–296 (2012).
674 26. Terao, C. *et al.* Genetic landscape of interactive effects ofHLA-DRB1alleles on
675 susceptibility to ACPA(+) rheumatoid arthritis and ACPA levels in Japanese
676 population. *J. Med. Genet.* **54,** 853–858 (2017).
677 27. Terao, C., Raychaudhuri, S. & Gregersen, P. K. Recent Advances in Defining the
678 Genetic Basis of Rheumatoid Arthritis. *Annu. Rev. Genomics Hum. Genet.* **17,**
679 273–301 (2016).
680 28. Hu, X. *et al.* Integrating Autoimmune Risk Loci with Gene-Expression Data
681 Identifies Specific Pathogenic Immune Cell Subsets. *Am. J. Hum. Genet.* **89,**
682 496–506 (2011).
683 29. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to

684 Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J.*
685 *Hum. Genet.* **97,** 139–152 (2015).
686 30. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of Effector CD4 T Cell
687 Populations. *Annu. Rev. Immunol.* **28,** 445–489 (2010).
688 31. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug
689 discovery. *Nature* **506,** 376–381 (2014).
690 32. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population
691 links cell types to complex human diseases. *Nat. Genet.* **50,** 390–400 (2018).
692 33. Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes
693 in T cells. *Nature* **520,** 558–562 (2015).
694 34. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the
695 genetic architecture of diseases and complex traits. *Nat. Genet.* 1 (2018).
696 doi:10.1038/s41588-018-0148-2
697 35. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci
698 and evidence for colocalization of causal variants with lymphoid gene enhancers.
699 *Nat. Genet.* **47,** 381–6 (2015).
700 36. Oliveira, T. V., Maniero, F., Santos, M. H. H., Bydlowski, S. P. & Maranhão, R. C.
701 Impact of high cholesterol intake on tissue cholesterol content and lipid transfers
702 to high-density lipoprotein. *Nutrition* **27,** 713–718 (2011).
703 37. Wang, C. *et al.* Genome-wide profiling of target genes for the systemic lupus
704 erythematosus-associated transcription factors IRF5 and STAT4. *Ann. Rheum.*
705 *Dis.* **72,** 96–103 (2013).
706 38. Mokhtari, R. & Lachman, H. M. The Major Histocompatibility Complex (MHC) in

707 Schizophrenia: A Review. *J. Clin. Cell. Immunol.* **7,** (2016).
708 39. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for
709 rheumatoid arthritis. *Nat. Genet.* **44,** 1336–1340 (2012).
710 40. Westra, H.-J. *et al.* Fine-mapping identifies causal variants for RA and T1D in
711 DNASE1L3, SIRPG, MEG3, TNFAIP3 and CD28/CTLA4 loci. *bioRxiv* 151423
712 (2017). doi:10.1101/151423
713 41. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk
714 loci. *Bioinformatics* **33,** 248–255 (2017).
715 42. Chen, W., Mcdonnell, S. K., Thibodeau, S. N., Tillmans, L. S. & Schaid, D. J.
716 Incorporating Functional Annotations for Fine-Mapping Causal Variants in a
717 Bayesian Framework Using Summary Statistics. (2016).
718 doi:10.1534/genetics.116.188953
719 43. Qiao, Y. *et al.* Synergistic Activation of Inflammatory Cytokine Genes by
720 Interferon-γ-Induced Chromatin Remodeling and Toll-like Receptor Signaling.
721 *Immunity* **39,** 454–469 (2013).
722 44. Pham, T.-H. *et al.* Dynamic epigenetic enhancer signatures reveal key
723 transcription factors associated with monocytic differentiation states. *Blood* **119,**
724 e161–e171 (2012).
725 45. Dimitrova, L. *et al.* PAX5 overexpression is not enough to reestablish the mature
726 B-cell phenotype in classical Hodgkin lymphoma. *Leukemia* **28,** 213–216 (2014).
727 46. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans.
728 *Science (80-.).* **328,** 232–235 (2010).
729 47. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,**

730 R137 (2008).
731 48. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized
732 Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** 1–22 (2010).
733 49. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription
734 Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell
735 Identities. *Mol. Cell* **38,** 576–589 (2010).
736 50. Gibbs, R. A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–
737 74 (2015).
738 51. Fonseka, C. Y., Rao, D. A. & Raychaudhuri, S. Leveraging blood and tissue CD4+
739 T cell heterogeneity at the single cell level to identify mechanisms of disease in
740 rheumatoid arthritis. *Curr. Opin. Immunol.* **49,** 27–36 (2017).
741
742
743
744
745
746
747
748

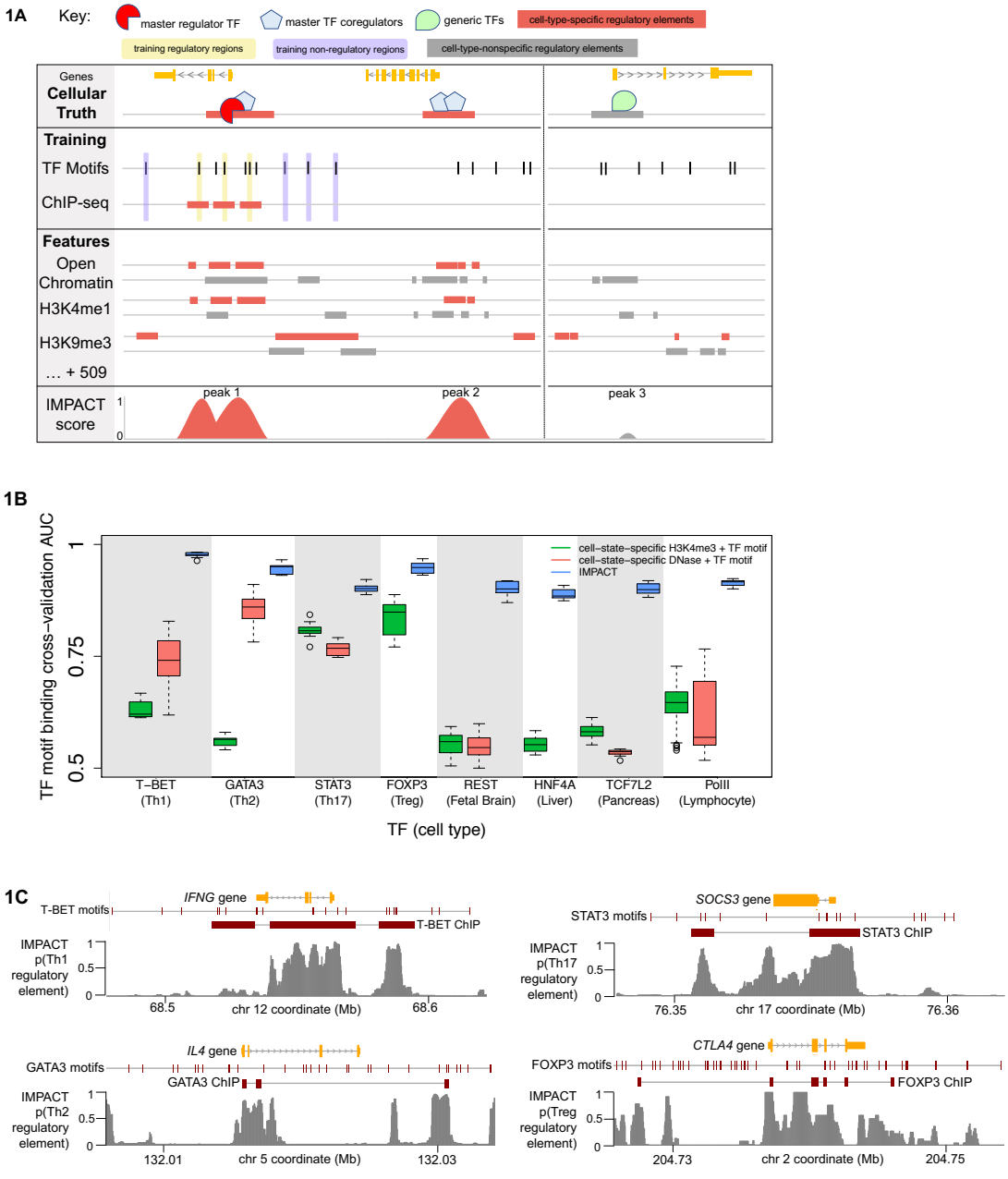


Figure 1.

(a) IMPACT learns a chromatin profile of cell-type-specific regulation, characteristic of the master regulator TF (red) “gold standard” regulatory elements (TF motif and ChIP-seq peak, yellow) and non-regulatory elements (TF motif, no ChIP-seq peak, purple). In this toy example, IMPACT learns that cell-type-specific open chromatin and H3K4me1 are strong predictors of cell-type-specific regulatory elements, while cell-type-

758 nonspecific DHS and H3K4me1 are less informative. IMPACT also learns that
759 H3K9me3 is a strong predictor of non-regulatory elements. IMPACT is expected to re-
760 identify regulatory elements marked by master TF binding (ChIP-seq and motif) [peak
761 1], while also identifying others where the chromatin profile is similar, perhaps
762 representing cell-type-specific transcriptional processes [peak 2]. IMPACT is not
763 expected to predict regulation at cell-type-nonspecific elements [peak 3], such as
764 promoters of generic housekeeping genes, assuming these elements have different
765 chromatin profiles. **(b)** IMPACT significantly outperforms cell-state-specific active
766 promoter (H3K4me3, green) and open chromatin (DNase, red) annotations in predicting
767 TF binding on a motif over 10 trials measured by computing the average ROC AUC
768 (receiver operator characteristic area under the curve). As there is no Treg DNase data,
769 there is no comparison AUC distribution. **(c)** Cell-state-specific regulatory element
770 IMPACT predictions for canonical target genes of T-BET, GATA3, STAT3, and FOXP3.
771
772
773
774

2

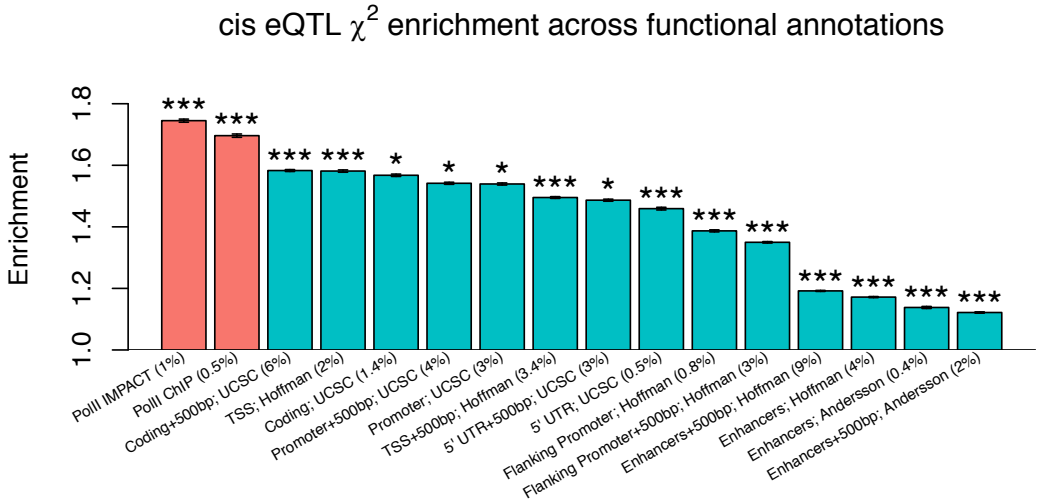


Figure 2.

Enrichment of cis eQTL chi-squared statistics, measured in 3,754 peripheral blood samples, across functional annotations including Pol II IMPACT, Pol II ChIP, used to train IMPACT, and various sequence annotations. Values in parentheses after annotation name are the average annotation value across all common variants and represent the effective genome-wide size of the annotation. No asterisk denotes permutation $p>0.05$, 1 asterisk permutation $p<0.05$, 2 asterisks permutation $p<0.01$, 3 asterisks permutation $p<0.001$. Intervals at the top of each bar represent the 95% confidence interval of the enrichment estimate, which was averaged over 7,025 unique genes genome-wide.

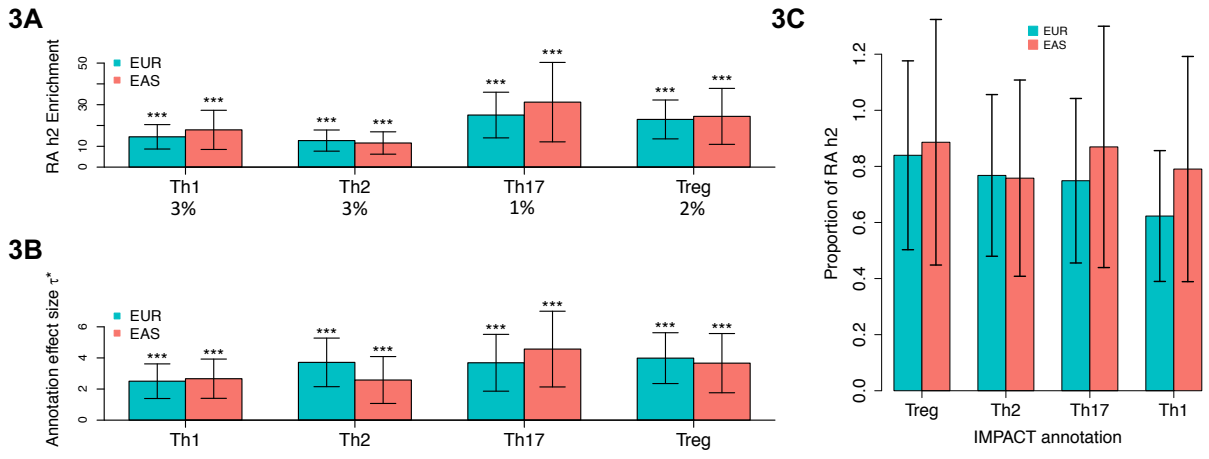


Figure 3.

(a) Enrichment of RA h2 in CD4+ T IMPACT for EUR and EAS populations. Values below cell-states are the average annotation value across all common variants and represent the effective genome-wide size of the annotation. **(b)** Annotation effect size (τ*) of each annotation separately conditioned on the baseline-LD. **(c)** Proportion of total causal RA h2 explained by the top 5% of SNPs in each IMPACT annotation. For all panels, 95% CI represented by black lines. For panels a and b, no asterisk denotes p>0.05, 1 asterisk p<0.05, 2 asterisks p<0.01, 3 asterisks p<0.001.

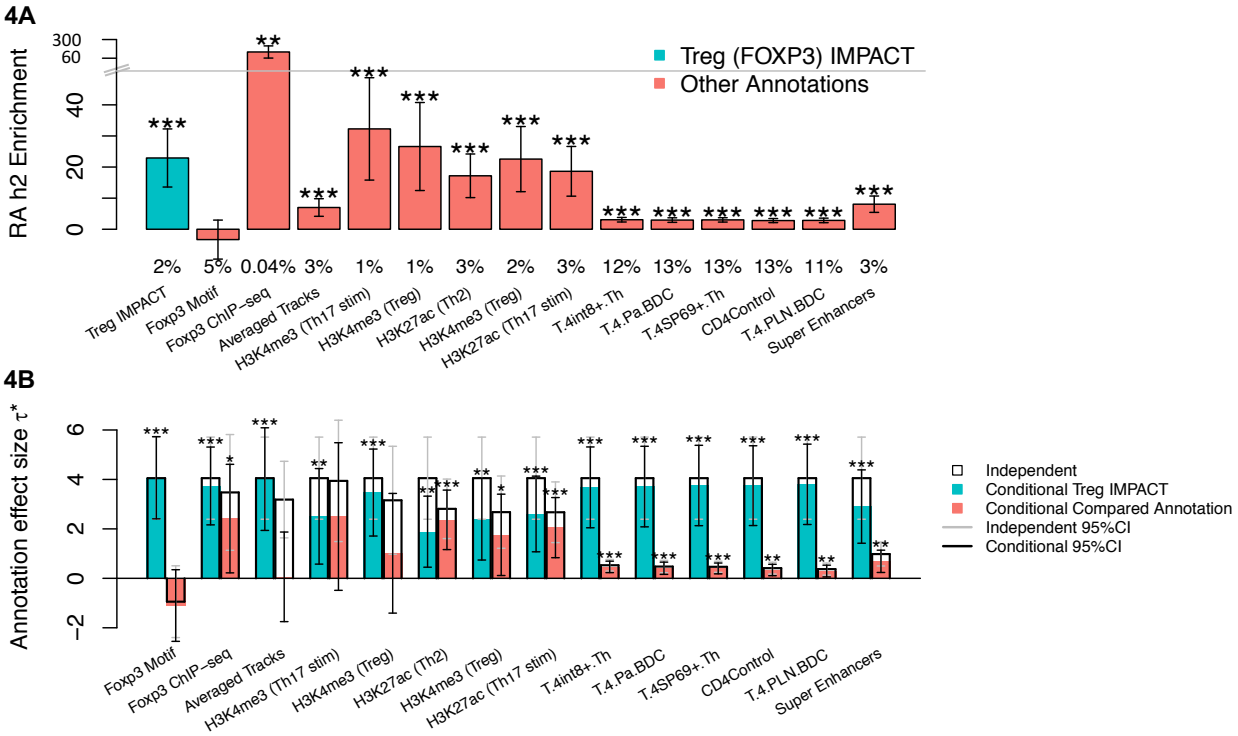


Figure 4.

(a) RA h2 enrichment of CD4+ Treg related annotations and compared T cell functional annotations. Values below cell-states represent the effective genome-wide size of the annotation. From left to right, we compare Treg IMPACT to genome-wide FOXP3 motifs, FOXP3 ChIP-seq, a genome-wide averaged track of features in the IMPACT framework (Averaged Tracks), the top 5, in terms of independent τ^*, cell-type-specific histone modification annotations[7], the top 5, in terms of independent τ^*, cell-type-specifically expressed gene sets (URLS)[8], and T cell super enhancers[33]. **(b)** CD4+ Treg IMPACT annotation standardized effect size (τ^*) consistently significantly greater than zero when conditioned on other T cell related functional annotations. τ^* for independent (e.g. non-conditional) analyses are denoted by the top of each black bar, as a reference for the conditional analyses, denoted by the top of each colored bar. The Treg IMPACT annotation captures a significant amount of RA h2, denoted by significantly positive τ^*,

regardless of the conditioned annotation. For panels a and b, no asterisk denotes $p>0.05$, 1 asterisk $p<0.05$, 2 asterisks $p<0.01$, 3 asterisks $p<0.001$.

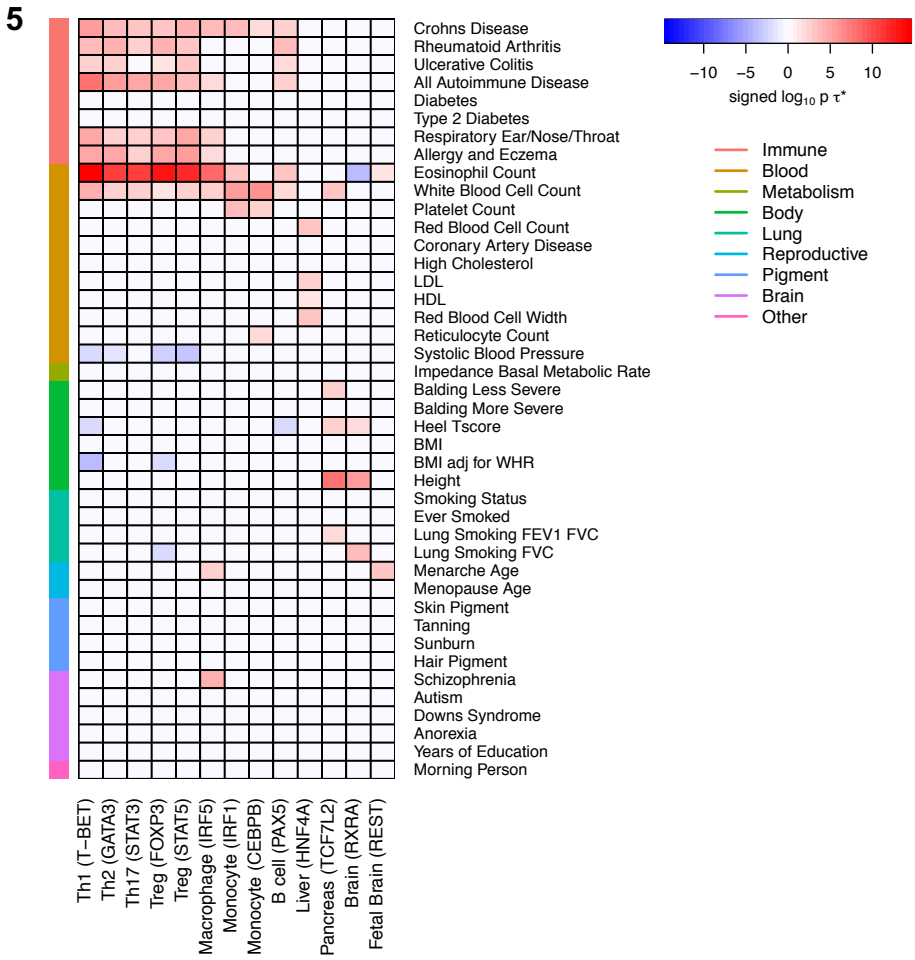


Figure 5.

Signed $\log_{10}$ p-values of τ^* for 42 traits across the four CD4+ T IMPACT and various other cell types for comparison. Annotations capture h2 in distinct sets of complex traits, shown by significantly positive τ^*. Color shown only if p-value of $\tau^* < 0.025$ after multiple hypothesis correction.

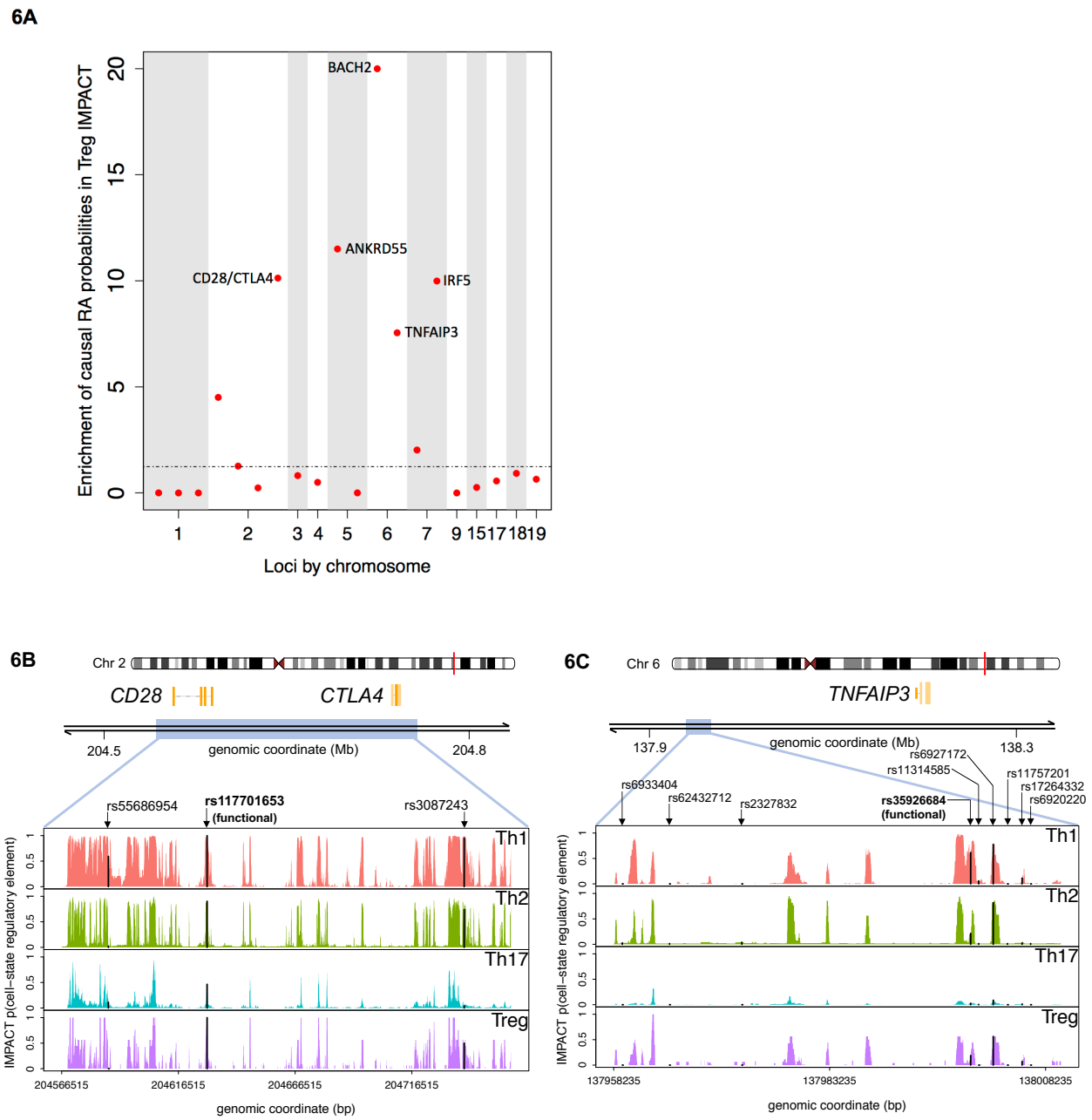
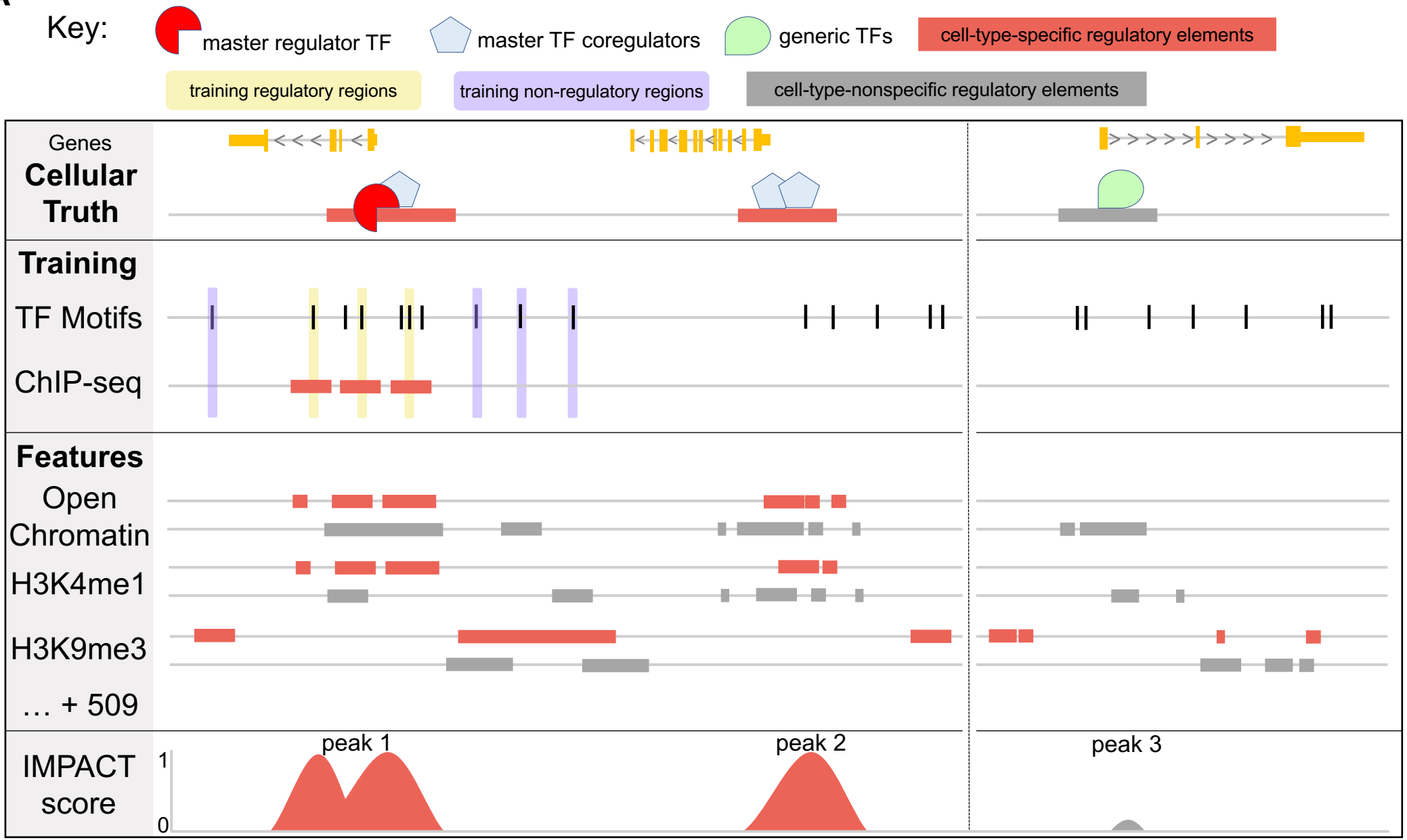


Figure 6.

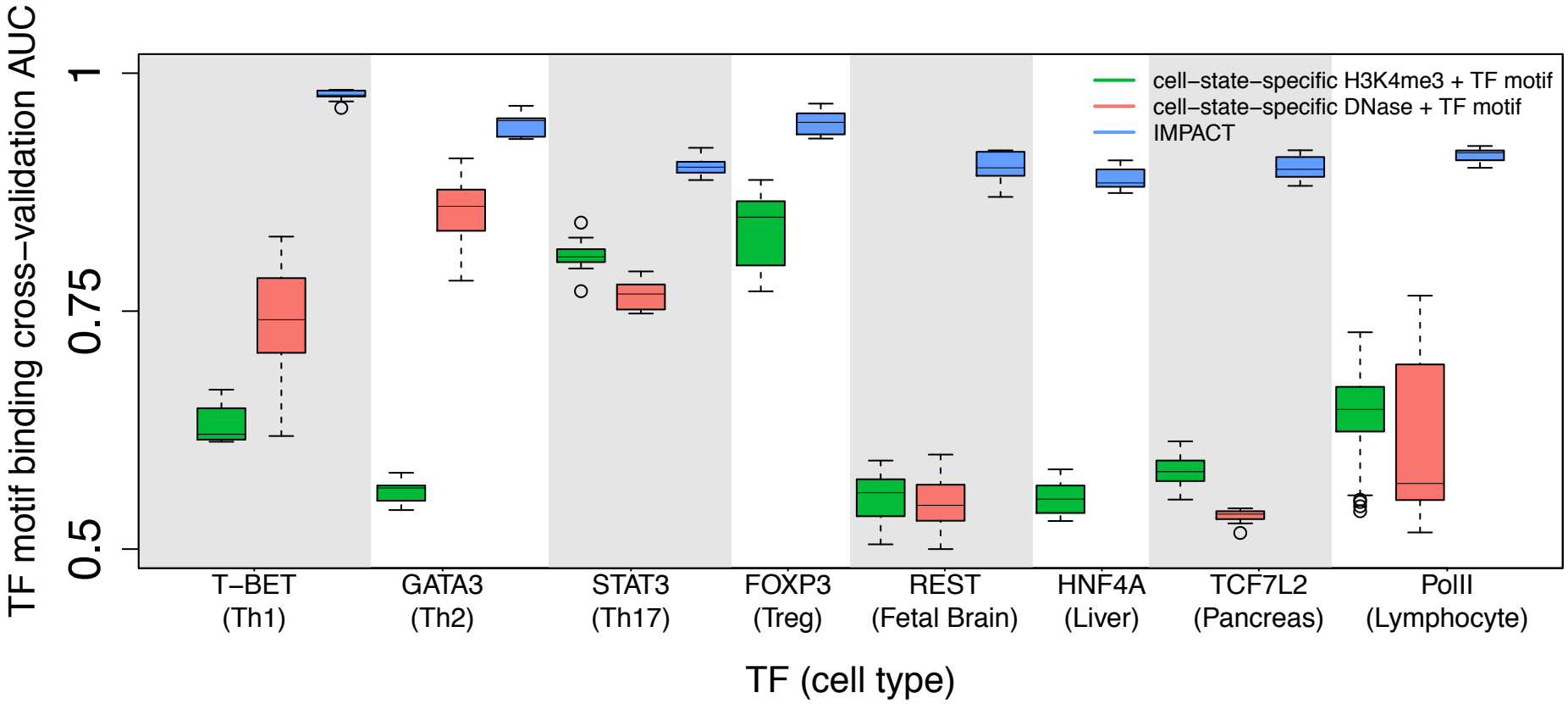
(a) Significant enrichment of posterior probabilities of putatively causal RA SNPs in the top 1% of SNPs annotated with CD4+ Treg regulatory element probabilities, highlighting particularly strong enrichment at the *BACH2*, *ANKRD55*, *CTLA4/CD28*, *IRF5*, and *TNFAIP3* loci. **(b,c)** IMPACT corroborates experimental validations of putatively causal

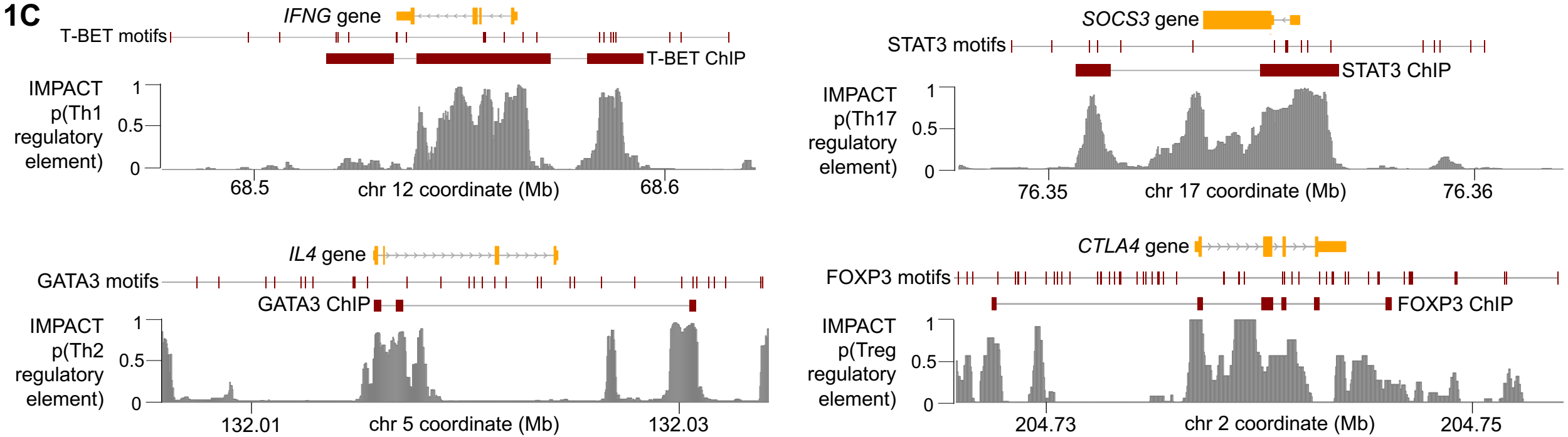
854 RA SNPs. For two RA-associated loci, *CTLA4/CD28* and *TNFAIP3*, we examine the
855 putatively causal SNP with experimentally validated differential enhancer activity
856 (bolded) and other 90% credible set SNPs (unbolded)[40]. IMPACT scores at these SNPs
857 are highlighted with a black line. **(b)** We observe high probability IMPACT regulatory
858 elements in all four CD4+ T cell-states for the functional SNP rs117701653 in the
859 *CD28/CTLA4* locus. **(c)** We observe a high probability Th1-specific IMPACT regulatory
860 element for the functional SNP rs35926684 in the *TNFAIP3* locus.

1A

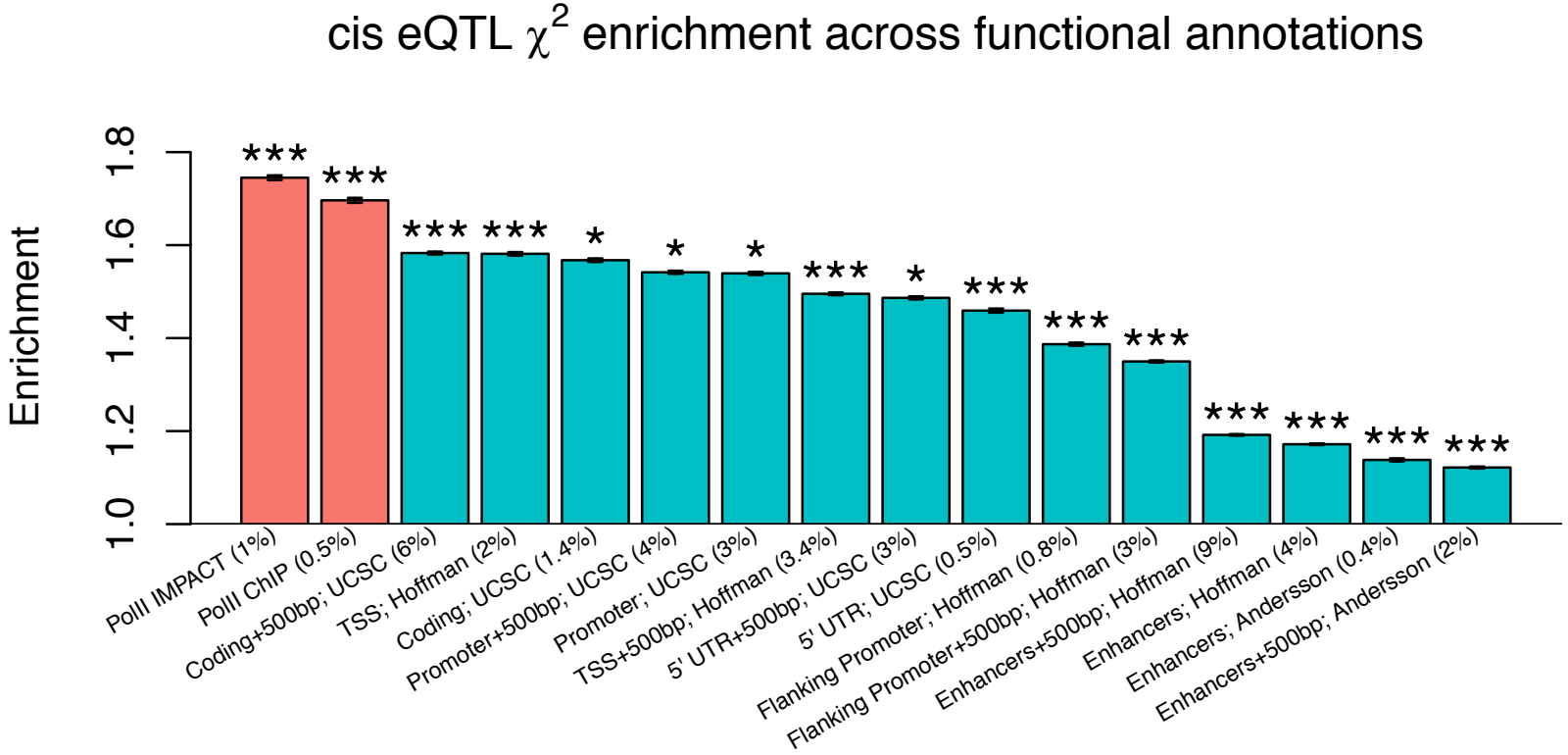

Key:
master regulator TF
master TF coregulators
generic TFs
cell-type-specific regulatory elements
training regulatory regions
training non-regulatory regions
cell-type-nonspecific regulatory elements
Genes
Cellular Truth
Training
TF Motifs
ChIP-seq
Features
Open Chromatin
H3K4me1
H3K9me3
… + 509
IMPACT score
1
0
peak 1
peak 2
peak 3

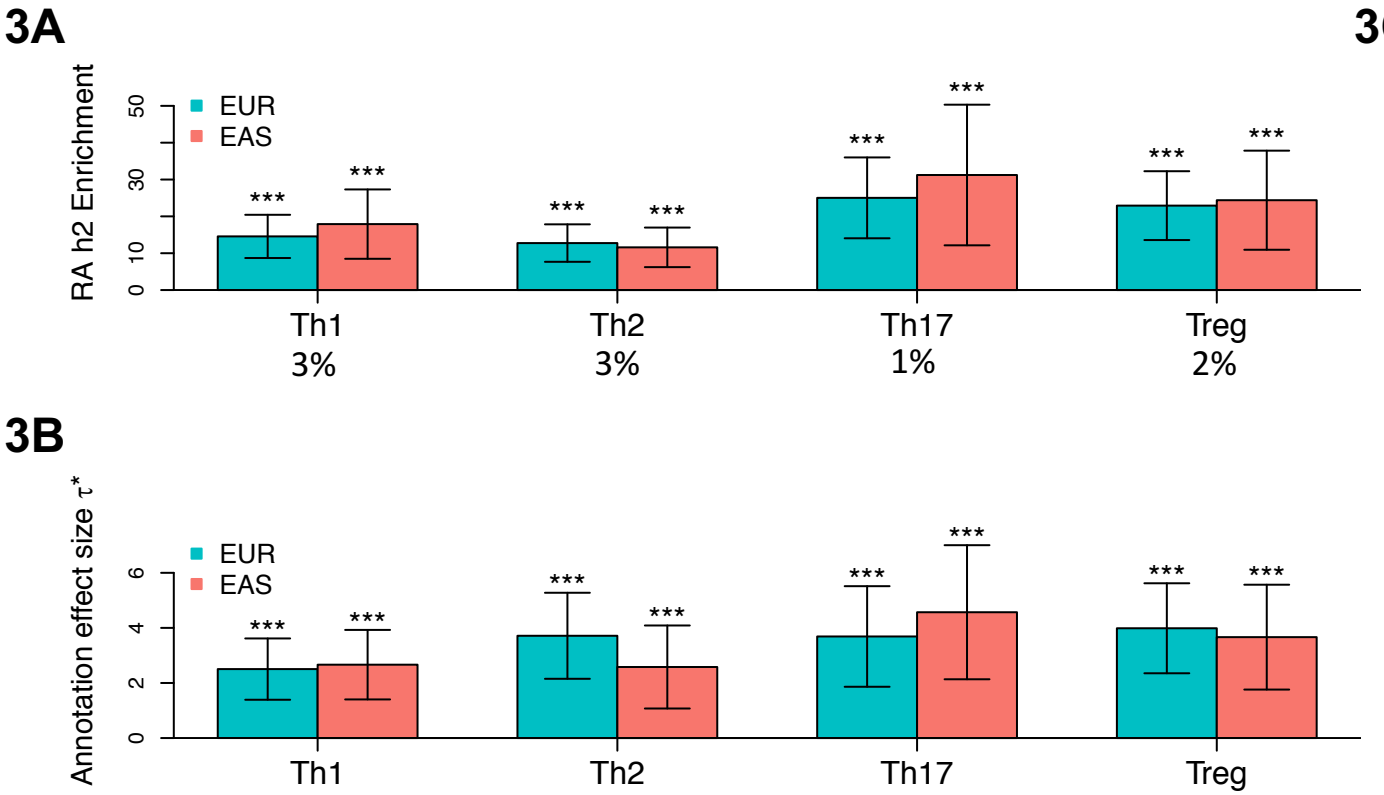
1B

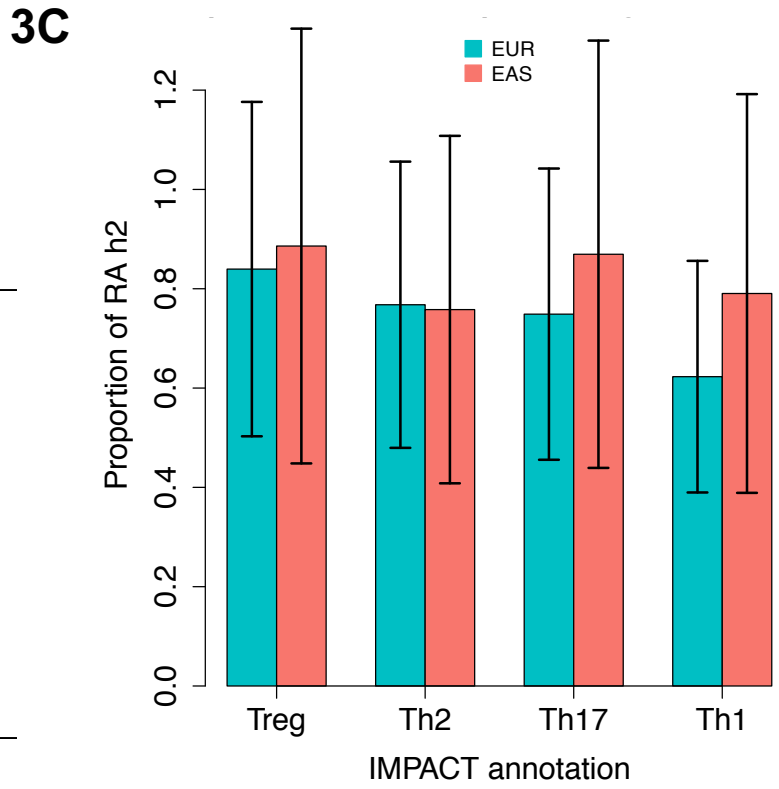

TF motif binding cross−validation AUC
1
0.75
0.5
cell−state−specific H3K4me3 + TF motif
cell−state−specific DNase + TF motif
IMPACT
T−BET (Th1)
GATA3 (Th2)
STAT3 (Th17)
FOXP3 (Treg)
REST (Fetal Brain)
HNF4A (Liver)
TCF7L2 (Pancreas)
PolII (Lymphocyte)
TF (cell type)


1C
IFNG gene
T-BET motifs
T-BET ChIP
IMPACT p(Th1 regulatory element)
1
0.5
0
68.5
chr 12 coordinate (Mb)
68.6
SOCS3 gene
STAT3 motifs
STAT3 ChIP
IMPACT p(Th17 regulatory element)
1
0.5
0
76.35
chr 17 coordinate (Mb)
76.36
IL4 gene
GATA3 motifs
GATA3 ChIP
IMPACT p(Th2 regulatory element)
1
0.5
0
132.01
chr 5 coordinate (Mb)
132.03
CTLA4 gene
FOXP3 motifs
FOXP3 ChIP
IMPACT p(Treg regulatory element)
1
0.5
0
204.73
chr 2 coordinate (Mb)
204.75

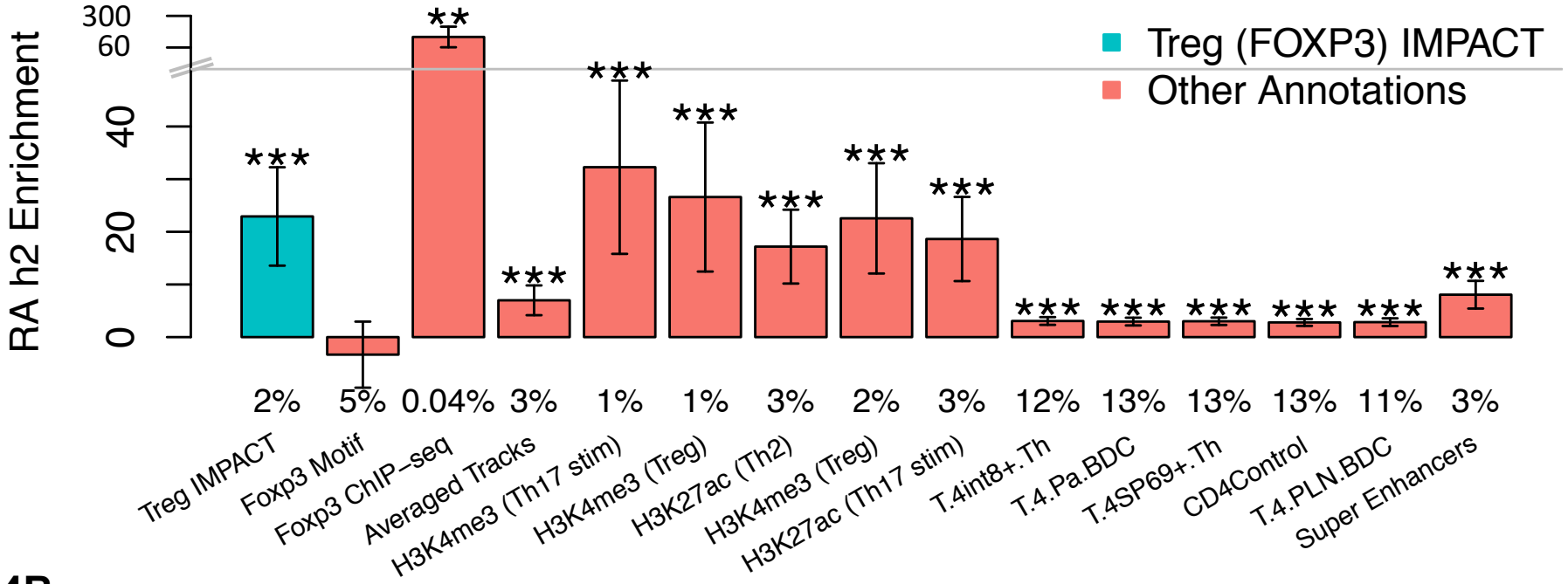
2


cis eQTL χ^2 enrichment across functional annotations
Enrichment
1.0
1.2
1.4
1.6
1.8
***
***
***
***
*
*
*
***
*
***
***
***
***
***
***
***
PolII IMPACT (1%)
PolII ChIP (0.5%)
Coding+500bp; UCSC (6%)
TSS; Hoffman (2%)
Coding; UCSC (1.4%)
Promoter+500bp; UCSC (4%)
Promoter; UCSC (3%)
TSS+500bp; Hoffman (3.4%)
5' UTR+500bp; UCSC (3%)
5' UTR; UCSC (0.5%)
Flanking Promoter; Hoffman (0.8%)
Flanking Promoter+500bp; Hoffman (3%)
Enhancers+500bp; Hoffman (9%)
Enhancers; Hoffman (4%)
Enhancers; Andersson (0.4%)
Enhancers+500bp; Andersson (2%)

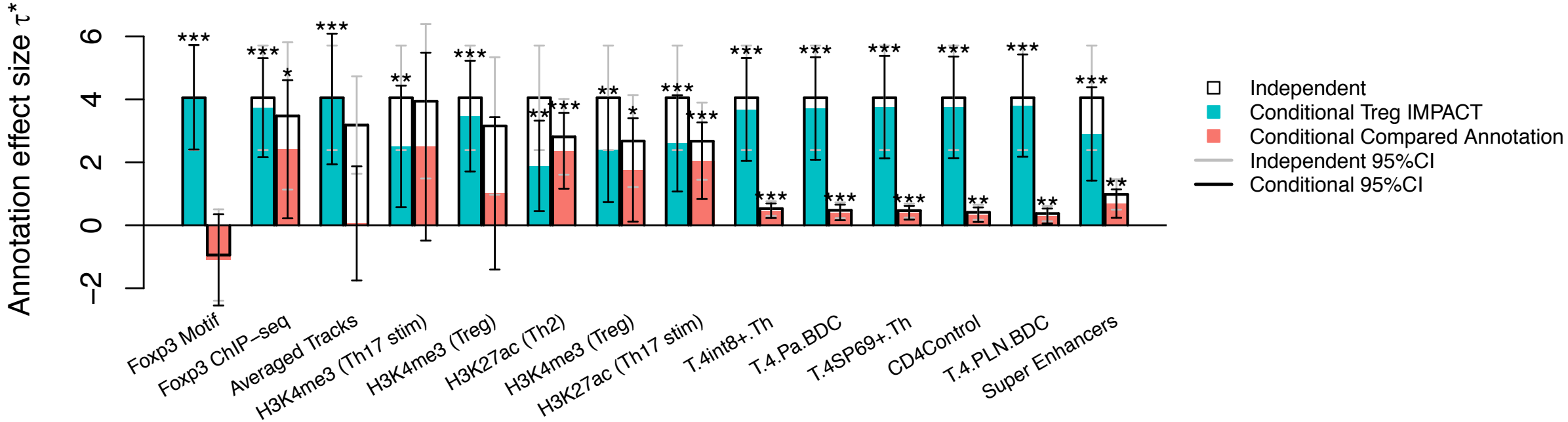

3A
RA h2 Enrichment
EUR
EAS
0
10
30
50
***
Th1
3%
Th2
3%
Th17
1%
Treg
2%
3B
Annotation effect size τ*
0
2
4
6
Th1
Th2
Th17
Treg


3C
EUR
EAS
Proportion of RA h2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Treg
Th2
Th17
Th1
IMPACT annotation

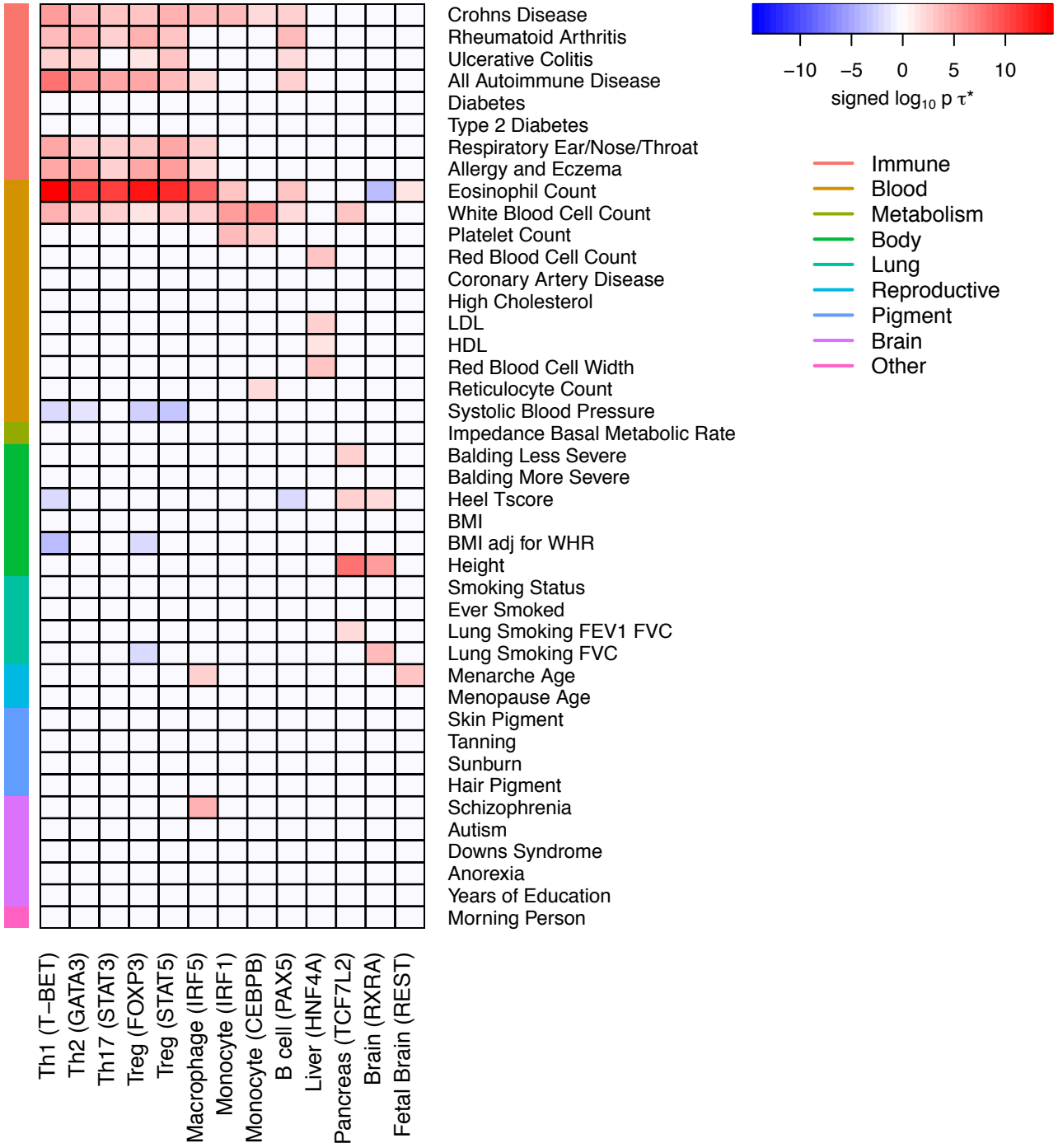
4A


RA h2 Enrichment
300
60
40
20
0
Treg (FOXP3) IMPACT
Other Annotations
2%
5%
0.04%
3%
1%
1%
3%
2%
3%
12%
13%
13%
13%
11%
3%
Treg IMPACT
Foxp3 Motif
Foxp3 ChIP−seq
Averaged Tracks
H3K4me3 (Th17 stim)
H3K4me3 (Treg)
H3K27ac (Th2)
H3K4me3 (Treg)
H3K27ac (Th17 stim)
T.4int8+.Th
T.4.Pa.BDC
T.4SP69+.Th
CD4Control
T.4.PLN.BDC
Super Enhancers

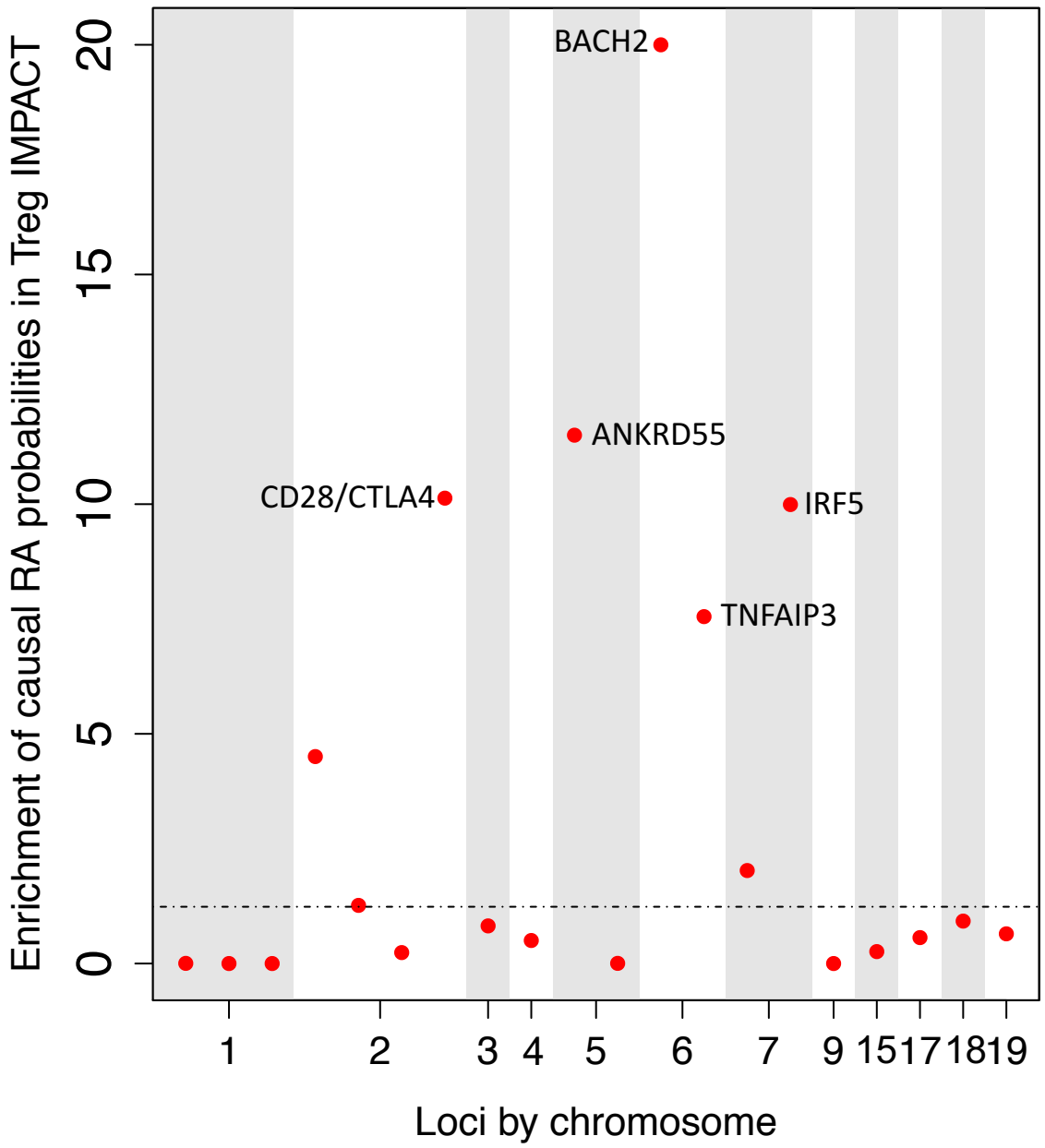
4B

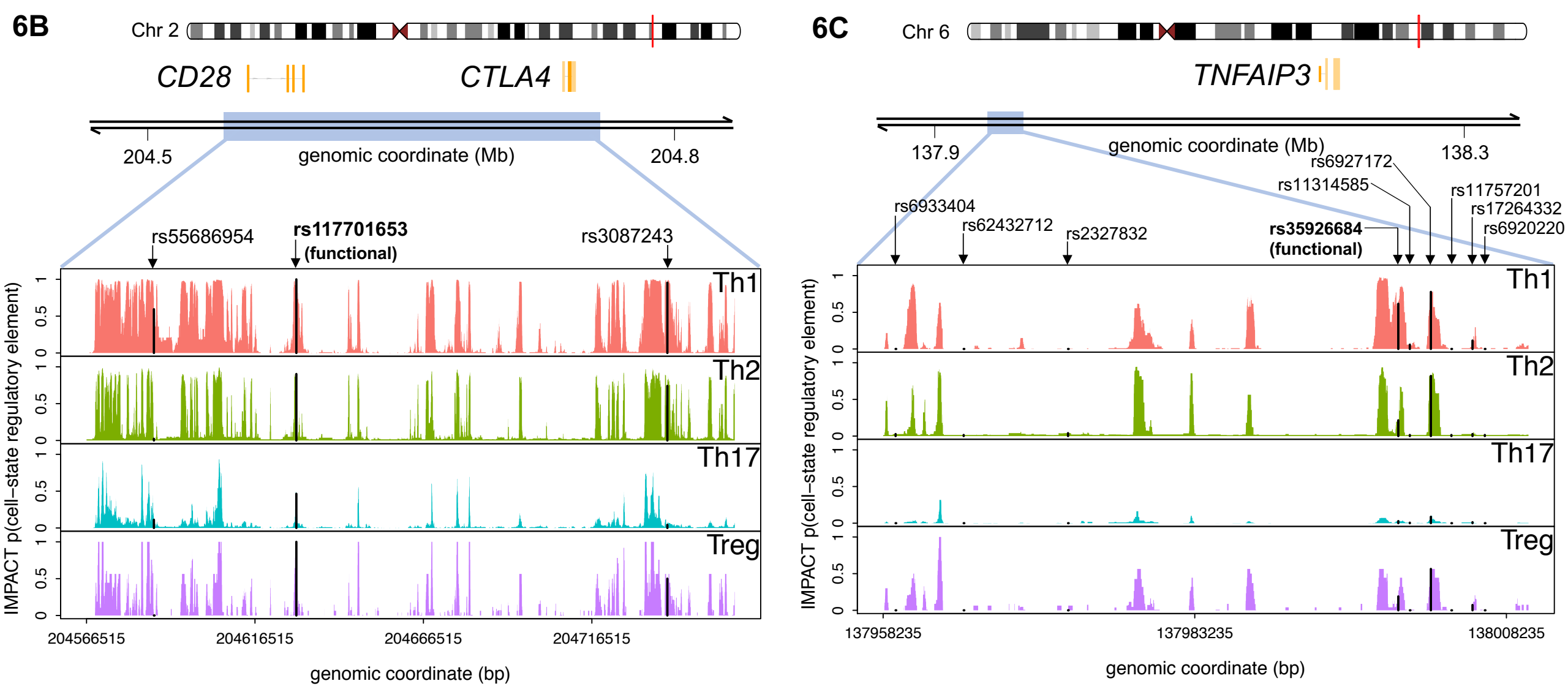

Annotation effect size τ*
6
4
2
0
−2
Independent
Conditional Treg IMPACT
Conditional Compared Annotation
Independent 95%CI
Conditional 95%CI
Foxp3 Motif
Foxp3 ChIP−seq
Averaged Tracks
H3K4me3 (Th17 stim)
H3K4me3 (Treg)
H3K27ac (Th2)
H3K4me3 (Treg)
H3K27ac (Th17 stim)
T.4int8+.Th
T.4.Pa.BDC
T.4SP69+.Th
CD4Control
T.4.PLN.BDC
Super Enhancers

5


Crohns Disease
Rheumatoid Arthritis
Ulcerative Colitis
All Autoimmune Disease
Diabetes
Type 2 Diabetes
Respiratory Ear/Nose/Throat
Allergy and Eczema
Eosinophil Count
White Blood Cell Count
Platelet Count
Red Blood Cell Count
Coronary Artery Disease
High Cholesterol
LDL
HDL
Red Blood Cell Width
Reticulocyte Count
Systolic Blood Pressure
Impedance Basal Metabolic Rate
Balding Less Severe
Balding More Severe
Heel Tscore
BMI
BMI adj for WHR
Height
Smoking Status
Ever Smoked
Lung Smoking FEV1 FVC
Lung Smoking FVC
Menarche Age
Menopause Age
Skin Pigment
Tanning
Sunburn
Hair Pigment
Schizophrenia
Autism
Downs Syndrome
Anorexia
Years of Education
Morning Person
Th1 (T–BET)
Th2 (GATA3)
Th17 (STAT3)
Treg (FOXP3)
Treg (STAT5)
Macrophage (IRF5)
Monocyte (IRF1)
Monocyte (CEBPB)
B cell (PAX5)
Liver (HNF4A)
Pancreas (TCF7L2)
Brain (RXRA)
Fetal Brain (REST)
−10
−5
0
5
10
signed $\log_{10}$ p τ^*
Immune
Blood
Metabolism
Body
Lung
Reproductive
Pigment
Brain
Other

6A


Enrichment of causal RA probabilities in Treg IMPACT
0
5
10
15
20
BACH2
ANKRD55
CD28/CTLA4
IRF5
TNFAIP3
1
2
3
4
5
6
7
9
15
17
18
19
Loci by chromosome


6B
Chr 2
CD28
CTLA4
204.5
genomic coordinate (Mb)
204.8
rs55686954
rs117701653
(functional)
rs3087243
Th1
Th2
Th17
Treg
IMPACT p(cell−state regulatory element)
204566515
204616515
204666515
204716515
genomic coordinate (bp)
6C
Chr 6
TNFAIP3
137.9
genomic coordinate (Mb)
138.3
rs6927172
rs11314585
rs11757201
rs17264332
rs6920220
rs6933404
rs62432712
rs2327832
rs35926684
(functional)
Th1
Th2
Th17
Treg
IMPACT p(cell−state regulatory element)
137958235
137983235
138008235
genomic coordinate (bp)