Rheumatoid arthritis heritability is concentrated in regulatory elements with CD4+ T cell-state-specific transcription factor binding profiles

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

Active regulatory elements within CD4+ T cells harbor disproportionate heritability (h2) for rheumatoid arthritis (RA). We hypothesized that regulatory elements specific to pathogenic CD4+ T cell-states better capture RA h2; however, defining these elements is challenging. To this end, we introduce IMPACT, a genome annotation strategy to identify cell-state-specific regulatory elements defined by key transcription factor binding profiles, learned from 398 chromatin and sequence annotations. Integrating IMPACT annotations of four CD4+ T cell-states with RA summary statistics of 38,242 Europeans and 22,515 East Asians, we observe that on average the top 5% of Treg predicted regulatory elements explain 85.7% (s.e. 19.4%, enrichment *p*<1.6e-05) of RA h2, and other cell-states explain a similar proportion. IMPACT captures RA h2 better than active CD4+ T cell regulatory elements, including super enhancers and specifically expressed genes (all *p*<0.05). IMPACT is generalizable to non-immune cell types and can identify other complex trait associated regulatory elements.

It is now well recognized that complex trait diseases have disproportionate heritability (h2) in subsets of genes and regulatory elements, particularly those that are specifically expressed in pathogenic cell and tissue types. For example, genetic studies of rheumatoid arthritis (RA), which is an autoimmune disease attacking synovial joint tissue leading to permanent joint damage and disability[1], have suggested a critical role by CD4+ T cells[2–10]. However, developing safe and effective therapeutics requires knowledge of the key cellular subsets, or cell-states, underlying disease risk and their regulatory elements. Naive CD4+ T cells may differentiate into memory T cells, and then into effector T cells Th(T helper)1, Th2, and Th17 and T regulatory (Treg) cells, requiring the action of a limited number of key transcription factors (TFs): T-bet or Stat4, Gata3 or Stat6, Stat3 or ROR-gamma-t, FoxP3 or Stat5, respectively[11]. Recently, we and others have identified CD4+ T cell-states significantly associated with RA: Tph (T peripheral helper)[12] and Th1 super-effector[13] cells.

We hypothesized that regulatory elements specific to pathogenic CD4+ T cell-states capture RA h2 better than cell-state-nonspecific CD4+ T cell elements. However, defining cell-state-specific gene regulatory elements is challenging; identification of active promoters and enhancers through open chromatin assays in the cell-state of interest will highlight both cell-state-specific and nonspecific regulatory elements.

To overcome this challenge, we take a two-step approach where we first choose a single cell-state-specific canonical TF and identify experimental binding regions. Then, IMPACT (Inference and Modeling of Phenotype-related ACtive Transcription) predicts TF occupancy at binding motifs by aggregating 386 cell-type-specific epigenomic features and 12 other sequence features in an elastic net logistic regression model (**Methods***: IMPACT Model*). Epigenomic features include histone mark ChIP-seq, ATAC-seq, DNase-seq and HiChIP (**Table S1**) assayed in immune, muscular, skeletal, brain, and other cell types, while sequence features include coding, intergenic, etc. From this regression we learn a TF binding chromatin profile, which we use to probabilistically define other genomic regions as cell-state-specific regulatory elements at nucleotide-resolution, distinguishing from nonspecific or non-regulatory elements (**Figure 1A**).

We used IMPACT to predict regulatory elements in four CD4+ T cell-states: Th1, Th2, Th17, and Treg, in each case selecting a single regulator based on availability of high quality genome-wide TF occupancy data (ChIP-seq): T-bet, Gata3, Stat3, and FoxP3, respectively (**Table S2**)[14–18]. IMPACT predicts TF occupancy at binding motifs with high accuracy (mean AUC 0.94 (s.e. 0.03), 10-fold cross validation performed on 80% of training data, AUC evaluated on the withheld 20%, **Figure 1B**, **Figure S1, Table S3**). We find that cell-state-specific chromatin features are often the most important predictor variables (**Figure S2**). IMPACT performs significantly better than naive strategies that predict TF motif binding based on active promoters (H3K4me3 ChIP-seq) or cell-state-specific open chromatin (DNase-seq) (all p<1.5e-15, **Figure 1B**). Restricting feature categories in the model, such as removing cell-state-specific features or features of the same epigenomic assay, revealed that only Th2 prediction accuracy significantly declines from exclusion of Th2-specific features (**Table S3**). While experimental binding data from ChIP-seq can only identify broad regions of binding signal (~300 bp), IMPACT provides nucleotide resolution, illustrated by predictions at canonical TF-specific targets (**Figure 1C**, **Figure S3**). We find that IMPACT annotations are highly correlated with one another and with particular epigenomic training annotations, such as Th2 open chromatin (**Figure S4**). In particular, Th17 IMPACT is correlated with H3K4me3 annotations and T-bet IMPACT with H3K4me1 annotations, reflecting CD4+ T promoter co-localization.

To test our hypothesis that IMPACT captures RA h2, we used stratified LD (linkage disequilibrium) score regression (S-LDSC)[9] with publicly available European (EUR, N = 38,242)[9] and East Asian (EAS, N = 22,515)[19] RA GWAS summary statistics to partition the common SNP h2 of RA, which is estimated to be about 18% for EUR and 21% for EAS (**Methods***: S-LDSC*). We computed CD4+ T cell-state-specific (Th1, Th2, Th17, and Treg) regulatory element probabilities at all common SNPs (MAF ≥ 0.05) genome-wide, excluding the major histocompatibility complex (MHC) due to its outlier LD structure. We then created S-LDSC models, composed of one or more IMPACT annotations and a set of 69 baseline annotations, controlling for cell-type-nonspecific functional, LD-related, and MAF associations, referred to as the baseline-LD model[20]. We found that each CD4+ T cell-state-specific IMPACT annotation is significantly enriched with RA h2 in both populations (average enrichment = 20.05, all p<1.9e-04, **Figure 2A**, **Table S4**). The standardized annotation effect size, τ^*, is defined as the proportionate change in per-SNP h2 associated with a one standard deviation increase in the value of the annotation. The τ^* of an annotation that captures RA h2 will be significantly greater than zero and the greater the τ^*, the more that annotation captures RA h2. τ^* is significantly positive for all CD4+ IMPACT annotations (all p<2.1e-03, **Figure 2B,** conditional analysis in **Figure S5**), demonstrating that IMPACT captures RA h2 not explained by the baseline-LD annotations. We then created annotations consisting of the top 5% of regulatory SNPs according to each IMPACT cell-state and show that the Treg annotation explains the greatest proportion of RA h2, 84.0% (s.e. 17.1%) in Europeans and 88.6% (s.e. 22.3%) in East Asians (**Figure 2C**). Lastly, we show that while IMPACT annotations are strongly correlated with one another, they are weakly correlated with baseline-LD annotations (**Figure 2D**), consistent with significantly positive τ* in **Figure 2B**.

We hypothesized that the CD4+ T cell-state IMPACT annotations would capture RA h2 better than other T cell functional annotations. To this end, we compared the enrichments and τ* of each CD4+ IMPACT annotation to that of various functional annotations, in EUR: TF binding motifs, genome-wide TF occupancy (ChIP-seq), an annotation that assigns each SNP a value proportional to the number of IMPACT epigenomic features it overlaps (Averaged Tracks), the five largest τ* CD4+ T cell-specific histone mark annotations[9], the five largest τ* CD4+ T cell-specifically expressed gene sets and their regulatory elements[6], and T cell super enhancers[21] (**Figure 3, Figure S6**). Then, we computed the annotation τ* while pairwise conditioned on each other and on baseline-LD. We found that, even when conditioned on annotations with highly significant enrichment, the CD4+ Treg and Th2 IMPACT τ* are significantly positive (all *p*<0.01) and are more significant, thereby capturing RA h2 better than all other annotations, except H3K27ac in Th2 cells.

We next hypothesized that IMPACT would inform functional variant fine-mapping. Using a GWAS of 11,475 European RA cases and 15,870 controls[22], an independent study from the European summary statistics[23] used in our h2 analyses, our group recently fine-mapped a subset of 20 RA risk loci, each with a manageable number of putatively causal variants, and created 90% credible sets of these SNPs[24]. We computed the enrichment of fine-mapped causal probabilities across these 20 loci in the top 1% of our CD4+ T cell-state-specific IMPACT annotations (**Methods**: *Posterior Probability Enrichment*). We found that only the Treg annotation is significantly enriched (2.87, permutation p<8.6e-03, **Figure 4A, Table S5**), suggesting this annotation may be useful to prune putatively causal RA variants. Furthermore, we observe uniquely high Treg enrichment in the *BACH2* and *IRF5* loci (16.2 and 8.1, respectively), suggesting putatively causal SNPs in these loci may function in a Treg-specific context.

In related work, our group observed both differential binding of CD4+ nuclear extract via EMSA and differential enhancer activity via luciferase assays at two credible set SNPs, narrowing down the list of putatively causal variants in the *CD28/CTLA4* and *TNFAIP3* loci[24]. We observed that both variants with functional activity were located at predicted IMPACT regulatory elements, suggesting that IMPACT may be used to narrow down credible sets to reduce the amount of experimental follow up. First, at the *CD28/CTLA4* locus, we observe high probability regulatory elements across the four CD4+ T cell-states at the functional SNP rs117701653 and lower probability regulatory elements at other credible set SNPs rs55686954 and rs3087243 (**Figure 4B**). Second, at the *TNFAIP3* locus, we observe high probability regulatory elements at the functional SNP rs35926684 and other credible set SNP rs6927172 (**Figure 4C**) and do not predict regulatory elements at the other 7 credible set SNPs. The CD4+ Th1 specific regulatory element at rs35926684 suggests that this SNP may alter gene regulation specifically in Th1 cells and hence, we suggest any functional follow-up be done in this cell-state. Few other credible set SNPs in the other 18 fine-mapped loci have high IMPACT cell-state-specific regulatory element probabilities (**Figure S7**). We note that disease-relevant IMPACT functional annotations may be integrated with existing functional fine mapping methods, like PAINTOR[25], to assign causal posterior probabilities to variants.

We then applied our CD4+ T IMPACT annotations to 41 additional polygenic traits[20,26] and observed significantly positive τ^* for autoimmune traits, Crohn's and ulcerative colitis (both p<0.04), and several blood traits, eosinophil and white blood cell counts (both p<0.02), but not for non-immune-mediated traits (**Figure 5, Table S6**). For comparison, we built an HNF4A (hepatocyte nuclear factor 4A) IMPACT annotation to target LDL and HDL, liver-associated traits[27]. As expected, HNF4A IMPACT τ^* is not significant for autoimmune-mediated traits, but is significantly positive for LDL and HDL (both p<0.02), suggesting that IMPACT can find complex trait associated regulatory elements specific to a range of cell types.

In summary, IMPACT predicts cell-state-specific regulatory elements based on chromatin profiles of experimental TF binding and may help elucidate complex trait biology. First, we observed highly significant enrichments of RA h2 in IMPACT regulatory elements and explain the majority of RA h2 with just the top 5%. Second, we found that IMPACT annotations capture RA h2 better than most CD4+ T cell functional annotations. Lastly, we have briefly shown another utility of IMPACT to narrow down credible sets to reduce functional follow-up. We recognize several limitations to our work: 1) we have not experimentally validated the activity of any of our predicted regulatory elements, 2) predicted regulatory elements are limited to genomic regions that have been epigenetically assayed, 3) while IMPACT models appear unique between cell-states in terms of which epigenomic features best indicate TF motif binding, predicted regulatory element probabilities across cell-states are highly correlated due to the correlated nature of epigenomic features. In light of these limitations, IMPACT is an emerging strategy for identifying trait associated regulatory elements and generating hypotheses about the cell-states in which variants may be functional, motivating the need to develop therapeutics that target specific disease-driving cell-states.

Methods

Rationale

To explore if we could predict cell-state-specific regulatory elements, we selected RA and CD4+ T cells as the trait and cell type for our model because multiple lines of evidence support the role of CD4+ T cells in the pathogenesis of RA. First, classical HLA typing studies implicated the MHC region[2], more recently fine-mapped to the individual amino acid residues of HLA-DRB1[3], which presents antigens to CD4+ T cells. Second, we[7,8] and others[9,10] have demonstrated that there is an enrichment of RA causal variation in active regulatory elements of CD4+ T cells. Similarly, genes within RA risk loci are enriched for specific expression in CD4+ T cells[5] and genes expressed specifically within CD4+ T cell subpopulations are enriched for RA h2[9]. Finally, recent fine-mapping efforts have identified putatively causal variants within the *CD28*/*CTLA4* and *TNFAIP3* loci that have differential enhancer effects *in-vivo*[24].

Data

Genome-wide Annotation Data. We obtained publicly available genome-wide annotations in a broad range of cell types for the GRCh37 (hg19) assembly. The accession numbers and or file names for features downloaded from NCBI, Blueprint, Roadmap, and chromHMM are listed in **Table S1**. Features from Finucane et al 2015[9] are labeled as they were in supplemental tables of this study. Cell-type-specific annotation types include ATAC-seq, DNase-seq, FAIRE-seq, HiChIP, polymerase and elongation factor ChIP-seq, and histone modification ChIP-seq. Sequence annotations were downloaded from UCSC's publicly available bedfiles and include conservation, exons, introns, intergenic regions, 3'UTR, 5'UTR, promoter-TSS, TTS, and CpG islands). All genome-wide feature data, except conservation, is represented in standard 6-column bedfile format. Conservation is represented in bedgraph format, in which the average score is reported for each 1024 bp window genome-wide.

TF ChIP-seq data. We determined genome-wide TF occupancy from publicly available ChIP-seq (**Table S2**) of the four key regulators (T-bet[14,15], Gata3[16], Stat3[17], and FoxP3[18]) assayed in their respective primary cell-states Th1, Th2, Th17, and Tregs. ChIP-seq peaks were called by macs[28] [v1.4.2 20120305] (FDR<0.05).

IMPACT Model

We build a model that predicts TF binding on a motif by learning the epigenomic profiles of the TF binding sites. We use logistic regression to model the log odds of TF binding based on a linear combination of the effects β_j of the *j* epigenomic features, where β_0 is an intercept.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j$$

From the log odds, which ranges from negative to positive infinity, we compute the probability of TF binding, ranging from 0 to 1.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j)}}$$

We use an elastic net logistic regression framework implemented by the *cv.glmnet* R [v1.0.143] package[29], 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.

Training IMPACT. We train IMPACT on gold standard regulatory and non-regulatory elements of a particular TF, meaning that there is one IMPACT model per TF/cell-state pair. To build the regulatory class, we scanned the TF ChIP-seq peaks, mentioned above, for matches to the TF-specific binding motif, using HOMER[30] [v4.8.3]. Each match must receive a sequence similarity score greater than or equal to the threshold provided by the PWMs (position weight matrices) in the Jaspar database (**Table S2**). We only scan for the TF motif of the corresponding TF ChIP-seq dataset, e.g. we only looked for T-bet motifs in the T-bet ChIP-seq data. We retained the highest scoring motif match for each ChIP-seq peak and used the genomic coordinates of the center two nucleotides to create a *GenomicRanges* object in R. For the T-bet data, there were

10,086 such ranges, 1,224 for Gata3, 1,181 for Stat3, and 1,005 for FoxP3. For every run of IMPACT, 1,000 regulatory ranges are randomly selected, labeled with a 1, to train the model. Controlling the number of ranges used will standardize the logistic regression output such that predictions and model fits will be more comparable between cell-state models.

To build the non-regulatory class, we scanned the entire genome for TF motif matches, again using HOMER, and selected motif matches with no overlap with that TF's ChIP-seq peaks, e.g. we scan for T-bet motifs, and only retain regions not overlapping T-bet ChIP-seq peaks. We do not check for overlap with other TF (i.e. Gata3, Stat3, 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, for each region we annotate, we iterate through all genome-wide features, asking if there is overlap. In the distal case, we look for feature overlap of a more distal nucleotide (i.e. 1,000 base pairs up *or* downstream, such that overlap at either distal position will count and we do not distinguish between the upstream or downstream overlap in the feature matrix). Our rationale is that although a nucleotide may not intersect a particular feature, it may be informative to know that there is one nearby. The upstream and downstream distal coordinates (parameter set to 1,000 bp) are computed by subtracting or adding, respectively, the parameter value to these coordinates. If the computed distal coordinate is negative, the value is replaced with 1. If the computed distal coordinate is larger than the length of the chromosome, the value is replaced with the length of the chromosome. We set our distal parameter to 1,000 bp (**Table S3**) and do not check for overlap of any additional nucleotides within this distance. The feature matrix, on which the model is trained, may look like the following with dimensions (11,000 rows by 2+2X columns):

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

We applied IMPACT genome-wide to assign nucleotide-resolution regulatory probabilities, using the model β learned from the elastic net logistic regression CV. Model β from the run of IMPACT used to produce genome-wide annotations as well as the averaged model β over 10 runs can be found in **Table S1**. We show the largest magnitude β for each cell-state IMPACT model in **Figure S2**.

Stratified-LD Score Regression (S-LDSC)

Genome-wide association data. We collected RA GWAS summary statistics[23] for 38,242 European individuals, combined cases and controls, and 22,515 East Asian individuals, comprised of 4,873 RA cases and 17,642 controls[19]. Reference SNPs, used to estimate European LD scores, were the set of 9,997,231 SNPs with minor allele count greater or equal than five in a set of 659 European samples from phase 3 of 1000 Genomes Projects[31]. The regression coefficients were estimated using 1,125,060 HapMap3 SNPs and heritability was partitioned for the 5,961,159 reference SNPs with MAF ≥ 0.05. Reference SNPs, used to estimate East Asian LD scores, were the set of 8,768,561 SNPs with minor allele count greater or equal than five in a set of 105 East Asian samples from phase 3 of 1000 Genomes Projects[31]. The regression coefficients were estimated using 1,026,051 HapMap3 SNPs and heritability was partitioned for the 5,469,053 reference SNPs with MAF ≥ 0.05. Frequency and weight files (1000G EUR phase3, 1000G EAS phase3) are publicly available and may be found in our URLs.

Methodology. We apply S-LDSC[9] [v1.0.0], a method developed to partition polygenic trait heritability by one or more functional annotations, to quantify the contribution of IMPACT cell-state-specific regulatory annotations to RA and other autoimmune disease heritability. We annotate common SNPs (MAF ≥ 0.05) with multiple cell-state-specific IMPACT models, assigning a regulatory element score to each variant. Then S-LDSC was run on the annotated SNPs to compute LD scores. Here, the two statistics we use to evaluate each annotation's contribution to disease heritability are enrichment and standardized effect size (τ*).

If a_{cj} is the value of annotation *c* for SNP *j*, we assume the variance of the effect size of SNP *j* depends linearly on the contribution of each annotation *c*:

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

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

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

Considering the expectation of X_j^2 and following the derivation from Gazal et al 2017[20],

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

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

where *N* is the sample size of the GWAS, $l(j,c)$ is the LD score of SNP *j* with respect to annotation *c*, and r_{jk}^2 is the true, e.g. population-wide, genetic correlation of SNPs *j* and *k*. Since τ_c is not comparable between annotations or traits, Gazal et al 2017[20] defines τ_c^* as the per-annotation standardized effect size, a function of the standard deviation of the annotation *c*, $sd(c)$, the trait-specific SNP-heritability estimated by LDSC h_g^2, and the total number of reference common SNPs used to compute h_g^2, *M* = 5,961,159 in EUR and 5,469,053 in EAS:

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

We define enrichment of an annotation as the proportion of heritability explained by the annotation divided by the average value of the annotation across the *M* common (MAF

≤ 0.05) SNPs. Enrichment may be computed for binary or continuous annotations according to the equation below, where $h_g^2\,(c)$ is the h2 explained by SNPs in annotation *c*.

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

Enrichment does not quantify effects that are unique to a given annotation, whereas τ* does.

Each S-LDSC analysis involves conditioning IMPACT annotations on 69 baseline annotations, referred to as the baseline-LD model, consisting of 53 cell-type-nonspecific annotations[9], which include histone marks and open chromatin, 10 MAF bins, and 6 LD-related annotations[20] to assess if functional enrichment is cell-type-specific and to control for the effect of MAF and LD architecture. Consistent inclusion of MAF and LD associated annotations in the baseline model is a standard recommended practice of S-LDSC. When conditionally comparing two annotations, say A and B, in a single S-LDSC model, the two annotations may have similar enrichments if they are highly correlated. However, the τ* for the annotation with greater true causal variant membership will be larger and more statistically significant (e.g. > 0). Specifically, a τ* of 0, means that the annotation does not change per-SNP h2. A strongly negative $\tau *$ means that membership to the categorical annotation decreases per-SNP h2, while a strongly positive τ* means that membership to the annotation increases per-SNP h2. The significance of τ* is computed based on a test of how different from 0 the $\tau *$ is.

MHC exclusion. We note that S-LDSC excludes the MHC (major histocompatibility complex) due to its extremely high gene density and outlier LD structure, which is thought to be the strongest contributor to RA disease h2[13]. However, our work supports the notion that there is an undeniably large amount of RA h2 located outside of the MHC.

Posterior probability enrichment

Previous work from our group aimed to define the most likely causal RA variant for each locus harboring a genome-wide significant variant[24]. To this end, posterior probabilities were computed with the approximate Bayesian factor (ABF), assuming one causal variant per locus. The posteriors were defined as:

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

where *i* is the *ith* variant, and *n* is the total number of variants in the locus. As such, the ABF over all variants in a locus sum to 1. Then, for each of the defined 20 RA-associated loci[24], we computed the enrichment of high posterior probabilities in the top 1% of cell-state-specific IMPACT regulatory elements (**Table S5**). For each RA-associated locus *l,*

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

where $P_c(i)$ is the posterior causal probability of SNP *i*, such that *i* belongs to the top 1% of the cell-state-specific IMPACT annotation *c*, M_l is the number of SNPs in locus *l* for which we have a computed posterior probability. The denominator formulates the null hypothesis that each SNP in a locus is equally causal. We computed the average of these enrichment values over the 20 RA-associated loci. A permutation distribution was calculated by computing an average enrichment value over these 20 loci, in 1,000 trials, in which random posterior probabilities (of the same quantity M_l) were selected. The permutation *p*-value was calculated by comparing the quantile value of the IMPACT enrichment to the assumed-normal permutation distribution defined by its mean and standard deviation, using the *pnorm* function in R.

Caveats

This approach might be easily applied to a wide-range of other diseases, recognizing several important caveats. First, it relies on choosing a key regulator TF with a known consensus binding motif. While for CD4+ T cells there is extensive literature of cell-states and their key regulatory drivers, this may not be readily available for all cell types. Second, 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. Most immune populations are difficult to sort ex-vivo, however IMPACT models 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 relative to the other cells, and as such 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 of this 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

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.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Firestein, G. S. Evolving concepts of rheumatoid arthritis. *Nature* **423,** 356–361 (2003).
2. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44,** 291–296 (2012).
3. Terao, C. *et al.* Genetic landscape of interactive effects ofHLA-DRB1alleles on susceptibility to ACPA(+) rheumatoid arthritis and ACPA levels in Japanese population. *J. Med. Genet.* **54,** 853–858 (2017).

4. Terao, C., Raychaudhuri, S. & Gregersen, P. K. Recent Advances in Defining the Genetic Basis of Rheumatoid Arthritis. *Annu. Rev. Genomics Hum. Genet.* **17,** 273–301 (2016).

5. Hu, X. *et al.* Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets. *Am. J. Hum. Genet.* **89,** 496–506 (2011).

6. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50,** 621–629 (2018).

7. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet.* **97,** 139–152 (2015).

8. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45,** 124–130 (2013).

9. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

10. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2014).

11. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of Effector CD4 T Cell Populations. *Annu. Rev. Immunol.* **28,** 445–489 (2010).

12. Rao, D. A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* **542,** 110–114 (2017).

13. Fonseka, C. Y., Rao, D. A. & Raychaudhuri, S. Leveraging blood and tissue CD4+ T cell heterogeneity at the single cell level to identify mechanisms of disease in

rheumatoid arthritis. *Curr. Opin. Immunol.* **49,** 27–36 (2017).

14. Soderquest, K. *et al.* Genetic variants alter T-bet binding and gene expression in mucosal inflammatory disease. *PLOS Genet.* **13,** e1006587 (2017).

15. Hertweck, A. *et al.* T-bet Activates Th1 Genes through Mediator and the Super Elongation Complex. *Cell Rep.* **15,** 2756–2770 (2016).

16. Gustafsson, M. *et al.* A validated gene regulatory network and GWAS identifies early regulators of T cell–associated diseases. *Sci. Transl. Med.* **7,** 313ra178-313ra178 (2015).

17. Tripathi, S. K. *et al.* Genome-wide Analysis of STAT3-Mediated Transcription during Early Human Th17 Cell Differentiation. *Cell Rep.* **19,** 1888–1901 (2017).

18. Schmidl, C. *et al.* The enhancer and promoter landscape of human regulatory and conventional T-cell subpopulations. *Blood* **123,** e68–e78 (2014).

19. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50,** 390–400 (2018).

20. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49,** 1421–1427 (2017).

21. Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520,** 558–562 (2015).

22. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44,** 1336–1340 (2012).

23. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).

24. Westra, H.-J. *et al.* Fine-mapping identifies causal variants for RA and T1D in

DNASE1L3, SIRPG, MEG3, TNFAIP3 and CD28/CTLA4 loci. *bioRxiv* 151423 (2017). doi:10.1101/151423

25. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33,** 248–255 (2017).

26. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0148-2

27. Gertz, J. *et al.* Distinct Properties of Cell-Type-Specific and Shared Transcription Factor Binding Sites. *Mol. Cell* **52,** 25–36 (2013).

28. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

29. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** 1–22 (2010).

30. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).

31. Gibbs, R. A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

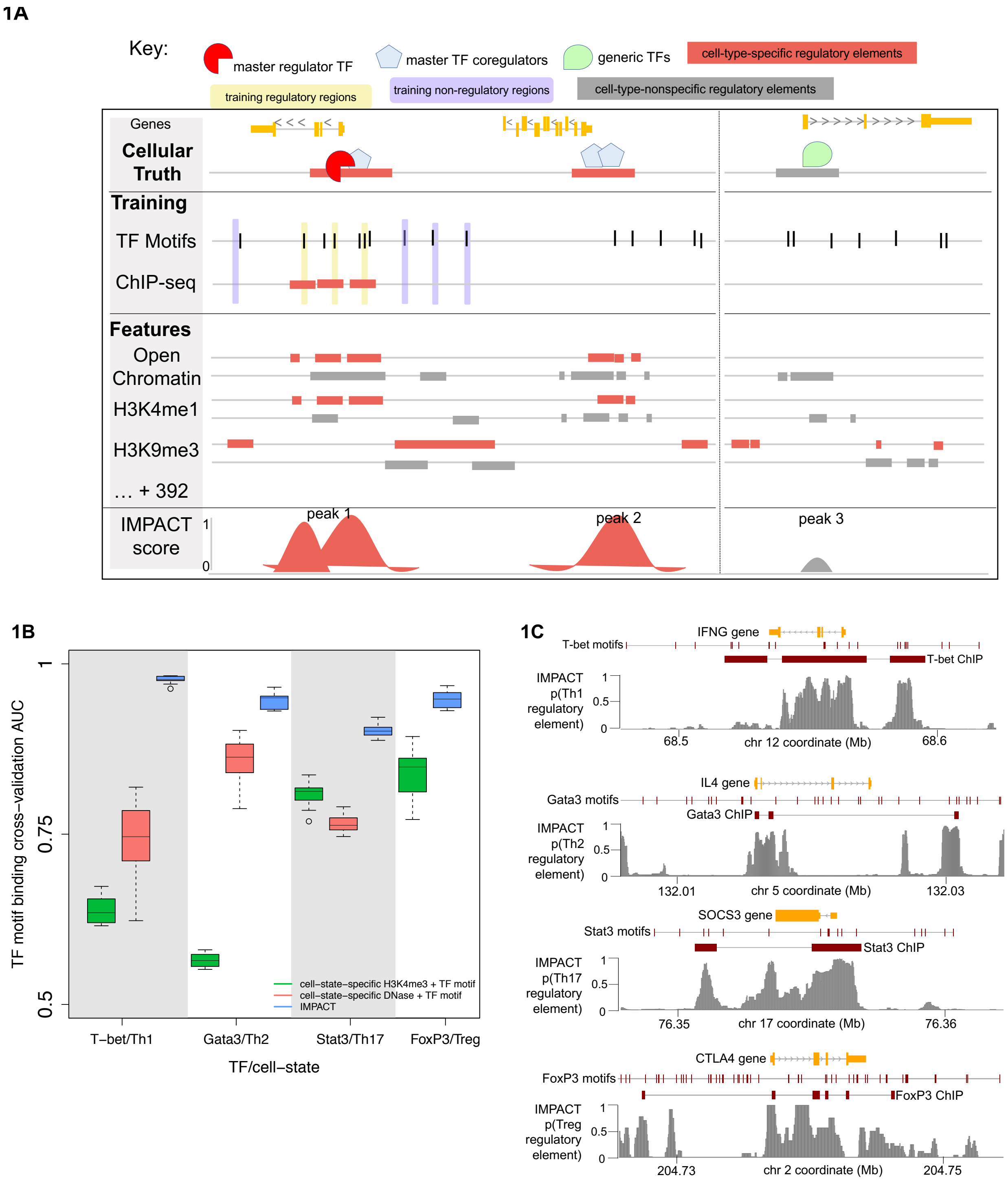

Figure 1.

(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-nonspecific DHS and H3K4me1 are less informative. IMPACT also learns that H3K9me3 is a strong predictor of non-regulatory elements. IMPACT is expected to re-identify regulatory elements marked by master TF binding (ChIP-seq and motif) [peak 1], while also identifying others where the chromatin profile is similar, perhaps representing cell-type-specific transcriptional processes [peak 2]. IMPACT is not expected to predict regulation at cell-type-nonspecific elements [peak 3], such as promoters of generic housekeeping genes, assuming these elements have different chromatin profiles. **(b)** IMPACT significantly outperforms cell-state-specific active promoter (H3K4me3, green) and open chromatin (DNase, red) annotations in predicting TF binding on a motif over 10 trials measured by computing the average ROC AUC (receiver operator characteristic area under the curve). As there is no Treg DNase data, there is no comparison AUC distribution. **(c)** Cell-state-specific regulatory element IMPACT predictions for canonical target genes of T-bet, Gata3, Stat3, and FoxP3.

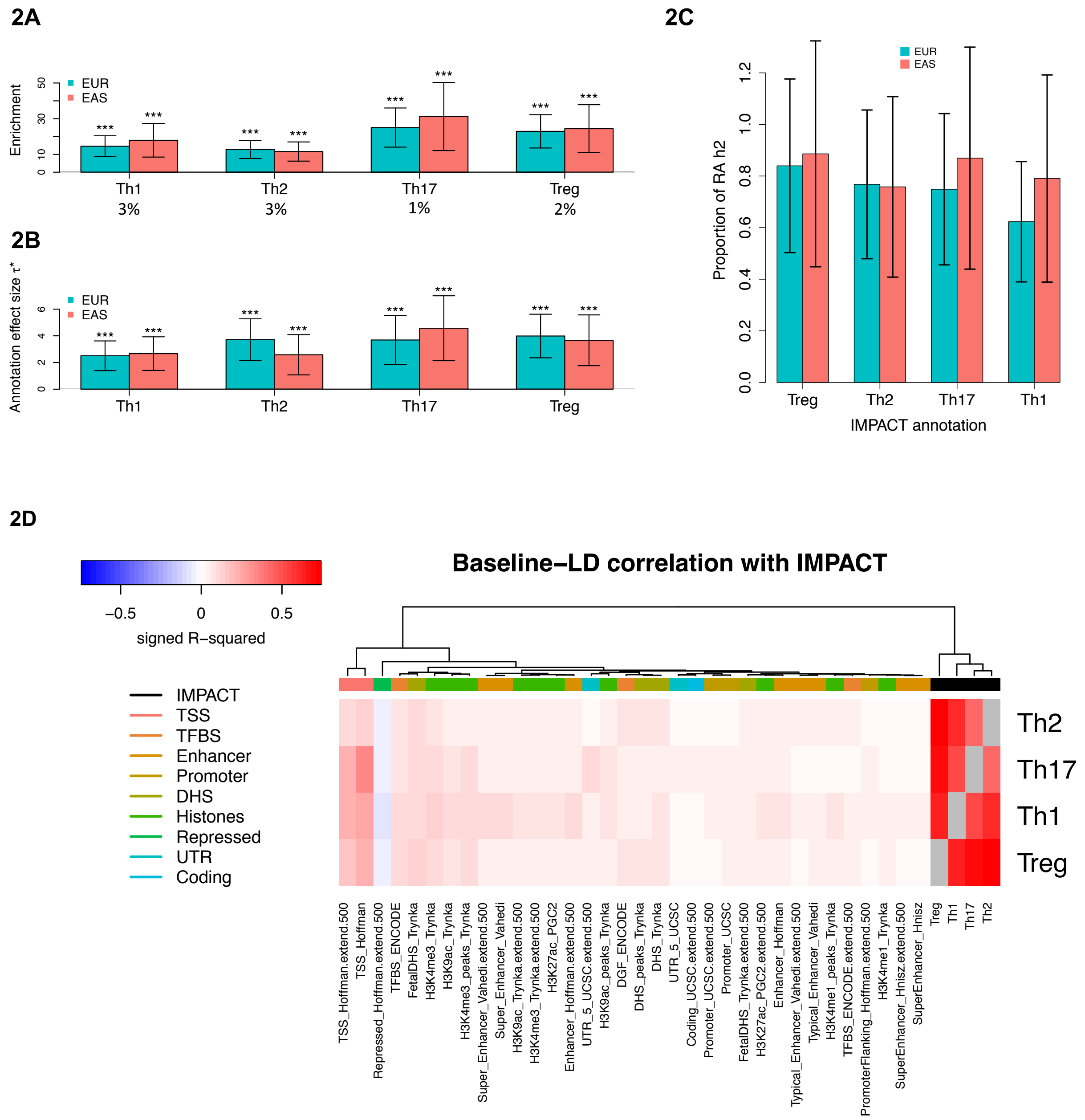

Figure 2.

(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. **(d)** 2D hierarchical clustering of pairwise signed Pearson R-squared correlations between CD4+ IMPACT annotations and most strongly correlated baseline-LD annotations.

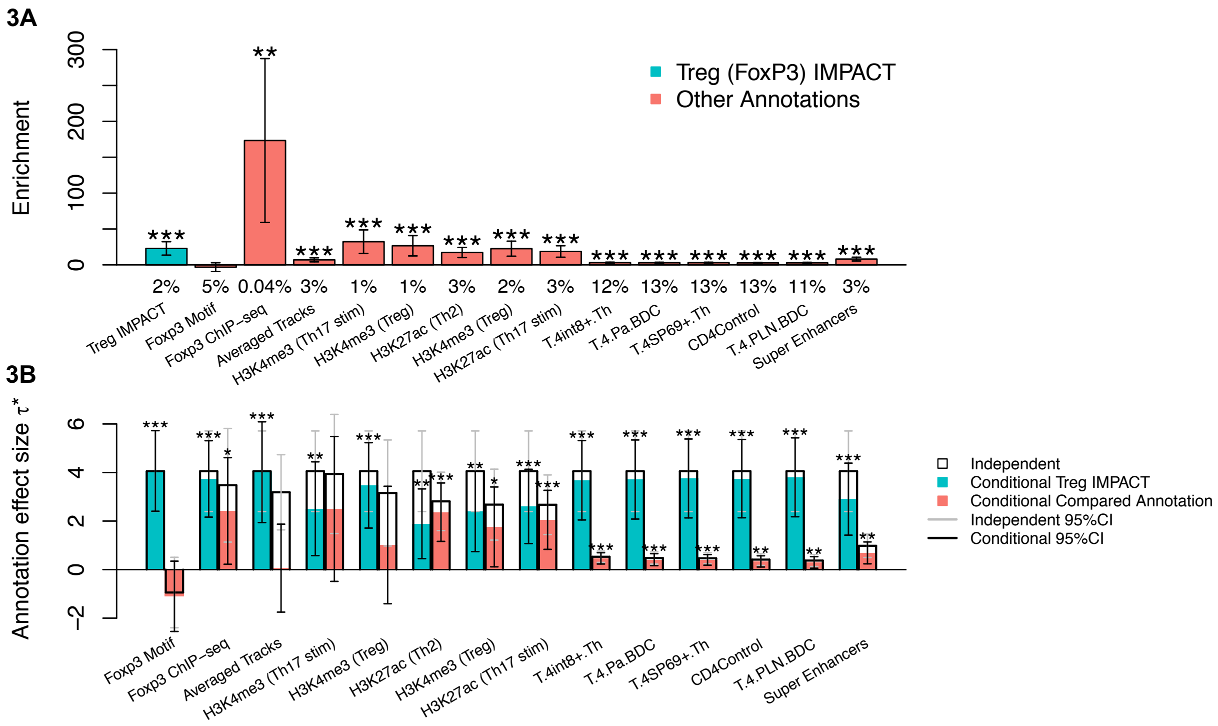

Figure 3.

(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[9], the top 5, in terms of independent τ*, cell-type-specifically expressed gene sets (URLS)[6], and T cell super enhancers[21]. **(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.

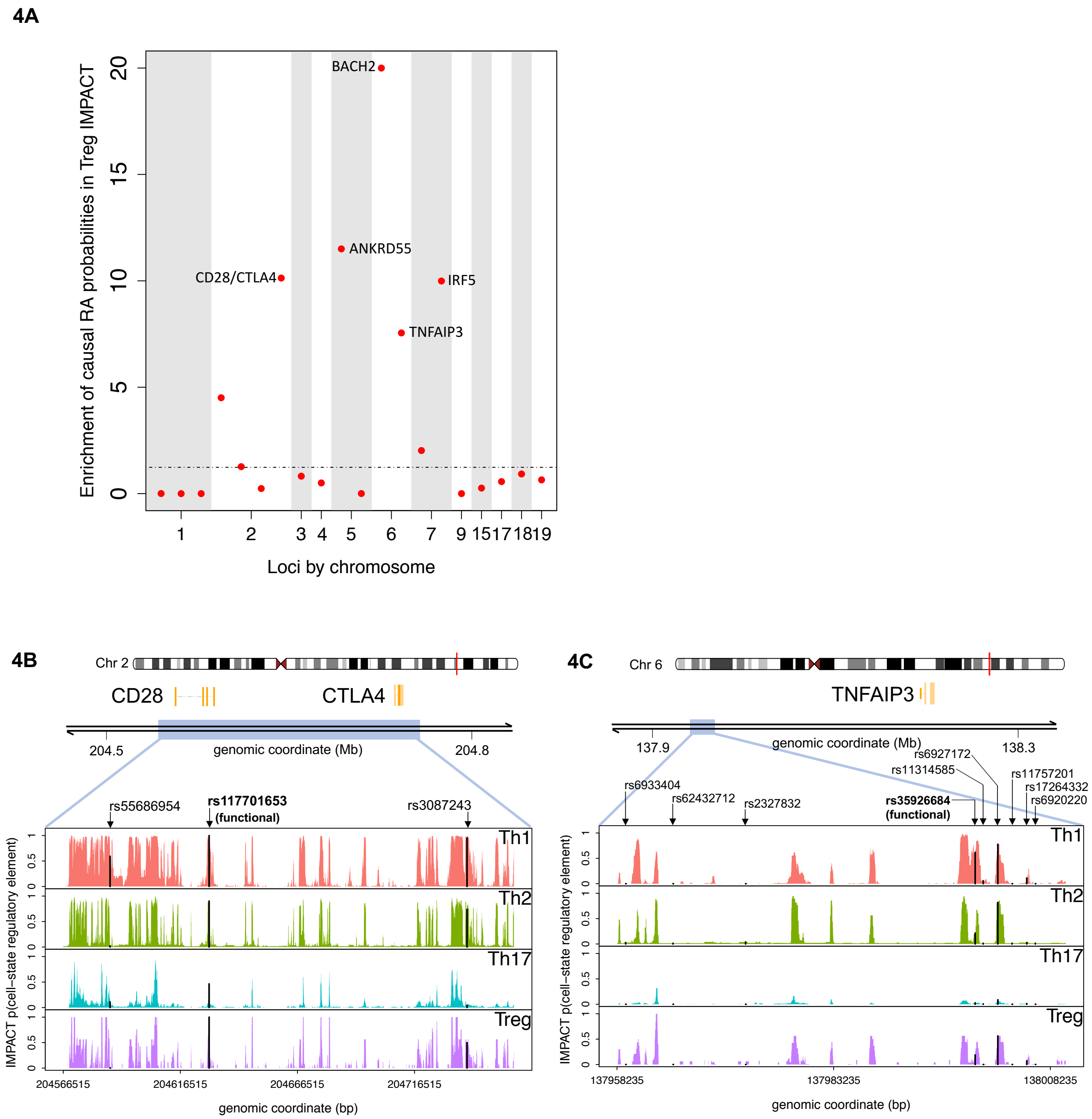

Figure 4.

(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

RA SNPs. For two RA-associated loci, *CTLA4/CD28* and *TNFAIP3*, we examine the putatively causal SNP with experimentally validated differential enhancer activity (bolded) and other 90% credible set SNPs (unbolded)[24]. IMPACT scores at these SNPs are highlighted with a black line. **(b)** We observe high probability IMPACT regulatory elements in all four CD4+ T cell-states for the functional SNP rs117701653 in the *CD28/CTLA4* locus. **(c)** We observe a high probability Th1-specific IMPACT regulatory element for the functional SNP rs35926684 in the *TNFAIP3* locus.

5

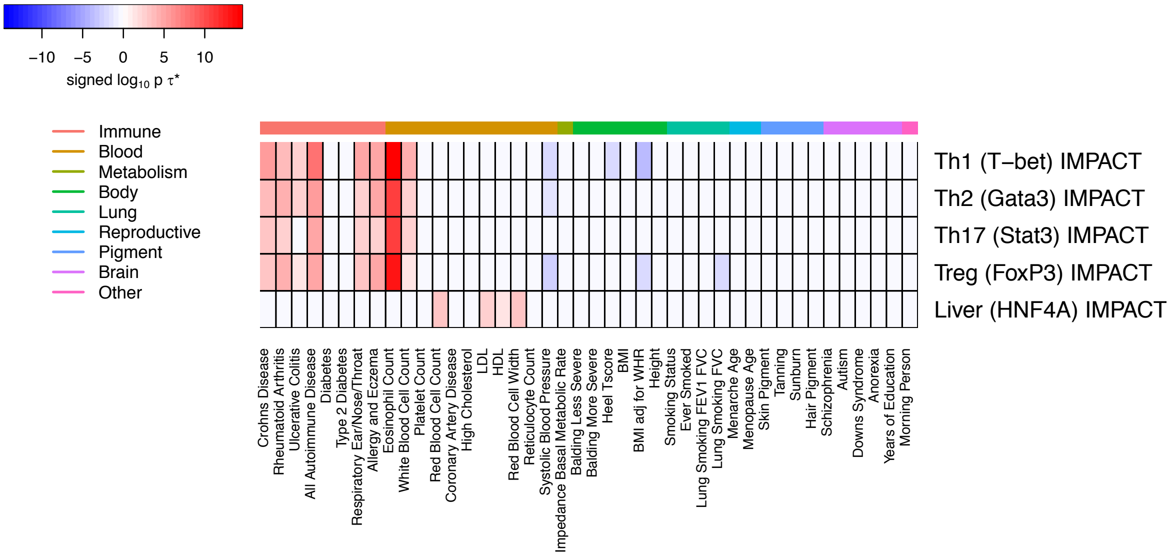

Figure 5.

Signed log10 p-values of τ* for 42 traits across the four CD4+ T IMPACT annotations and Liver (HNF4A) IMPACT for comparison. Both sets of annotations capture h2 in distinct sets of complex traits, with significantly positive τ* for expectedly related traits. Color shown only if p-value of τ* < 0.025.