WhichTF is dominant in your open chromatin data?

Yosuke Tanigawa\textsuperscript{1,*}, Ethan S. Dyer\textsuperscript{2,3,7,*}, Gill Bejerano\textsuperscript{1,4,5,6,+}

1. Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, CA, USA.
2. Stanford Institute for Theoretical Physics, Stanford University, Stanford, CA, USA.
3. Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD, USA.
4. Department of Developmental Biology, Stanford University, Stanford, CA, USA.
5. Department of Computer Science, Stanford University, Stanford, CA, USA.
6. Department of Pediatrics, Stanford University School of Medicine, Stanford University, Stanford, CA, USA.
7. Current address: Google, Mountain View, CA, USA.

* These authors contributed equally and ordered by age
+ Corresponding author (G.B. bejerano@stanford.edu)

Abstract

We present WhichTF, a novel computational method to identify dominant transcription factors (TFs) from chromatin accessibility measurements. To rank TFs, WhichTF integrates high-confidence genome-wide computational prediction of TF binding sites based on evolutionary sequence conservation, putative gene-regulatory models, and ontology-based gene annotations. Applying WhichTF, we find that the identified dominant TFs have been implicated as functionally important in well-studied cell types, such as NF-κB family members in lymphocytes and GATA factors in cardiac tissue. To distinguish the transcriptional regulatory landscape in closely related samples, we devise a differential analysis framework and demonstrate its utility in lymphocyte, mesoderm developmental, and disease cells. We also find TFs known for stress response in multiple samples, suggesting routine experimental caveats that warrant careful consideration. WhichTF yields biological insight into known and novel molecular mechanisms of TF-mediated transcriptional regulation in diverse contexts, including human and mouse cell types, cell fate trajectories, and disease-associated tissues.
Introduction

Transcription factors (TFs) are the master regulators of development. They define, refine, and can even divert cellular trajectories. TFs perform these important tasks by binding to specific DNA sequences in open chromatin, where they recruit additional cofactors and together modulate expression of downstream genes. TFs regulate biological processes in healthy adult tissues, and mutations to both TF genes and their genomic binding sites have been linked with human disease\textsuperscript{1,2}.

The advent of next generation sequencing has paved the way for chromatin immunoprecipitation followed by sequencing (ChIP-seq)-based methods for the discovery of genome-wide loci where a given TF binds DNA in a given cell population\textsuperscript{3}. Tools developed for the analysis of ChIP-seq data, such as GREAT\textsuperscript{4} (Gene Regulatory Enrichment of Annotations Tool), have discovered and leveraged a compelling phenomenon: when a TF is functionally important for the progression of a certain process, such that its perturbation leads to the disruption of this process, the binding sites for this TF are often highly enriched in the gene regulatory domains of the “downstream” target genes that drive this process\textsuperscript{4}.

TFs work in different combinations to enact a vast repertoire of cellular fates and responses\textsuperscript{5}. Between 1,500-2,000 TFs are thought to be encoded in the human genome\textsuperscript{1}.

Performing ChIP-seq for more than a handful of TFs in any cellular context is an expensive laborious procedure, while the assaying of hundreds of TFs even in the same cell state is impractical except in a handful of settings, by the most lavishly funded consortia.

To obtain a more comprehensive view of transcriptional regulation in action, experimental focus has turned from the assaying of individual TFs to the assaying of all open chromatin in a given cellular context. These DNase-seq, ATAC-seq, or single-cell
ATAC-seq accessibility profiles offer a proxy for all cis-regulatory elements active in a given cellular state\(^6\)–\(^8\).

While assaying all TFs is infeasible, many hundreds of TFs have been studied in one or more cellular contexts, or via complementary methods (such as protein binding microarrays or high-throughput SELEX), to obtain the DNA binding preference of the TF\(^1\). These hundreds of TF binding motifs can then be used to predict transcription factor binding sites (TFBSs) for all characterized TFs in various context-specific sets of accessible chromatin.

Very often, biological processes of interest are conserved at the genome sequence level across closely related species, such as primates or mammals. As such, computational tools like PRISM\(^9\) (Predicting Regulatory Information for Single Motifs) can be used to obtain a rarefied subset of binding site predictions that are both observed to be positioned in open chromatin and conserved orthologously in additional species. Because these sites evolve under purifying selection, they are more likely to be individually important in the probed context\(^9\).

Here, we innovate on the foundation of two tools our group previously developed: PRISM\(^9\) for the prediction of evolutionarily conserved binding sites for hundreds of human and mouse TFs, and GREAT\(^4\) for the detection of functions enriched in gene regulatory regions. We use insights from both to develop WhichTF, a tool that applies a novel statistical test to identify the most dominant TFs within a set of user-specified open chromatin regions. In this work, dominant TFs refer to TFs whose conserved binding sites are enriched within functionally-coherent regions of the input open chromatin regions. We show that our molecular definition of dominance successfully predicts biologically
important factors in the context of different cell types, differentiation pathways, and even
disease associated cellular sets.

Results

WhichTF Approach Overview

In order to predict dominant TFs, WhichTF relies on both functional genome
annotations from GREAT and pre-curated, conservation-based predictions of TFBSs from
PRISM. As such, we use GREAT in conjunction with the mouse genome informatics (MGI)
phenotype ontology to annotate all genes in the human GRCh38 (hg38) and mouse
GRCm38 (mm10) genomes with a canonical transcription start site (TSS), a putative gene
regulatory domain, and any MGI phenotypes known to be affected by mutations to the
associated gene. This procedure yields more than 700,000 gene-phenotype relationships for
each genome (Fig. 1a, step 1). We also use PRISM to predict mammalian conserved
TFBSs using 672 manually curated PWMs from 569 TFs across the entire genome. The
updated PRISM predictions resulted in 268 million and 161 million putative TFBSs for the
human and mouse genomes, respectively (Fig. 1a, step 2).

To confirm the utility of restricting ourselves to regulatory domains of highly
enriched ontology terms, we evaluated the relative enrichment in the number of TFBSs
within the input open chromatin region as a baseline method (Online Methods). We found
the baseline results are often overloaded with TFs associated with general housekeeping
processes (Supplementary Table S1). We therefore turned to focus on the top 100 enriched
terms (Online Methods).
For a given query (Fig. 1a, step 3), WhichTF uses functional annotations to enhance its prediction of dominant transcription factors. This is accomplished by computing TF enrichments in only a restricted, particularly relevant, subset of the user’s input. Specifically, WhichTF uses GREAT to identify enriched ontology terms within the user’s input query. Each term is associated with a region of the genome corresponding to all of the regulatory domains of genes annotated with that term. WhichTF selects the top 100 ontology terms. For each term and every TF, WhichTF counts the number of binding sites falling in the intersection of the user-specified accessible regions and the region of the genome associated to the term of interest (Fig. 1a, step 4), and computes enrichment statistics, represented as a TF-by-term enrichment matrix (Fig. 1b). Aggregating over the functional terms, WhichTF computes a novel score and significance used for ranking TFs (Fig. 1c, Online Methods). The top-ranked TFs are hypothesized to be functionally relevant TFs in a cell exhibiting the indicated accessibility profile.

**WhichTF identifies functionally important TFs across diverse cell types**

To test the ability of WhichTF to identify functionally important TFs across different cell types, we applied WhichTF to DNase-seq profiles and found that the predicted dominant TFs are often confirmed to be functionally relevant by perturbation studies (Fig. 2a). In B- and T-cells, for example, we identified TFs in the NF-κB pathway, which are key factors in lymphocyte development and adaptive immunity. In embryonic heart tissue, we found GATA-4, -5, and, -6 – known regulators of cardiac development and growth that, when perturbed, have been implicated in human congenital heart disease. In embryonic hindbrain tissue, we found SOX2, a critical regulator of neural progenitor pluripotency and differentiation in embryogenesis and later development, including adult hippocampal
neurogenesis. WhichTF yielded similar biologically meaningful results from the corresponding cell types for mouse DNase-seq datasets (Supplementary Table S2), suggesting that WhichTF can highlight both the identity and evolutionarily conserved binding sites of key TFs from open chromatin in diverse contexts across species.

WhichTF robustly quantifies biologically meaningful similarities and differences in TF-mediated transcriptional programs

Precise knowledge of cell state and identity is crucial for understanding normal development and disease. To assess whether WhichTF can quantitatively and robustly capture biologically meaningful similarities and differences in TF-mediated transcriptional programs, we applied a t-distributed stochastic neighbor embedding (t-SNE) analysis to WhichTF score vectors computed for 90 samples across 7 cell types. We found brain, lung, and hematopoietic cells are mapped to distinct regions (Fig. 2b). Furthermore, we saw fine-grained substructures among closely related samples. For example, we observed a clear separation of GM12878, B-cells, and T-cells. Reassuringly, different samples from the same biological tissue, such as left ventricle, right ventricle, and heart, showed no clear separation.

WhichTF identifies differentially dominant TFs for closely related cell types

B-cells and T-cells share a closely related developmental trajectory. As Fig. 2a shows, WhichTF identified NF-κB family members NFKB1, RELA, and RELB as shared dominant TFs. WhichTF also identified lineage-specific factors, such as SPI-B for B-cells and RUNX3 for T-cells (Fig. 2a). SPI-B is an ETS family TF known to play a key role in B-cell development and function, and environmental response. RUNX3, in contrast, play T-cell-specific functional roles, such as in CD4 versus CD8 thymocyte commitment, helper
versus killer T-cell specification, and helper type selection\textsuperscript{22}. These differential roles for SPI-B and RUNX3 are corroborated by their cell-type-specific expression in B-cells and T-cells, respectively (Fig. 3a)\textsuperscript{23}.

Although we identified multiple TFs distinguishing B- and T-cells, the results are dominated by common factors. This is reasonable, as they share most of their developmental program\textsuperscript{13}. To identify TFs with relative dominance from a given pair of samples, we developed a differential analysis framework focusing on uniquely accessible regions only in one sample (Online Methods). In B-cells, the differential analysis highlighted additional ETS family members, PU.1 and SPI-C. These TFs are essential for healthy B-cell differentiation and function (Fig. 3b). In T-cells, we saw an additional RUNX family member, RUNX1, as well as CBFβ (Fig. 3b) – both are functionally relevant in T-cells. Indeed, RUNX1, RUNX3 and CBFβ form a complex and are crucial for the healthy function of T-lymphocytes\textsuperscript{32}.

\textbf{WhichTF identifies differentially dominant TFs along developmental trajectories}

TFs regulate cell fate decisions in animal developmental programs\textsuperscript{1}. To gain insights into the molecular mechanisms influencing cellular differentiation, we applied WhichTF to ATAC-seq data from timepoints along mesoderm development to identify differentially dominant TFs that distinguish cell fates at each step along the trajectory, from human embryonic stem cells (ESCs) to early somite vs. cardiac mesoderm (Fig. 4)\textsuperscript{24}.

The first step of mesoderm development is the differentiation from ESCs to anterior (APS) or mid (MPS) primitive streak (PS) cells. In both APS and MPS cells, we found WNT signaling TFs, such as TCF7L2 and LEF1, as well as T-box family TFs, such as TBX-2 and -3 (Fig. 4a-b). WNT signaling is involved in PS differentiation and is crucial in
inducing PS cell types. T-box family members also play key roles in PS development. TBX6 is a canonical PS marker, and the specific loss of Eomes (a.k.a. Tbr2), causes ectopic primitive streak formation in mice. The specific T-box family member TBX3, ranked third in APS cells, has been implicated in early stage of differentiation towards mesoderm from ESCs in mouse and Xenopus and has been reported for its functional redundancy with Tbx2 during Xenopus gastrulation. RUNX3, our top hit for APS, shows conserved expression in mouse neuromesodermal progenitor (NMP) cells and human D3-NMP-like cells. Interestingly, we also found previously unreported T-box family TFs, TBX15 and TBR1, of which TBX15 is linked to decreased skeletal muscle mass in mouse and known for tissue-specific expression in muscle, a tissue developed from the mesoderm lineage (Supplementary Figure S1).

In paraxial mesoderm, we found WNT signaling TFs, which promote paraxial and suppress lateral mesoderm (Fig. 4c). We also find HOXC13, necessary for proper development of the paraxial mesoderm into the presomatic mesoderm. In early somites, we found MEIS2 and ZIC2, which are required in development of cranial and cardiac neural crest and somite cells, respectively (Fig. 4d).

In lateral mesoderm, we found multiple GATA family members, of which GATA4 is a downstream effector of BMP signaling in lateral mesoderm (Fig. 4e). We also saw RUNX3, which is co-expressed with RUNXI in lateral mesoderm; both are necessary for hematopoiesis. GLI1, a key TF in hedgehog (HH) signaling, is necessary for establishing left-right asymmetry in lateral mesoderm. In cardiac mesoderm, we found FOS TFs, GATA TFs, and GLI1 (Fig. 4f). Interestingly, FOSL2 regulates the rate of myocardial differentiation, and HH signaling via GLI1 is required for secondary heart
field development\textsuperscript{35}. As mentioned above, GATA factors are canonical drivers of cardiac development and all the GATA family members identified for mesoderm development (GATA-1, -2, -4, and -6) are implicated in Human cardiovascular diseases\textsuperscript{14,2}.

**WhichTF identifies potentially disease-relevant TFs**

Transcriptional mis-regulation has a broad impact on human diseases\textsuperscript{2}. To assess whether WhichTF can shed light on the transcriptional regulatory molecular basis of human disorders, we examined systemic lupus erythematosus (SLE) as a case study. SLE is a heterogeneous and chronic autoimmune disorder most prevalent in young women and affecting 0.1\% of the population. Its genetic and epi-genetic bases are poorly understood with known genetic associations accounting for only 10-20\% of the observed heritability.

While SLE is characterized by mis-regulated immune response in T- and B-cells, few TFs have been identified to play functionally relevant roles in SLE\textsuperscript{36}.

To better understand the regulatory landscape of SLE, we identified differentially dominant TFs in healthy B-cells compared to SLE-affected B-cells and vice versa by applying WhichTF to ATAC-seq datasets\textsuperscript{37}. We found BCL6 as a differentially dominant TF in healthy vs. SLE B-cells (Table 1). BCL6 is an important marker of T-helper follicular cells, a T-cell subtype which has been found to be mis-regulated in SLE\textsuperscript{38}. Other differentially dominant TFs and their corresponding genes are implicated in autoimmune disorders (Table 1). A sonic hedgehog (SHH)-Gli signaling pathway member GLI1 is involved in pathogenesis of rheumatoid arthritis through synovial fibroblast proliferation\textsuperscript{39}.

A common genetic variant in *TCF7L2*, which is known for type 2 diabetes risk allele, discriminates autoimmune from non-autoimmune type 1 diabetes in young patients\textsuperscript{40}. In a
model system to study multiple sclerosis, ZEB1 is suggested as a regulator of experimental autoimmune encephalomyelitis\(^4\).

**WhichTF uncovers stress response signatures**

Context-specific measurements of open chromatin typically require purification of the desired cell type through mechanical and enzymatic tissue dissociation, which can be quite taxing on the cells. Indeed, it has been reported that stress response factors are often highly expressed in dissociated tissues\(^4\). Corroborating these observations, WhichTF often identifies canonical stress-associated TFs as some of the most dominant TFs in multiple very different contexts. As an illustration, we present WhichTF results for additional DNase-seq datasets (**Table 2**). For three endothelial cell types and adrenal gland cells, we found many members of FOS/AP-1 and NF-κB TFs, which are both known for their roles in stress response. We also found ZFP410 (also known as ZNF410), a poorly characterized Zinc finger TF, among the top hits across multiple cell types, suggesting its potential role in stress response. Even in the samples dominated by stress-associated TFs, we still found well-known context-specific players among the top hits, such as GATA3 and WT-1 in kidney cells and SOX and FOX TFs in endothelial cells\(^4\)–\(^6\). We also found that the boundary between stress response and cell-type specific functions can be ambiguous, or at least context dependent. For example, we found FOS/AP-1 and NF-κB dominant in keratinocytes and B-cells, respectively which, in addition to being stress-associated, are also known for their context-specific functions\(^13,46\).
We present WhichTF, a novel computational method to identify and rank known or novel dominant TFs in any given set of accessible chromatin regions or through pairwise differential analysis of related samples. The WhichTF score is built on high confidence PRISM\textsuperscript{9} predictions of conserved TFBSs as well as gene regulatory domain and ontological annotation models from GREAT\textsuperscript{4}. Applying WhichTF to dozens of samples across diverse biological contexts, such as multiple cell types, developmental programs, and disease samples, we found that the functional relevance of the identified dominant TFs is often supported or suggested by published literature.

WhichTF identifies not only cell-type specific TFs, but factors reflecting biological processes shared among multiple samples. One such example in our result, corroborated by previous expression profiling, suggests stress response due to cellular dissociation is a shared process\textsuperscript{42}. In addition to previously identified factors, we report an under-characterized Zinc finger protein, ZNF410, as a TF potentially involved in cellular stress response. The identification of stress associated TFs suggests WhichTF may serve as a useful quality control of chromatin accessibility data.

As we have demonstrated above, WhichTF is broadly applicable. WhichTF takes as input any form of chromatin accessibility measurement for either human or mouse, the two most studied genomes. Our illustrative examples span both species and assay types, such as DNase-seq and ATAC-seq. When combined with emerging single-cell accessibility profiling technologies\textsuperscript{8}, WhichTF will provide systematic characterization of dominant TFs across a spectrum of cell-types. For example, application of WhichTF to datasets from large-scale projects, such as the Human Cell Atlas project\textsuperscript{47}, has the potential to discover
dominant TFs for each cell type and binding sites of those TFs. Moreover, our differential
analysis framework will help in understanding how closely related cell types diverge by
providing hypotheses of differentially important TFs.

The resources made available with this study, including WhichTF and the GREAT
update, provide an excellent foundation for investigating the molecular mechanisms of TF-
mediated cis-regulation. Together, these results highlight the benefit of combining
experimental characterization of chromatin accessibility, high-quality TFBS reference
datasets, and ontological genome annotation, suggesting that systematic identification of
dominant TFs across a large number of samples will be a powerful approach to understand
molecular mechanisms of gene regulation and their influence on cell type differentiation,
development, and disease.
Online Methods

GREAT v.4.0.4 update

We performed a major update of Genomic Regions Enrichment of Annotations Tool (GREAT)\(^4\) and released it as version 4.0.4. GREAT currently supports the human (\textit{Homo sapiens} GRCh38 and GRCh37/hg19) and mouse (\textit{Mus musculus} GRCm38/mm10 and NCBIM37/mm9) genomes. We obtained Ensembl gene sets from the following Ensembl\(^48\) versions:

- Human GRCh38: Ensembl version 90
- Human GRCh37: Ensembl for GRCh37 version 90
- Mouse GRCm38: Ensembl version 90
- Mouse NCBIM37: Ensembl version 67

By focusing on the set of genes with at least one Gene Ontology (GO) annotation\(^10,11\) as described before\(^4\), we defined putative gene regulatory domains for 18,777 (GRCh38), 18,549 (GRCh37/hg19), 21,395 (GRCm38/mm10), and 19,996 (NCBIM37/mm9) genes’ canonical transcription start sites.

We also updated the ontology reference data. GREAT currently supports the most recent versions of the following ontologies at the time of analysis: Ensembl genes, Gene Ontology (GO)\(^10,11\), human phenotype ontology\(^49\), and mouse genome informatics (MGI) phenotype ontology\(^12\) (Supplementary Table S3). The new Ensembl genes ontology is a “flat” ontology that makes every gene into a term, facilitating the testing of cis-regulatory elements congregation in the regulatory domains of individual genes. For MGI phenotype ontology, we mapped MGI gene identifiers to Ensembl human gene IDs using one-to-one orthology mappings from Ensembl Biomart\(^48\) version 90. In total, we compiled 2,861,656,
285 2,846,384, 2,734,172, and 2,675,691 gene-term relationships for GRCh38, GRCh37, GRCm38, and NCBIM37 genome assemblies, respectively (Supplementary Table S3).

287 **Computational TFBS prediction with PRISM**

To take advantage of growing sequence data from both multiple species and functional genomics datasets, we updated our computationally predicted PRISM conserved transcription factor binding sites (TFBSs) for the human (*Homo sapiens* GRCh38 and GRCh37) and mouse (*Mus musculus* GRCm38 and NCBIM37) genomes. Briefly, PRISM predicts TFBSs based on evolutionary conservation of TF motif matches\(^9\). The GRCh37 and NCBIM37 tracks are derived using liftOver\(^50\) from that of GRCh38 and GRCm38, respectively.

We used the following multiple alignment from the UCSC genome browser\(^50\):  
- Human GRCh38: Hg38 100-way conservation alignment (lastz)  
- Mouse GRCm38: Mm10 60-way conservation alignment (lastz)

We removed Killer whale (*Orcinus orca*, orcOrc1) from the human alignment because of chromosome name mismatch. We further subset the alignments to Eutherian species\(^9\), resulting in 57 and 40 species for human and mouse, respectively. Using our manually curated TF monomer motif library\(^51\), we applied PRISM\(^9\) with the default parameters and focused on the top 10,000 predicted TFBSs for each TF in our analyses. We used GNU parallel in our analysis\(^52\).

**Baseline TF enrichment method without functional annotation**

We computed the binomial p-value of each TFBS set, using the total number of TFBS predictions, the number intersecting the query and the fraction of the genome covered
by the open chromatin region. We ranked the TFs by their binomial fold (Supplementary Table S1).

**WhichTF analysis protocol**

WhichTF combines user specified accessibility measures, such as ATAC-seq or DNase-seq peaks with precomputed reference datasets to produce a ranked list of context specific, dominant TFs. The reference datasets consist of GREAT regulatory domain models, MGI mouse phenotype ontology-based gene annotations, and PRISM TFBS predictions.

WhichTF first identifies the top 100 ontology terms \( (\pi_1, \ldots, \pi_{100}) \) based on the GREAT enrichment test on the input query set with the default “basal plus extension” association rule and a filter that terms must be associated with no fewer than two genes and no more than 500 genes associated to them. For each TF in the PRISM TFBS prediction library of \( N \) TFs, WhichTF takes an intersection of the TFBS prediction track and the user submitted open regions using overlapSelect \(^{50}\).

Each TF in the PRISM library has a different number of TFBSs and regulatory domains of different total sizes associated with each term. To capture the relative importance of different TFs within different contexts, WhichTF computes a few measures of statistical significance for each transcription factor and term and summarizes these measures in TF by term summary statistic matrices. Specifically, we apply hypergeometric and binomial tests defined below:
TF hypergeometric test

Let’s define the GREAT gene regulatory domain for term $\pi_j$ as RegDom$_j$, PRISM

TFBS prediction for TF$_i$ as TFBS$_i$, and user’s input query as QUERY. We define $n_i$, $k_{ij}$, $N_i$, and $K_{ij}$ as follows:

1. $n_i = \#\{\text{TFBS}_i \cap \text{QUERY}\}$
2. $k_{ij} = \#\{(\text{TFBS}_i \cap \text{QUERY}) \cap \text{RegDom}_j\}$
3. $N = \#\{(\bigcup_k \text{TFBS}_k) \cap \text{QUERY}\}$
4. $K_j = \#\{(\bigcup_k \text{TFBS}_k) \cap \text{QUERY} \cap \text{RegDom}_j\}$

where, $\cap$ denotes genomic intersection operation and $\#\{G\}$ denotes a function to count the number of elements in genomic regions, $G$. With these parameters, we compute the hypergeometric p-value for each pair of TF$_i$ and term $\pi_j$:

$$\min(n_i, K_j) \sum_{k = k_{ij}}^{n_i} \binom{K_j}{k} \binom{N - K_j}{n_i - k} \binom{N}{n_i}$$

TF binomial test

Using the intersection track, TFBS$_i \cap \text{QUERY}$, we compute the GREAT binomial p-value for each pair of TF$_i$ and term $\pi_j$:

$$\sum_{k = k_{ij}}^{n_i} \binom{n_i}{k} p_{\pi_j}^k (1 - p_{\pi_j})^{n_i - k}$$

where, $p_{\pi}$ denotes the probability of drawing a base annotated with term $\pi$ from non-gap genomic sequences under the uniform distribution$^4$. 
Adaptive TF significance threshold

To eliminate false positives, WhichTF focuses on terms where the most significant TF characterized by both hypergeometric and binomial p-value match. Using the enrichment statistics, WhichTF selects dominant TFs for each selected ontology term. We compute the adaptive threshold for each of the hypergeometric and binomial test by finding a leap in the p-values of the top 10 TFs for each term using the following procedure. Let’s denote the top 10 hypergeometric p-values for a fixed functional term $\pi$ as $p_1 \leq p_2 \leq \cdots \leq p_{10}$. We define the difference of adjacent negative log of p-values as $d_k = -\log \frac{p_k}{p_{k+1}}$. We define $m$, the index with the largest leap in p-value as $m = \arg\max_k d_k$. Our adaptive threshold is $p_m$ and we only keep TFs with hypergeometric p-values that satisfies $p \leq p_m$. We define the adaptive threshold for binomial p-values in the same way. We say TF$_i$ is significant for term $\pi_j$ when it passes the adaptive thresholds for both TF hypergeometric and TF binomial tests.

WhichTF scores

For each TF, WhichTF computes the score by the following equation. Let $(\pi_1, \ldots, \pi_K)$ be the set of terms selected from step 1 in the order of relevance with $\pi_1$ as the top hit. Let $\text{Rank}(\text{TF}_i, \pi_j)$ be the rank of $\text{TF}_i$ for term $\pi_j$. Let $\text{Significant}(\text{TF}_i, \pi_j)$ denote a Boolean variable that indicates whether $\text{TF}_1$ passes the filters described above for term $\pi_j$ (i.e. Significant is 1 if the TF passes the significance filter and zero otherwise). With this notation, we define the WhichTF score of TF$_i$ as:

$$\text{WhichTF score (TF}_i) = \sum_j \frac{\text{Significant}(\text{TF}_i, \pi_j)}{\sqrt{j \cdot \text{Rank}(\text{TF}_i, \pi_j)}}.$$
WhichTF conditional p-values

WhichTF computes the statistical significance of a WhichTF score based on a null model that any ordering of TFs within each term is equally likely. Thus, the probability of a given score is determined by the relative number of configurations with the score. To enumerate the number of configurations with a given score in polynomial time, we devised a dynamic programing approach\(^3\) which acts recursively on the number of functional terms, \(K\). This procedure first discretizes each contribution to the summand in the definition of the WhichTF score defined above. Let \(\{s_{j1}, s_{j2}, \ldots, s_{jM_j}\}\) be the set of all the possible cumulative scores up to term \(\pi_j\), that is the scores gotten by computing the above sum only up to term \(\pi_j\). Here, \(M_j\) is the number of distinct discretized scores up to term \(\pi_j\). Let \(n_{j1}\) represent the number of different ways of getting each such score, \(s_{j1}\), and let \(S_j = \{(s_{j1}, n_{j1}), (s_{j2}, n_{j2}), \ldots, (s_{jM_j}, n_{jM_j})\}\) be the set of all tuples of scores and number of configurations. Finally, let \(\{t_{j1}, t_{j2}, \ldots, t_{jM_j}\}\) denote the individual summands at term \(\pi_j\).

The p-value of each score is computed directly from \(S_K\), the full set of cumulative scores and number of configurations, by dividing the number of configurations with scores greater than or equal to a given score by the total number of configurations. This list of tuples, \(S_j\), can be computed recursively with the base case of \(S_0 = \{(0, 1)\}\). The set of scores at level \(j+1\) is given by all combinations, \(s_{j1} + t_{j+1,k}\), with the number of configurations given by aggregating over all combinations of \(s\) and \(t\) that yield the same cumulative score.

Given that the WhichTF scores of multiple TFs are not independent, we apply the procedure defined above from the top scoring TF to the TF with the lowest score and compute conditional statistical significance. This means that for the computation of...
statistical significance of the $i$-th ranking TF, we remove TFs whose rank is smaller than $i$
and apply the recursive procedure defined above.

**Application of WhichTF in diverse functional contexts**

**Multiple cell types from the ENCODE/Roadmap project**

From the ENCODE/Roadmap data portal, we obtained “hotspot” files derived from DNase-seq experiments\textsuperscript{54,55}. All coordinates are provided in GRCh37. We present analysis spanning 95 samples from 12 cell types and tissues (Supplementary Table S4).

We systematically applied WhichTF to each sample and obtained the ranked list of TFs as well as a vector of WhichTF scores across all TFs in the library (Figure 2a, Table 2).

We applied t-SNE, a non-linear dimension reduction method\textsuperscript{18}, implemented in Python Scikit Learn library\textsuperscript{56} with perplexity 10 (Figure 2b).

Using mouse ENCODE DNase-seq datasets provided in GRCm38 from the four cell types used for the human analysis (Figure 2a, Supplementary Table S5), we applied WhichTF using mouse GRCm38 reference dataset (Supplementary Table S2).

**Cell type-specific expression analysis**

We presented cell type-specific RNA-seq data from the GEO database (GSE118165)\textsuperscript{23}. We subseted this dataset to the unstimulated samples and plotted the expression of SPIB and RUNX3 for lymphoid cells in T and B cell lineages (Figure 3a).

**WhichTF for differential analysis**

To find TFs dominant in an input set A compared to another input set B, we defined set A and set B regions as foreground and background, respectively. We used bedtools\textsuperscript{57}
“subtract” to keep a subset of A that does not overlap with B. We applied WhichTF single run mode (above) on the identified differentially accessible regions (Figure 3b).

Mesoderm lineage dataset

Using ATAC-seq datasets (SRP073808 from NCBI GEO database) of mesoderm development\textsuperscript{24} (Supplementary Table S6), we applied WhichTF differential analysis following the diagram of sequential differentiation (Figure 4).

Systemic lupus erythematosus dataset

Eight sets (4 SLE and 4 healthy controls [HC]) were taken from the NCBI sequence read archive (SRA, Supplementary Table S7). Paired end reads were mapped using bowtie2 with the outer distance flag (-X) set to 1000 and otherwise default settings\textsuperscript{58}. Samtools was used to generate a sorted bam file and MACS2 was used to call peaks with shift set to 37, extension size set to 72 and broad and keep-dup flags on\textsuperscript{59,60}. Given that some of the samples in this dataset are from a biobank, we conservatively defined differentially accessible regions shown below and applied WhichTF differential analysis (Table 1):

\begin{itemize}
  \item SLE – HC := SRR3158183 – \bigcup_{x \in SRR3158176-9} x
  \item HC – SLE := \bigcap_{x \in SRR3158176-9} x – \bigcup_{x \in SRR3158180-3} x
\end{itemize}

Tissue-specific gene expression of the identified TF

Using the data obtained from the GTEx Portal\textsuperscript{61} on 05/24/2019 (phs000424.v7.p2), we investigated whether the identified TFs in have a tissue-specific expression (Supplementary Figure S1).
Data availability

All datasets analyzed in this study are publicly available through the ENCODE/Roadmap portal [https://www.encodeproject.org/], NCBI GEO database [https://www.ncbi.nlm.nih.gov/geo/], NCBI sequence read archive [NCBI sequence read archive], or the GTEx Portal [https://gtexportal.org] with identifiers included in Supplementary Tables S4-S7 and in Online Methods.

Code availability

WhichTF program and analysis scripts are available at our Bitbucket repository: https://bitbucket.org/bejerano/whichtf

GREAT version 4.0.4: https://great.stanford.edu

References

1. Lambert, S. A. et al. The Human Transcription Factors. Cell 172, 650–665 (2018).
2. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. Cell 152, 1237–1251 (2013).
3. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497–1502 (2007).
4. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501 (2010).
5. Rosenfeld, M. G. Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. Genes Dev. 20, 1405–1428 (2006).
6. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

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

8. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

9. Wenger, A. M. *et al.* PRISM offers a comprehensive genomic approach to transcription factor function prediction. *Genome Res.* **23**, 889–904 (2013).

10. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

11. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).

12. Smith, C. L. & Eppig, J. T. Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *J. Biomed. Semant.* **6**, 11 (2015).

13. Gerondakis, S. & Siebenlist, U. Roles of the NF-κB Pathway in Lymphocyte Development and Function. *Cold Spring Harb. Perspect. Biol.* **2**, a000182 (2010).

14. Pikkarainen, S., Tokola, H., Kerkelä, R. & Ruskoaho, H. GATA transcription factors in the developing and adult heart. *Cardiovasc. Res.* **63**, 196–207 (2004).
15. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

16. Thiel, G. *Transcription Factors in the Nervous System: Development, Brain Function, and Diseases*. (John Wiley & Sons, 2006).

17. Hodge, R. D. & Hevner, R. F. Expression and actions of transcription factors in adult hippocampal neurogenesis. *Dev. Neurobiol.* **71**, 680–689 (2011).

18. Maaten, L. V. D. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

19. DeKoter, R. P. *et al.* Regulation of Follicular B Cell Differentiation by the Related E26 Transformation-Specific Transcription Factors PU.1, Spi-B, and Spi-C. *J. Immunol.* **1001413** (2010). doi:10.4049/jimmunol.1001413

20. Sokalski, K. M. *et al.* Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. *Blood* blood-2011-02-335539 (2011). doi:10.1182/blood-2011-02-335539

21. Willis, S. N. *et al.* Environmental sensing by mature B cells is controlled by the transcription factors PU.1 and SpiB. *Nat. Commun.* **8**, 1426 (2017).

22. Woolf, E. *et al.* Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 7731–7736 (2003).

23. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *bioRxiv* 409722 (2018). doi:10.1101/409722

24. Loh, K. M. *et al.* Mapping the Pairwise Choices Leading from Pluripotency to Human Bone, Heart, and Other Mesoderm Cell Types. *Cell* **166**, 451–467 (2016).
25. Papaioannou, V. E. The T-box gene family: emerging roles in development, stem cells and cancer. *Development* **141**, 3819–3833 (2014).

26. Weidgang, C. E. *et al.* TBX3 Directs Cell-Fate Decision toward Mesendoderm. *Stem Cell Rep.* **1**, 248–265 (2013).

27. Denans, N., Iimura, T. & Pourquié, O. Hox genes control vertebrate body elongation by collinear Wnt repression. *eLife* (2015). doi:10.7554/eLife.04379

28. Machon, O., Masek, J., Machonova, O., Krauss, S. & Kozmik, Z. Meis2 is essential for cranial and cardiac neural crest development. *BMC Dev. Biol.* **15**, 40 (2015).

29. Inoue, T., Ota, M., Mikoshiba, K. & Aruga, J. Zic2 and Zic3 synergistically control neurulation and segmentation of paraxial mesoderm in mouse embryo. *Dev. Biol.* **306**, 669–684 (2007).

30. Rojas, A. *et al.* Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element. *Development* **132**, 3405–3417 (2005).

31. Park, B.-Y. & Saint-Jeannet, J.-P. Expression analysis of Runx3 and other Runx family members during Xenopus development. *Gene Expr. Patterns GEP* **10**, 159–166 (2010).

32. Kalev-Zylinska, M. L. *et al.* Runx3 is required for hematopoietic development in zebrafish. *Dev. Dyn.* **228**, 323–336 (2003).

33. Tsiairis, C. D. & McMahon, A. P. A Hh-dependent Pathway in Lateral Plate Mesoderm Enables The Generation Of Left-Right Asymmetry. *Curr. Biol. CB* **19**, 1912–1917 (2009).
34. Jahangiri, L. et al. The AP-1 transcription factor component Fosl2 potentiates the rate of myocardial differentiation from the zebrafish second heart field. *Dev. Camb. Engl.* **143**, 113–122 (2016).

35. Dyer, L. A. & Kirby, M. L. Sonic hedgehog maintains proliferation in secondary heart field progenitors and is required for normal arterial pole formation. *Dev. Biol.* **330**, 305–317 (2009).

36. Tsokos, G. C. Systemic Lupus Erythematosus. *N. Engl. J. Med.* **365**, 2110–2121 (2011).

37. Scharer, C. D. et al. ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. *Sci. Rep.* **6**, 27030 (2016).

38. Fujikura, D. et al. Death receptor 6 contributes to autoimmunity in lupus-prone mice. *Nat. Commun.* **8**, 13957 (2017).

39. Qin, S. et al. The Effect of SHH-Gli Signaling Pathway on the Synovial Fibroblast Proliferation in Rheumatoid Arthritis. *Inflammation* **39**, 503–512 (2016).

40. Bakhtadze, E. et al. Common variants in the TCF7L2 gene help to differentiate autoimmune from non-autoimmune diabetes in young (15–34 years) but not in middle-aged (40–59 years) diabetic patients. *Diabetologia* **51**, 2224–2232 (2008).

41. Stridh, P. et al. Fine-Mapping Resolves Eae23 into Two QTLs and Implicates ZEB1 as a Candidate Gene Regulating Experimental Neuroinflammation in Rat. *PLOS ONE* **5**, e12716 (2010).

42. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
Van Esch, H. & Bilous, R. W. GATA3 and kidney development: why case reports are still important. *Nephrol. Dial. Transplant.* **16**, 2130–2132 (2001).

Kreidberg, J. A. WT1 and kidney progenitor cells. *Organogenesis* **6**, 61–70 (2010).

Park, C., Kim, T. M. & Malik, A. B. Transcriptional regulation of endothelial cell and vascular development. *Circ. Res.* **112**, 1380–1400 (2013).

Eckert, R. L. *et al.* AP1 transcription factors in epidermal differentiation and skin cancer. *J. Skin Cancer* **2013**, 537028 (2013).

Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).

Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky1105

Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

Guturu, H., Doxey, A. C., Wenger, A. M. & Bejerano, G. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**, 20130029 (2013).

Tange, O. GNU Parallel - The Command-Line Power Tool. *Login USENIX Mag.* **36**, 42–47 (2011).

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to Algorithms, 3rd Edition.* (The MIT Press, 2009).

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
55. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

56. Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* **12**, 2825–2830 (2011).

57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

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

61. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
Acknowledgements

We thank Stanford’s Kyle M. Loh, as well as Heidi Chen, Alex M. Tseng, and other members of the Bejerano Lab for useful discussions, feedback and advice. Y.T. is supported by a Funai Overseas Scholarship from the Funai Foundation for Information Technology and the Stanford University School of Medicine. E.S.D. was supported in part by the Simons Collaboration Grant on the Non-Perturbative Bootstrap. This work was supported by National Institute of Mental Health (NIMH) of the National Institutes of Health (NIH) under awards U01MH105949 to G.B. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author information

Author contributions

E.S.D., Y.T. and G.B. conceived and designed the study. Y.T. updated GREAT. E.S.D. conceived of and developed the WhichTF algorithm with support from Y.T. and G.B. E.S.D. and Y.T. performed the computational analyses. Y.T. led the completion of the manuscript with support from E.S.D. and oversight from G.B. Y.T. and E.S.D. contributed equally to the project and author list is ordered by age. The manuscript was written and approved by all authors.

Competing interests

The authors declare no competing interests.
Figures and Tables

Figure 1

Fig. 1 WhichTF identifies dominant TFs for given experimental measurements of chromatin accessibility. (a) WhichTF uses gene regulatory domain models and ontologies from the genomic region enrichment analysis tool (GREAT) (step 1) and conservation-based PRISM predictions of TFBSs (step 2). Given a user-defined set of genomic regions (step 3), WhichTF considers the top-\(K\) GREAT functional terms \((\pi_1, \ldots, \pi_K)\) enriched in the query regions. For all pairwise combinations of top-\(K\) term and TF, WhichTF counts the number of TFBSs within the specified query regions (step 4). (b) The binomial and hypergeometric TFBS enrichment \(p\)-values for each ontology term are compiled in a TF-by-term summary statistic matrix. (c) Aggregating the summary statistics over terms, WhichTF returns a ranked list of TFs, ordered by predicted functional importance in the user-specific chromatin environment, with the corresponding scores and statistics (Online Methods). TSS, transcription start site.
Fig. 2 Which TF identifies dominant TFs in diverse cell types. (a) The top 5 identified dominant TFs for B-, T-, heart, and brain cells are shown with the corresponding negative
log conditional probability (-log CP), a statistical significance of the score of each TF, conditioned on the TFs with higher score (Online Methods). The importance and PubMed ID (PMID) columns indicate whether existing literature supports the role of the identified TFs, typically through perturbation experiments. (b) For DNase-seq data tracks of 90 samples across 7 cell types, the WhichTF score vectors are projected to t-SNE plot. WhichTF quantitatively and robustly captures biological similarities and dissimilarities of TF-mediated transcriptional programs. The samples highlighted in (a) are annotated with arrows.
**Figure 3**

(a) Gene expression of the top differential TF genes, *SPI-B* and *RUNX3*, are shown (horizontal axis) across diverse lymphoid cell types (vertical axis) for up to four healthy donors. (b) The top 5 differential TFs for B-cells relative to T-cells (B-cell – T-cell) and vice versa (T-cell – B-cell) are shown with the corresponding statistical significance, negative log conditional probabilities (-log CP). The importance and PubMed ID (PMID) columns indicate whether existing literature supports the identified TFs.

| B cells – T cells | T cells – B cells |
|-------------------|-------------------|
| TF    | -log(CP) | Importance | PMID        | TF    | -log(CP) | Importance | PMID        |
|-------|----------|------------|-------------|-------|----------|------------|-------------|
| SPIB  | 28.4     | Confirmed  | 21057087    | RUNX3 | 171.1    | Confirmed  | 12796513    |
| SPI1  | 21.4     | Confirmed  | 21057087    | NFKB1 | 47.7     | Confirmed  | 20452952    |
| SPIC  | 17.1     | Confirmed  | 21057087    | RUNX1 | 36.5     | Confirmed  | 12796513    |
| REL   | 4.3      | Confirmed  | 20452952    | REL   | 8.0      | Confirmed  | 20452952    |
| RELB  | 2.8      | Confirmed  | 20452952    | CBFB  | 9.1      | Confirmed  | 17185462    |
**Figure 4**

Which TF identifies differentially dominant TFs compared to immediate progenitor cells along human mesoderm development pathway from ATAC-seq data. The top 5 TFs with the corresponding statistical significance, negative log conditional probabilities (-log CP) are shown. The importance (Imp.) and PubMed ID (PMID) columns indicate whether (i) existing literature supports the identified TFs (C: confirmed); (ii) literature reports closely related factors, such as co-factors and functionally related family members, or the identified TFs in related context (S: suggestive); or (iii) novel (N). ESCs, embryonic stem cells.

|   | Anterior primitive streak (APS) |     |    |    |
|---|---------------------------------|-----|----|----|
| a | TF                             | -log( CP) | Imp. | PMID |
|   | RUNX3                          | 11.7 | S   | 29899136 |
|   | TCF7L2                         | 10.0 | C   | 27419872 |
|   | TBX3                           | 7.4  | C   | 24319661 |
|   | CRX                            | 7.5  | S   | 17490610 |
|   | LEF1                           | 7.6  | C   | 27419872 |

|   | Mid primitive streak (MPS)     |     |    |    |
|---|---------------------------------|-----|----|----|
| b | TF                             | -log( CP) | Imp. | PMID |
|   | TCF7L2                         | 16.9 | C   | 27419872 |
|   | TBX2                           | 13.0 | S   | 24319661 |
|   | TBR1                           | 12.4 | N   |      |
|   | TBX15                          | 13.0 | N   |      |
|   | GATA4                          | 6.2  | S   | 27419872 |

|   | Paraxial mesoderm              |     |    |    |
|---|---------------------------------|-----|----|----|
| c | TF                             | -log( CP) | Imp. | PMID |
|   | TCF7L2                         | 7.7  | C   | 27419872 |
|   | LEF1                           | 6.9  | C   | 27419872 |
|   | HOXC13                         | 4.8  | C   | 25719209 |
|   | TCF7F1                         | 4.6  | C   | 27419872 |
|   | IKZF3                          | 3.6  | N   |      |

|   | Lateral mesoderm               |     |    |    |
|---|---------------------------------|-----|----|----|
| e | TF                             | -log( CP) | Imp. | PMID |
|   | RUNX3                          | 8.6  | C   | 20433948 |
|   | GLI1                           | 10.9 | C   | 19879143 |
|   | GATA4                          | 7.1  | C   | 27419872 |
|   | GATA6                          | 5.1  | C   | 27419872 |
|   | GATA2                          | 4.7  | C   | 27419872 |

|   | Early somite                   |     |    |    |
|---|---------------------------------|-----|----|----|
| d | TF                             | -log( CP) | Imp. | PMID |
|   | MEIS2                          | 6.9  | C   | 9337138 |
|   | ZIC2                           | 6.6  | C   | 17490632 |
|   | INSM1                          | 4.3  | S   | 25053427 |
|   | TEAD4                          | 4.1  | C   | 29636889 |
|   | FOXI1                          | 3.9  | N   |      |

|   | Cardiac mesoderm               |     |    |    |
|---|---------------------------------|-----|----|----|
| f | TF                             | -log( CP) | Imp. | PMID |
|   | GLI1                           | 10.9 | C   | 23873040 |
|   | FOS                            | 6.7  | C   | 11003651 |
|   | GATA4                          | 7.8  | C   | 24790981 |
|   | FOSL2                          | 6.5  | C   | 26732840 |
|   | GATA1                          | 5.2  | C   | 21464046 |
Table 1: WhichTF identifies disease relevant TFs

| TF       | HC - SLE -log(CP) | Imp. | PMID     | SLE - HC -log(CP) | Imp. | PMID     |
|----------|-------------------|------|----------|-------------------|------|----------|
| BCL6     | 28.7              | C    | 28045014 | GLI1              | 19.7 | S        |
| TFAP2B   | 19.3              | N    |          | ZFP143            | 11.0 | N        |
| ZEB1     | 16.6              | S    | 20856809 | TCF7L2            | 6.0  | S        |
| ZSCAN21  | 15.2              | N    |          | ONECUT2           | 5.2  | S        |
| ZSCAN20  | 14.2              | N    |          | DMRTC2            | 3.8  | N        |

Table 1 WhichTF identifies differentially dominant TFs from ATAC-seq measurement of B-cells from systemic lupus erythematosus (SLE) patients and healthy controls (HC). The top 5 TFs based on the analysis of HC with respect to SLE (HC - SLE) and vice versa (SLE - HC) are shown with the corresponding statistical significance, negative log conditional probabilities (-log CP). The importance (Imp.) and PubMed ID (PMID) columns indicate whether literature supports the identified TFs: confirmed (C), suggestive (S), or novel (N).
| Table 2: WhichTF identifies stress response factors in different samples |
|--------------------------------------------------------------|
| B-cell | Keratinocyte | Adrenal Gland | Lymphatic Vessel | Pulmonary Artery | Dermis Vessel |
|        | ENCCF719GOE | ENCCF047IIB | ENCCF212TPU | ENCCF354CZP | ENCCF596PRJ | ENCCF908DMH |
| 1 | SPIB | * | FOSB | * | ZFP410 | NFKB1 | + | FOSL1 | + | NFKB1 | + |
| 2 | NFKB1 | * | FOS | * | FOS | + | FOS | + | FOS | + | FOS | + |
| 3 | RELB | * | FOSL1 | * | FOSL1 | + | FOSL1 | + | FOSL2 | + | FOSL1 | + |
| 4 | RELA | * | JUND | * | NFKB1 | + | RELB | + | NFKB1 | + | RELA | + |
| 5 | SPI | * | BATF | + | BATF | + | BATF | + | BATF | + | BATF | + |
| 6 | SPI1 | * | FOSL2 | + | FOSL2 | + | JUND | + | RELB | + | BATF | + |
| 7 | ZFP410 | * | BACH2 | + | BACH1 | + | FOSL2 | + | BATF | + | FOSB | + |
| 8 | RUNX3 | * | JUND | + | JUND | + | REL | + | RELA | + | RELB | + |
| 9 | REL | * | BACH1 | + | RELB | + | RELA | + | SOX10 | * | JUND | + |
| 10 | STAT2 | * | JUN | * | BACH2 | + | SPIC | * | FOSB | + | SOX7 | * |
| 11 | WT1 | NFE2L2 | GATA3 | + | FOSB | + | BACH2 | + | ZFP410 | + |
| 12 | SNAI3 | NFKB1 | JUN | + | ZFP410 | BACH1 | + | BACH1 | + | BACH1 | + |
| 13 | ZEB2 | * | MZF1 | WT1 | * | SPIB | * | GATA4 | * | GATA4 | * |
| 14 | ATF6 | RELB | BATF | + | SOX30 | * | JUNB | * | SOX12 | * | SOX12 | * |
| 15 | E2F5 | ZFP217 | NFE2L2 | SOX7 | * | GATA5 | * | FOXD1 | * | FOXD1 | * |
| 16 | IKZF3 | * | ETS2 | GATA6 | SOX18 | * | SOX30 | * | FOX3 | * | SOX30 | * |
| 17 | ELF5 | PITX1 | FOSB | + | JUND | + | SPIB | * | SOX30 | * | SOX30 | * |
| 18 | SP100 | ATF6 | GATA4 | SOX12 | * | SOX18 | * | SOX18 | * | SOX18 | * |
| 19 | IF9 | * | MTF | * | BACH1 | + | JUN | * | FOXO6 | * | FOXO6 | * |
| 20 | SNAI1 | * | MTF | * | FOXP2 | * | FOXP3 | * | FOXP3 | * | FOXP4 | * |

Table 2: WhichTF identifies TFs known for stress response. The top 20 TFs identified by WhichTF are shown in ranked order for B-cells, keratinocytes, adrenal gland, lymphatic vessel endothelium, pulmonary artery endothelium, and dermis vessel endothelium cells. The TFs known to be involved in stress response signals are marked with plus (+), while TFs in families known to be functionally important in each context are marked with asterisk (*).
Supplementary materials

List of supplementary materials

Supplementary Figures

- Supplementary Figure S1: Gene expression profile of TBX15

Supplementary tables

- Supplementary Table S1: Baseline TF enrichment method
- Supplementary Table S2: Mouse ENCODE dataset analysis
- Supplementary Table S3: The update summary of GREAT ontologies
- Supplementary Table S4: Human ENCODE datasets
- Supplementary Table S5: Mouse ENCODE datasets
- Supplementary Table S6: Mesoderm development samples
- Supplementary Table S7: Sequence read archive accession IDs for systemic lupus erythematosus dataset
**Supplementary Figures**

Supplementary Figure S1. Tissue-specific gene expression profile of *TBX15* in muscle. The Human cell types are shown on x-axis and the expression (TPM) is shown on y-axis. The median and 25th and 75th percentiles are shown as box plots and data points are shown as outliers if they are above or below 1.5 times the interquartile range.

**Supplementary Tables**

Supplementary Table S1. Baseline TF enrichment method for the four human cell types from ENCODE and Roadmap DNase-seq datasets are shown. The top 5 identified TFs are shown for (a) B-cells, (b) T-cells, (c) heart cells, and (d) brain cells. ENCODE accession IDs for each sample and the dominant TFs and their corresponding -log10(p-value) are shown. There is less cell-type specificity in the identified results.

Supplementary Table S2. Mouse ENCODE dataset analysis. WhichTF identifies dominant TFs for four mouse cell types from ENCODE and Roadmap DNase-seq dataset. The top 5 identified dominant TFs are shown for (a) B-cells, (b) T-cells, (c) heart cells, and (d) hindbrain cells. The ENCODE accession IDs for each sample are shown on the top and the
dominant TFs and their corresponding statistical significance, conditional probabilities, are shown.

**Supplementary table S3.** The update summary of GREAT ontologies. Ensembl genes is a flat ontology defined from the set of genes with at least one meaningful annotation in gene ontology (Online Methods). GO: gene ontology. HPO: human phenotype ontology. MGI: mouse genome informatics.

**Supplementary Table S4.** Human ENCODE datasets. The list of ENCODE accession IDs used in our study and the corresponding cell type or tissues.

**Supplementary Table S5.** Mouse ENCODE datasets. The list of ENCODE accession IDs used in our study and the corresponding cell type or tissues.

**Supplementary Table S6.** Mesoderm development samples. The list of sample IDs, sample description, and the reference to the corresponding results.

**Supplementary Table S7.** Sequence read archive (SRA) accession IDs for systemic lupus erythematosus dataset. SLE indicates disease and HC indicates healthy control.