

Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells

Emily R. Miraldi,^{1,2} Maria Pokrovskii,³ Aaron Watters,⁴ Dayanne M. Castro,⁵ Nicholas De Veaux,⁴ Jason A. Hall,³ June-Yong Lee,³ Maria Ciofani,⁶ Aviv Madar,^{5,9} Nick Carriero,⁴ Dan R. Littman,^{3,7} and Richard Bonneau^{4,5,8}

¹Divisions of Immunobiology and Biomedical Informatics, Cincinnati Children's Hospital, Cincinnati, Ohio 45229, USA; ²Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio 45257, USA; ³Molecular Pathogenesis Program, The Kimmel Center for Biology and Medicine of the Skirball Institute, New York, New York 10016, USA; ⁴Center for Computational Biology, Flatiron Institute, New York, New York 10010, USA; ⁵Department of Biology, New York University, New York, New York 10012, USA; ⁶Department of Immunology, Duke University School of Medicine, Durham, North Carolina 27710, USA; ⁷The Howard Hughes Medical Institute; ⁸Center for Data Science, New York University, New York, New York 10010, USA

Transcriptional regulatory networks (TRNs) provide insight into cellular behavior by describing interactions between transcription factors (TFs) and their gene targets. The assay for transposase-accessible chromatin (ATAC)-seq, coupled with TF motif analysis, provides indirect evidence of chromatin binding for hundreds of TFs genome-wide. Here, we propose methods for TRN inference in a mammalian setting, using ATAC-seq data to improve gene expression modeling. We test our methods in the context of T Helper Cell Type 17 (Th17) differentiation, generating new ATAC-seq data to complement existing Th17 genomic resources. In this resource-rich mammalian setting, our extensive benchmarking provides quantitative, genome-scale evaluation of TRN inference, combining ATAC-seq and RNA-seq data. We refine and extend our previous Th17 TRN, using our new TRN inference methods to integrate all Th17 data (gene expression, ATAC-seq, TF knockouts, and ChIP-seq). We highlight newly discovered roles for individual TFs and groups of TFs ("TF-TF modules") in Th17 gene regulation. Given the popularity of ATAC-seq, which provides high-resolution with low sample input requirements, we anticipate that our methods will improve TRN inference in new mammalian systems, especially *in vivo*, for cells directly from humans and animal models.

[Supplemental material is available for this article.]

Advances in genome-scale measurement and mathematical modeling herald opportunities for high-quality reconstruction of transcriptional regulatory networks (TRNs). TRNs describe the control of gene expression patterns by transcription factors (TFs) (Hecker et al. 2009; Chai et al. 2014), providing mechanistic (and often genome-scale) insight into the complex regulation of cellular behavior (Bonneau et al. 2007). Measurements of chromatin state represent one such advance for TRN inference. For example, chromatin immunoprecipitation with sequencing (ChIP-seq) (Robertson et al. 2007) enables identification of an individual TF's binding sites genome-wide. These data provide evidence for regulatory interactions based on proximity of the TF binding site to the gene locus and have proved valuable for TRN inference (Lee et al. 2002; Ouyang et al. 2009; Ciofani et al. 2012). However, ChIP-seq might not be feasible for cell types and physiological settings in which sample material and a priori knowledge of key transcriptional regulators are scarce.

Genome-scale chromatin accessibility measurements (Giressi et al. 2007; Xi et al. 2007; Boyle et al. 2008; Buenrostro et al. 2013) and ChIP-seq for histone marks (Barski et al. 2007) correlate

with promoters, enhancers, and/or other locus control regions. These data can partially overcome limitations in a priori knowledge of cell-type-specific TF regulators if integrated with TF DNA-binding motifs (Pique-Regi et al. 2011). Large-scale efforts to characterize TF motifs are ongoing, with motifs currently available for approximately 1000 TFs in human (~60% coverage) (Jolma et al. 2010; Weirauch et al. 2014; Lambert et al. 2018). Thus, chromatin state experiments integrated with TF motif analysis provide indirect DNA-binding evidence for hundreds of TFs. This scale would be difficult to attain from individual TF ChIP-seq experiments. Of techniques available, the assay for transposase-accessible chromatin (ATAC)-seq (Buenrostro et al. 2013) best overcomes limitations in sample abundance, requiring two orders of magnitude fewer cells than a typical ChIP-seq, FAIRE-seq, or DNase I hypersensitive sites (DHS) experiment in standard, widely adopted protocols. ATAC-seq is also possible at single-cell resolution (Buenrostro et al. 2015).

In the context of TRN inference, chromatin state measurements provide an initial set of putative TF-gene interactions based on evidence of TF binding near a gene locus. Evidence, be it direct (TF ChIP-seq) or indirect (e.g., TF motif occurrence in accessible chromatin), can be used to refine gene expression modeling (Qin et al. 2014; Blatti et al. 2015; Wilkins et al. 2016). Integration of chromatin state data in TRN inference could mitigate false-positive

⁹Present address: Novartis Institutes for Biomedical Research, Cambridge, MA 02139, USA

Corresponding authors: emily.miraldi@cchmc.org; dan.littman@med.nyu.edu; rbonneau@flatironinstitute.org

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.238253.118>. Freely available online through the *Genome Research* Open Access option.

© 2019 Miraldi et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

and false-negative TF-gene interactions expected from chromatin-state data analyzed in isolation (Äijö and Bonneau 2016; Siahpirani and Roy 2016). Sources of false positives and negatives include (1) nonfunctional binding, (2) long-range interactions between genes and regulatory regions, (3) the limited availability of individual TF ChIP experiments and incomplete knowledge of TF motifs, and (4) nonbound accessible motifs. Thus, an initial TRN derived solely from chromatin state data can be considered a useful but noisy prior, to be integrated with other data types for TRN inference.

Genome-scale inference of TRNs in mammalian settings is an outstanding challenge, given the increased complexity of regulatory mechanisms relative to simpler eukaryotes. Thus, chromatin state data are especially important for mammalian TRN inference. Construction of a genome-scale TRN for T Helper Cell Type 17 (Th17) differentiation provided a proof-of-concept for this idea (Ciofani et al. 2012). Rich genomics data sets informed the Th17 TRN: 143 RNA-seq experiments (including knockout [KO] of 20 TFs), ChIP-seq of nine TFs, and microarray from the Immunological Genome Project (Heng et al. 2008). We used the Inferelator algorithm (Bonneau et al. 2006; Madar et al. 2010) to infer TRNs from the RNA-seq and microarray and used independent methods to build networks from TF ChIP and KO data. We showed that rank combination of the networks performed best at recovering known Th17 genes and GWAS disease genes associated with Th17 pathologies.

Since the original Th17 TRN publication, the Inferelator algorithm underwent developments that improve inference in unicellular organisms and are expected to improve TRN inference in a mammalian setting (Greenfield et al. 2013; Arrieta-Ortiz et al. 2015). Although the Inferelator's core model of transcriptional regulation still describes differential gene expression as a sparse multivariate linear function of TF activities (TFAs), the methods to solve for the TF-gene interaction terms and estimate TFAs have advanced. For example, the current Inferelator (Arrieta-Ortiz et al. 2015) uses a Bayesian approach to incorporate prior information (Greenfield et al. 2013).

The focus of this work is development of mammalian TRN inference methods from chromatin accessibility and gene expression, data types available or likely feasible for an ever-growing number of cell types, and biological conditions. In the context of mammalian TRN inference, several studies build TRNs directly from chromatin accessibility without further refinement by multivariate gene expression modeling (Neph et al. 2012; Rendeiro et al. 2016). Several other studies leverage variance in paired RNA-seq and ATAC-seq data sets; these TRN methods are exciting developments but require that ATAC-seq data for all or most RNA-seq conditions (Duren et al. 2017; Karwacz et al. 2017; Ramirez et al. 2017). In contrast, the present work is geared for TRN inference from RNA-seq

and ATAC-seq, in which ATAC-seq need not exist for more than one gene expression condition.

Development of any TRN inference method requires a comprehensive benchmark with a realistic experimental design, a recurrent challenge in computational biology. We previously developed substantial genomic resources in Th17 cells (Ciofani et al. 2012), and with the addition of ATAC-seq to these resources, Th17 could be a powerful system to compare network inference from RNA-seq and ATAC-seq to a “gold standard” (GS) network constructed through the more laborious approach of TF ChIP-seq and TF KO RNA-seq. Given the central role of Th17 cells in the etiology of autoimmune and inflammatory diseases (Littman and Rudensky 2010; Stadhouders et al. 2018), an updated map of transcriptional regulation in Th17 (incorporating new experimental and computational advances) could also enhance our understanding of Th17 biology in health and disease.

Results

Construction of Th17 benchmark for TRN inference from ATAC-seq and RNA-seq

To test the feasibility of TRN inference from chromatin accessibility and gene expression alone, we generated an ATAC-seq data set in Th17 cells and other in vitro polarized T Helper (Th) cells, matching a subset of experimental conditions from the original publication (Ciofani et al. 2012) (Fig. 1A). We identified 63,049 accessible regions (peaks), and clustering revealed that most

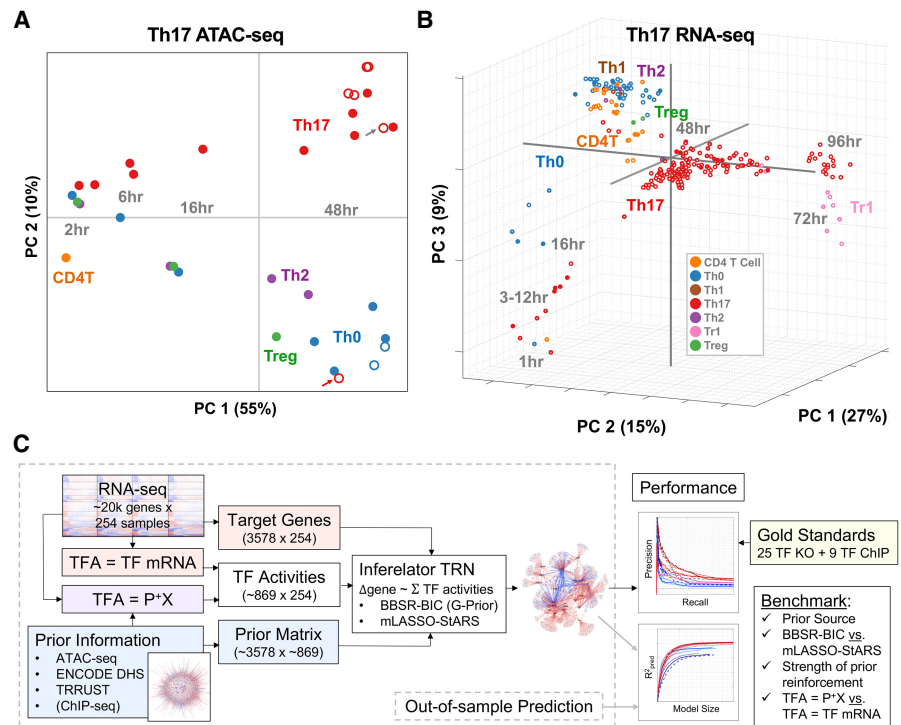


Figure 1. New and existing genomic benchmark resources for TRN inference in Th17. (A) PCA of chromatin accessibility profiles. The 33 ATAC-seq samples are plotted as a function of ATAC-seq peak intensities in PCA space, using the reference set of 63,049 ATAC-seq peaks identified. Open circles denote experimental conditions that deviate from the standard T cell differentiation conditions (e.g., gene deletion, additional cytokines). Gray and red arrows indicate *Maf* and *Stat3* KO Th17 conditions, respectively. (B) PCA of gene expression profiles. The 254 RNA-seq samples are plotted as a function of all genes in PCA space. (C) Study design (see text).

dynamically changed over the Th polarization time courses (Supplemental Fig. S1). These patterns are also apparent from principal component analysis (PCA) (Fig. 1A). Time was the most important driver of accessibility patterns. The first principal component (PC) explained 55% of the variance and captured peaks changing from two to 48 h in Th17, Th0, Th2, Treg. The second PC captured accessibility differences between Th17 and the other Th polarizations. The ATAC-seq data set contains additional perturbations, including TF KO of *Stat3* and *Maf* for Th17 and Th0 conditions (48 h). STAT3 is required for Th17 differentiation, and *Stat3* KO dramatically altered Th17 chromatin accessibility, leading to a Th0-like profile (Fig. 1A, red arrow; Supplemental Fig. S1), whereas *Maf* KO clustered with Th17 (Fig. 1A, gray arrow; Supplemental Fig. S1).

To the 143 RNA-seq experiments from the original publication, we added an additional 111 RNA-seq experiments for a total of 254 (Methods) (Fig. 1B). The majority (166 samples) were Th17, spanning 1–108 h and involving KO, siRNA knockdown, and/or drug inhibitors of TFs and signaling molecules. The study design also included other Th polarizations (Th0 [53], Tr1 [9], and Th1/Th2/Treg [two each]), as well as naive CD4⁺ T cells (25). Mirroring chromatin accessibility patterns, PCA of the gene expression data revealed time and T cell polarization conditions to be important drivers of transcriptional variation (Fig. 1B; Supplemental Fig. S2).

Although gene expression data are the only required input for the Inferelator, we hypothesized that inclusion of ATAC-seq data would improve TRN inference. Integration of ATAC-seq with TF motifs provides indirect evidence for the TF binding events driving altered chromatin state (Fig. 1A) and transcription (Fig. 1B). We generated two “prior” networks of TF–gene interactions from the ATAC-seq data: the A(Th17) prior, limited to Th17 48-h conditions, and the A(Th) prior, including all 33 Th samples (Methods). The ATAC priors contained more than 1 million putative TF–gene interactions for approximately 800 TFs (Supplemental Table S1A). To test these noisy priors in TRN inference, we developed a study design enabling quantitative performance evaluation of the resulting TRNs (Fig. 1C). Specifically, the TRNs are evaluated based on precision recall of an independent GS composed of TF–gene interactions supported by TF KO and TF ChIP data. As precision recall is limited to the TFs previously selected for KO (25 TFs) and/or ChIP-seq (nine TFs), we also evaluate TRN methods based on out-of-sample gene expression prediction.

By using precision recall and out-of-sample prediction metrics, we evaluate the effects of several key modeling decisions. Figure 1C outlines inputs to the Inferelator algorithm. From the gene expression data set, we seek to model the expression patterns of 3578 “target” genes (Methods) as functions of TFAs. Protein TFAs are rarely measured and technically infeasible for most TRN experimental designs. Thus, TFA is a hidden (or latent) variable in TRN inference (Liao et al. 2003; Fu et al. 2011). TF mRNA is the most common TFA estimate. However, many TF transcriptional activities require protein posttranslational modification. Thus, TF mRNA can be a poor proxy for protein TFA. TFA estimation based on prior knowledge of TF target genes provides an alluring alternative as it appears to be technically feasible, requiring only partial a priori knowledge of TF–gene interactions and gene expression data (Methods). “Prior-based” TFAs improved TRN inference in unicellular organisms (Arrieta-Ortiz et al. 2015; Tchourine et al. 2018). Here, we evaluate this approach in a mammalian setting.

We test two methods for model building: (1) Bayesian best subset regression with Bayesian information criteria for model selection (BBSR-BIC) (Arrieta-Ortiz et al. 2015) and (2) an alternative proposed here, modified least absolute shrinkage and selection op-

erator (Studham et al. 2014; Gustafsson et al. 2015) with stability approach to regularization selection (Liu et al. 2010) (mLASSO-StARS). We hypothesized that mLASSO-StARS would scale better with the increased transcriptional complexity of a mammalian setting (e.g., requiring larger models; Methods). Thus, we compare mLASSO-StARS and state-of-the-art BBSR-BIC.

Prior information can enter the inference procedure at two steps: (1) to estimate prior-based TFA (described above) and (2) to reinforce prior-supported TF–gene interactions at the multivariate regression step, using BBSR-BIC or mLASSO-StARS (Fig. 1C). The strength of prior reinforcement is an important TRN inference parameter; it controls the relative contribution of the prior (e.g., TF ChIP, ATAC-seq motif analysis) to evidence from the gene expression model (variance explained by individual TFs). Thus, we test several levels of reinforcement in our study design and compare sources of prior information, in addition to ATAC-seq.

mLASSO-StARS improves inference of a mammalian TRN

As illustrated in Figure 1C, we use precision recall to evaluate the impact of modeling decisions on TRN inference. This analysis depends on the quality of the GS. Both TF KO and TF ChIP-seq GS have caveats. Differential expression analysis of TF KOs yields an imperfect GS, as cellular TRNs adapt to the KO over time. Paralog compensation can lead to false negatives, whereas regulators downstream from the knocked-out TF can lead to secondary gene expression changes (false positives). The TF ChIP GS will also contain false positives (ChIP-seq peaks are not necessarily functional) and false negatives (peak–gene associations are based on linear proximity). Generating a GS from edges supported by both TF KO and TF ChIP reduces false positives but at the expense of false negatives. Thus, precision recall is a nuanced metric of method quality. For each GS, Supplemental Table S1A summarizes the number of edges, TFs, and target genes, as well as the percentage of overlap with other priors. Because we have KO data for 25 TFs but KO + ChIP data for only nine TFs, we also evaluate precision recall of the KO GS.

As expected, a TF ChIP-seq prior improves Th17 TRN inference (Fig. 2A, left). The ATAC prior boosts performance relative to the “no prior” control TRNs (Fig. 2A, central and right; Supplemental Fig. S3). In comparison to the ChIP-seq prior, the boost from the ATAC prior on the KO GS is smaller, likely reflecting increased levels of noise (e.g., from motif-based TF binding prediction). Also, in contrast to ChIP prior results, increasing the strength of prior reinforcement from moderate to high yields no advantage for the noisier ATAC prior. This suggests ATAC-seq prior reinforcement should be limited to moderate rather than high; gene expression data should be relied on to select a small subset of the regulatory hypothesis from the ATAC-seq prior network. For similar levels of prior reinforcement, prior-based TFA models outperform TF mRNA at low recall. For all ATAC TRNs and both GSs, mLASSO-StARS outperforms BBSR-BIC.

To explore experimental designs without context-specific chromatin accessibility, we tested two contrasting, publicly available prior information sources. The first is TF motif analysis of ENCODE DHS data from 25 mouse tissues, none of which include Th17 (Stergachis et al. 2014). The second is derived from the curated TRRUST database of human TF–gene interactions (Han et al. 2015). Although the ENCODE DHS prior includes about 1.5 million interactions between 546 TFs and about 17,000 genes (similar scale to the ATAC priors), the TRRUST prior is sparse: about 7000 interactions between 582 TFs and approximately 2000 genes

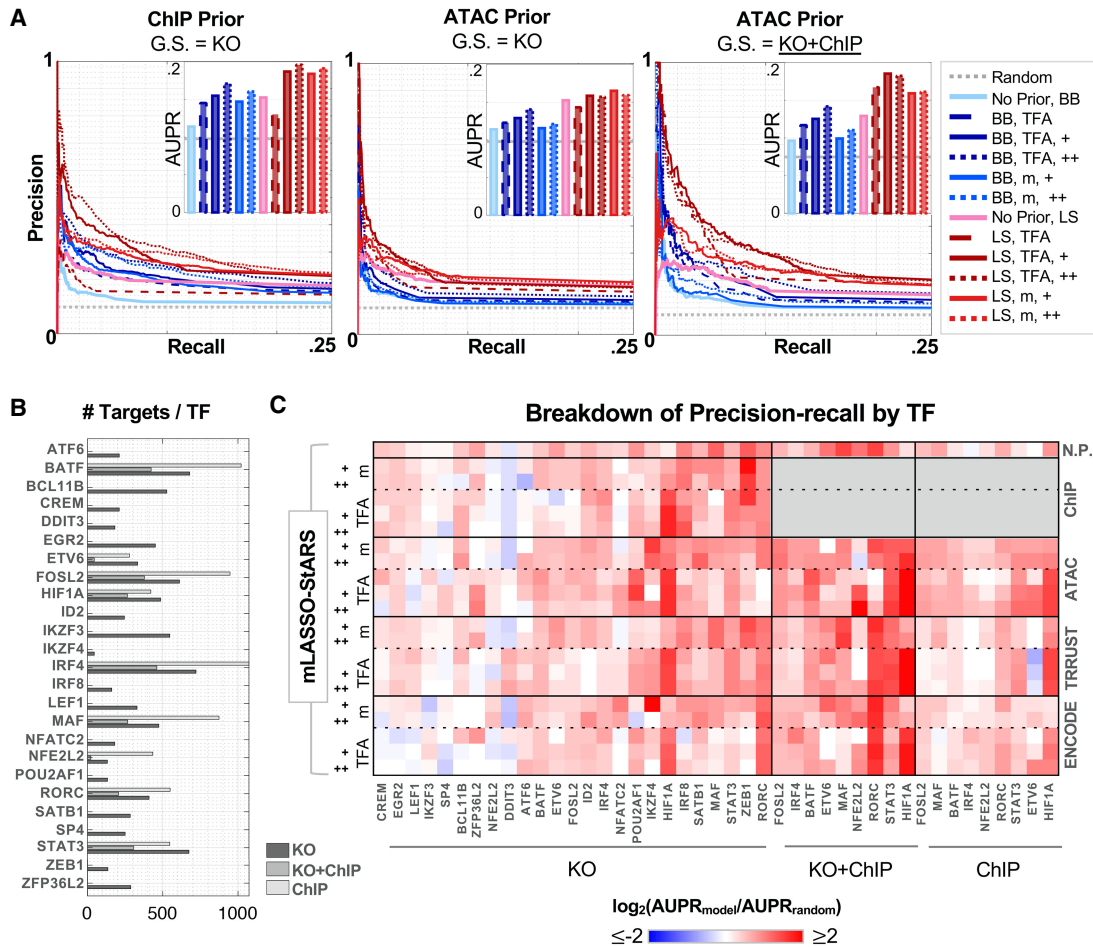


Figure 2. A prior network derived from ATAC-seq data improves TRN recovery of gold-standard TF-gene interactions. (A) Precision recall of Th17 TRNs. The left two panels enable comparison of TRNs built from ChIP versus Th17 ATAC priors, quantified by precision recall of the KO GS (25 TFs, 8875 interactions); insets display AUPR. The performance of several TRNs are plotted for each prior, based on Inferelator method. (LS) mLASSO-StARS; (BB) BBSR-BIC; (m) TF mRNA; (TFA) prior-based TFA; (+) indicates strength of prior reinforcement. Random and “no prior” TRNs serve as references. The right panel shows precision recall of the KO-ChIP GS (nine TFs, 2375 edges) for TRNs built from the Th17 ATAC prior. (B) Number of targets per TF in the GSs. Targets per TF are limited to the 3578 considered by the model. (C) TF-specific TRN performance. For each GS, AUPRs were calculated for each TF individually. TF-specific performance of TRNs is quantified as the \log_2 fold-change between AUPR of the TRN model relative to random. +, m, and TFA are as in A.

(Supplemental Table S1A). The TRRUST and ENCODE priors overlap less with the GSs than context-specific priors, and this is reflected in lower precision recall relative to ChIP and ATAC priors (Supplemental Fig. S3). However, use of either the ENCODE or TRRUST prior improves performance relative to the no prior control; this improvement is substantial for prior-based TFA models. Again, across priors, TFA methods, and levels of prior reinforcement, mLASSO-StARS outperformed BBSR-BIC.

To evaluate performance on experimental designs with fewer gene expression samples, we reduced the gene expression matrix from 254 to 50 randomly selected samples (Supplemental Fig. S4). The reduced sample size had a minor impact on precision recall, especially for context-specific ChIP and ATAC priors. These results bode well for extension of our methods to contexts in which gene expression data are less abundant.

We also evaluated how the different modeling decisions affect target prediction for each TF (Fig. 2C). There is nearly an order-of-magnitude difference in TF degree in the KO+ChIP GS (Fig. 2B), so this per-TF analysis additionally ensured that results were not dominated by a few high-degree TFs. Overall, mLASSO-

StARS also outperformed BBSR-BIC at TF-specific AUPR resolution (Supplemental Fig. S5).

AUPRs for many TFs were dependent on TFA estimation procedure (Fig. 2C). TF mRNA should work well for TFs whose main source of regulation is transcriptional, whereas for TFs regulated by posttranslational modification, prior-based TFA would be preferable. Consistent with this, prior-based TFA models have higher AUPR for STAT3, whereas prior-based TFA did not always improve prediction of RORC targets. With the ChIP prior (which included RORC TF ChIP), prior-based TFA AUPR was on par with TF mRNA AUPR. However, for the noisier ATAC-seq prior, prior-based TFA performed only slightly better than random, whereas TF mRNA models (including the “no prior” control) performed well across GSs. For ATAC-based TRN inference, target prediction for some TFs was better using prior-based TFA (HIF1A, STAT3, NFE2L2), whereas TF mRNA was better for some TFs (RORC, MAF, FOSL2) and roughly equivalent for others. Summarizing across priors and parameter sets, no TFA method dominates (Supplemental Fig. S5B). Based on these results, we later construct “final” Th17 TRNs using both TFA estimation methods.

Th17 TRN models predict out-of-sample gene expression patterns

We next evaluated whether the TRN models could predict out-of-sample gene expression patterns. In contrast to precision recall, gene expression prediction provides the opportunity to evaluate all interactions in the model. This evaluation method is especially important in poorly characterized cellular contexts, for which GSs do not exist. We chose three out-of-sample prediction leave-out sets, each with distinct patterns of gene expression (highlighted in Fig. 3A). The three leave-out sets were “early Th17” (all Th17 time points between 1–16 h; eight samples), “all Th0” (Th0 samples for all time points and perturbations; 53 samples), and “late Th17” (18 Th17 samples from 60 to 108 h after TCR stimulation). For both the BBSR-BIC and mLASSO-StARS methods, we tested prediction over a range of edge confidence values and corresponding model sizes. We quantified performance using r -squared of prediction, R^2_{pred} (Fig. 3B,C); $R^2_{\text{pred}} > 0$ indicates that the model has predictive benefit (Methods). Across all leave-out sets and methods,

out-of-sample prediction improved most as models expanded from average size of one to five TFs/gene (Fig. 3B). Most methods performed similarly well from zero to 10 TFs/gene, with the exception of BBSR-BIC models using prior-based TFA, in which prediction was worse. Predictive performance plateaued at about 10–15 TFs/gene, depending on the leave-out set (Fig. 3B, Supplemental Fig. S6). For model sizes of 10–15 TFs/genes, the mLASSO-StARS models outperformed BBSR-BIC models (Fig. 3B,C). These results, together with precision recall analyses, support mLASSO-StARS over BBSR-BIC for mammalian TRN inference.

Although we recommend StARS edge stabilities to rank interactions, we used the best quality metrics at hand (precision, recall, and R^2_{pred}) to guide selection of model-size cutoff for the “final” Th17 TRN (Methods). These quality metrics are plotted versus model size (Fig. 3D; Supplemental Figs. S6, S7) for TRNs with the Th17 ATAC (“ATAC-only”) or ChIP+KO+ATAC priors. (The ChIP+KO+ATAC prior [Methods] represents our best [combined] source of prior information and is later used to derive our

“final” Th17 TRN.) Once average model sizes reach approximately 15 TFs/gene (Supplemental Fig. S6), predictive performance plateaus, suggesting an average of 15 TFs/gene as a cutoff for edge inclusion in the network. Standardizing network sizes to 53,000 TF–gene interactions (about 15 TFs/gene), we calculated the percentage edge overlap among TRNs built from ATAC, ChIP, KO, ENCODE DHS, TRRUST, and combined priors. For each prior, we considered five modeling modes: prior-based TFA with no, moderate, or strong prior reinforcement and TF mRNA TFA with moderate or strong prior reinforcement. The percentage of shared edges between TRNs ranged from 83% to 10%. We clustered the networks to visualize how the modeling decisions affected resulting TRNs on a global scale (Supplemental Fig. S8; Supplemental Note 1).

“Core” Th17 TRNs contain literature-supported TF–gene interactions

Our primary objective is to assess the feasibility of high-quality TRN inference from gene expression and ATAC-seq data. Therefore, it is important to examine the Th17 TRNs at high resolution. Here, we focus analysis on TRN predictions for 18 “core” Th17 TFs and genes readily familiar to Th17 biologists (Fig. 4). TF–gene interactions in this “core” have been the focus of many studies (Christie and Zhu 2014; Li et al. 2014), which we leverage to evaluate the ATAC-based Th17 core TRNs.

From the literature and the KO-ChIP data (Fig. 4C), there is support for edges between RORC and several key Th17 cytokines and receptors: *Il17a*, *Il17f*, *Il22*, *Il1r1*, and *Il23r*. Two of these interactions

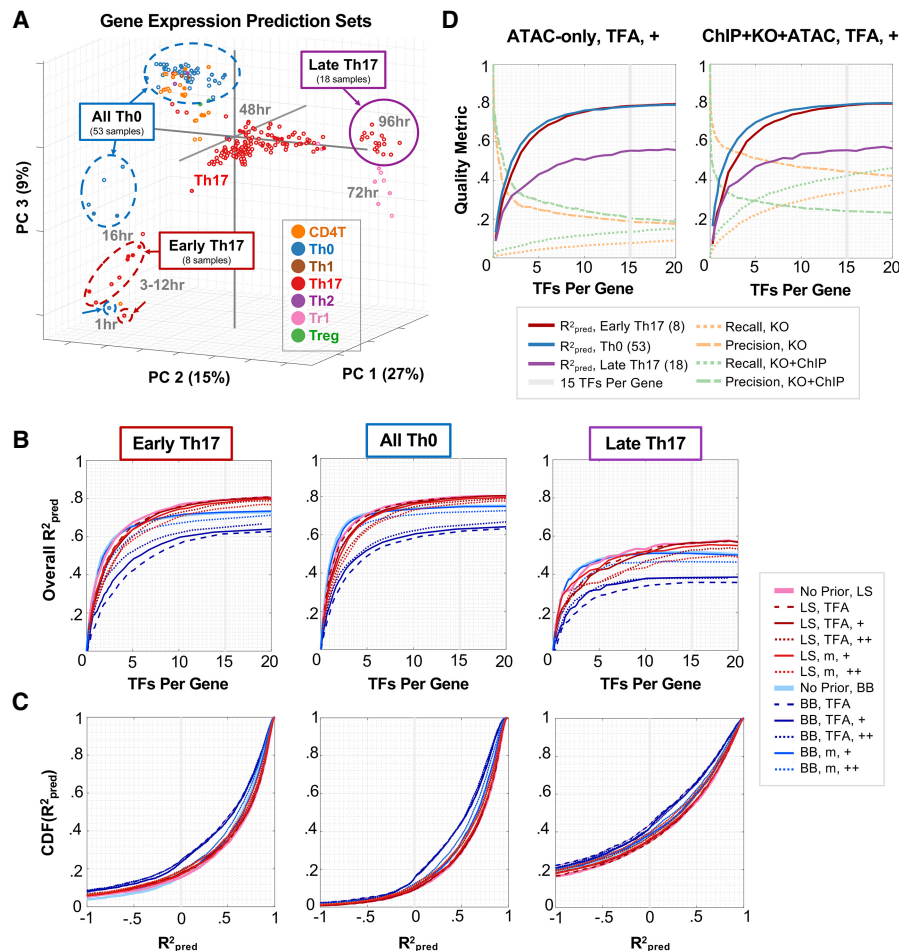


Figure 3. TRNs derived from RNA-seq and ATAC-seq data predict out-of-sample gene expression. (A) Leave-out sets plotted in PCA space. (B) Gene expression prediction. R^2_{pred} for each leave-out set is plotted as a function of mean number of TFs per gene. (LS) mLASSO-StARS; (BB) BBSR-BIC; (m) TF mRNA; (TFA) prior-based TFA; (+) indicates strength of prior reinforcement. The gray line corresponds to a model-size cutoff of mean 15 TFs per gene. (C) Distributions of R^2_{pred} values. Empirical cumulative distribution functions (CDFs) of per-gene R^2_{pred} values for each method (model-size cutoff = mean 15 TFs per gene). (D) Model quality metrics versus model size. For two TRN models built with Th17 ATAC (left) or ChIP+KO+ATAC (right) priors (mLASSO-StARS, bias = 0.5, TFA = P*X), the quality metrics (R^2_{pred} for each leave-out set, precision and recall) are plotted as a function of model size. The model size used for subsequent analyses is highlighted.

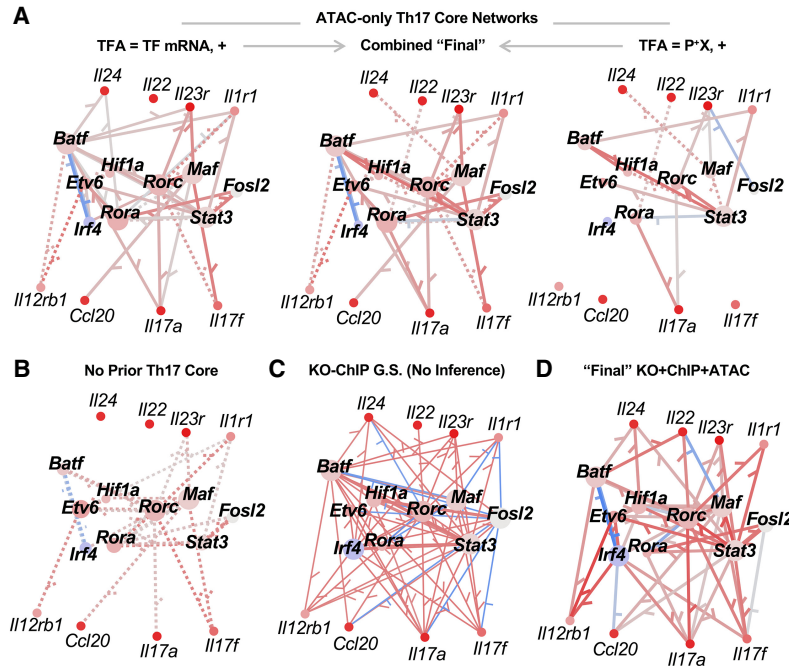


Figure 4. The Th17 TRNs recover key TF–gene interactions from the literature. (A–D) Th17 core TRN models. “Core” Th17 genes and TFs were selected from the literature for visual comparison with *jp_gene_viz* software. Network size was limited to an average of 15 TFs per gene for Inferelator networks using the following: (A) Th17 ATAC prior; (B) no prior; or (D) ChIP + ATAC + KO prior. The edges in Inferelator TRNs are colored according to partial correlation (red indicates positive; blue, negative) and weighted proportionally to edge stability. Solid edges have prior support, whereas dotted edges were learned from gene expression modeling alone. (C) The full KO-ChIP GS from Ciofani et al. (2012), where edge sign is based on differential gene expression analysis between TF KO and control. Nodes are colored according to z-scored gene expression at 48 h in Th17, relative to the other Th cell time points (red/blue indicates increased/decreased expression). The “final” KO + ChIP + ATAC (D) and ATAC-only (B) TRNs max-combine networks built using TF mRNA and prior-based TFA.

(*Il17a*, *Il23r*) were present in the ATAC prior and resulting TRNs (Fig. 4A). Gene expression modeling with TF mRNA was sufficient to recover four interactions (*Il17a*, *Il17f*, *Il1r1*, and *Il23r*) in the no prior TRN Figure 4B; these were also recovered in the ATAC TF mRNA TRN. The prior-based TFA ATAC TRN (Fig. 4A, right) recovers a data-driven edge between RORC and *Il22*. By combining predictions from both ATAC TRNs (Methods), all five RORC targets are recovered (Fig. 4A, central). As expected, inclusion of RORC ChIP and/or KO in the prior also leads to recovery of all five TF targets (Fig. 4D).

STAT3 is required for Th17 differentiation, playing a crucial role in driving *Rorc* expression. There is support for this interaction not only from context-specific ATAC-seq prior but also the ENCODE DHS prior. Consistent with the poor quality of mRNA-based STAT3 predictions (Fig. 2C), this interaction is not present in the no prior TRN. It is, however, recovered by all ATAC and ENCODE TRNs, even those built with TF mRNA, in which prior reinforcement likely overcomes weak correlation between TF mRNA and protein activity level (Fig. 4A–C; Supplemental Fig. S9).

MAF is another key regulator of Th17 cytokine and receptor expression, with KO-ChIP support for *Il17a*, *Il17f*, *Il23r*, and *Il1rb*. There is ATAC-seq support for MAF regulation of *Il17a*, *Il17f*, and *Il23r*. The prior-supported targets are recovered by the TF-mRNA ATAC models, but only the *Il23r* interaction is present in prior-based TFA models (Fig. 4A). Similar to *Rorc*, *Maf* mRNA might be the better proxy for TFA. Prior reinforcement also played a role, as only two of the four interactions are present in the no pri-

or TRN (Fig. 4B). In the absence of context-specific prior information and a strong signal from the gene expression model, only a single edge (*Il23r*) was recovered by one of the ENCODE models (Supplemental Fig. S9).

These results highlight the potential for TRN inference in new settings, in which integration of chromatin accessibility and gene expression is more feasible than sequential TF ChIP and KO experiments. Consistent with the TF-resolved AUPR analysis (Fig. 2C), they also suggest that there is value to building models from both TFA methods. For construction of the “final” Th17 TRN, we combine models based on both TFA methods (Methods). The literature-curated core of our final Th17 TRN contains the RORC and MAF cytokine and receptor interactions highlighted from the literature, as well as the established connection between STAT3 and *Rorc* (Fig. 4D).

ATAC-derived Th17 TRNs contain known and novel Th17 TFs

Having verified that the Th17 TRNs contain core Th17 TF–gene interactions from the literature, we develop a global, unbiased analysis of the final ChIP + ATAC + KO TRN to identify “core” Th17 regulators de novo. In addition, we extend our analysis to a final TRN using the

ATAC-only prior to simulate mammalian TRN inference in less well studied systems, in which KO and/or ChIP data might be unavailable. Overall, prior-supported edges make up 63% and 43% of the ~53,000 TF–gene interactions in ChIP + KO + ATAC and ATAC-only TRNs, respectively (Supplemental Table S5). Of the 715 potential TF regulators considered for final models, nearly all (~95%) have targets in the final TRNs, with positive interactions outnumbering negative nearly twofold (1.8:1 for ChIP + KO + ATAC or 1.9:1 for ATAC-only). TF degree varies dramatically (Supplemental Figs. S10, S11). Whereas the ATAC-only network democratizes TF degree distribution (no TF has more than 500 targets), the addition of ChIP and KO leads to very high degree for several TFs in the ChIP + KO + ATAC TRN (more than 500 targets for IRF4, BATF, MAF, SP4, FOSL2, and STAT3). Although there is a bias for TFs in the prior to have higher degree, several TFs without prior support have more than 100 targets in the final networks (four TFs for the ChIP + KO + ATAC TRN and 14 TFs for the ATAC-only TRN). Thus, prior information is not weighted so strongly as to preclude inclusion of TFs without known motifs. This aspect is important for discovery and holds for TFs with edges in the prior. Although only 16% or 7% of input prior edges remain in the TRN, 27% or 46% of learned regulatory interactions for TFs with prior information are new (not originally in the prior) in KO + ChIP + ATAC or ATAC-only TRNs, respectively. Thus, our method can reduce both false negatives and false positives found in prior networks. For example, motif analysis of the RORC ChIP data revealed that only about one-third of RORC peaks contained a

RORC motif (at motif occurrence cutoff $P_{\text{raw}} = 10^{-4}$). Although RORC can bind DNA directly, the RORC ChIP data suggest that RORC might also bind DNA indirectly (e.g., via TF complexes). Such indirect binding would be difficult to detect by ATAC-seq motif analysis alone, but their gene targets can be recovered via gene expression modeling.

Many of the TFs with highest degree are shared between the ChIP+KO+ATAC and ATAC-only TRNs (e.g., IRF4, BATF, SP4, RXRA, STAT3) (Supplemental Figs. S10, S11). We developed an unbiased approach to identify key regulators of the Th17 program. TFs were included in the set of “core” Th17 regulators if they met one of two criteria: (1) The TF promotes Th17 gene expression through activation of Th17 genes, or (2) the TF promotes Th17 expression through repression of non-Th17 genes (Methods). Similar de novo Th17 core TFs were recovered from both ChIP+KO+ATAC and ATAC-only TRNs, with several recognizable Th17-specific TFs from the literature (RORA, RORC, STAT3, MAF) in both networks (Fig. 5A,B). We note that this “core” TF analysis is robust to model-size cutoffs, as analysis of TRNs with average model-size of five or 10 TFs/gene yields similar results (Supplemental Fig. S12). Similarly, top-degree TFs per TRN are robust across model sizes (Supplemental Fig. S13).

TF–TF modules exhibit coordinated control of gene pathways in Th17

To aid in exploring the large Th17 TRNs (about 53,000 TF–gene interactions), we identified clusters of TFs with significant overlap in target genes (see Methods) (Fig. 5C; Supplemental Figs. S14, S15). We then applied a comprehensive gene-set enrichment analysis to predict functional roles for the “TF–TF clusters,” looking for consensus among pathway enrichments from five databases (Gene Ontology, Pathway Commons, KEGG, WikiPathways, and signatures from MSigDB) (Supplemental Figs. S16, S17; Kanehisa and Goto 2000; Gene Ontology Consortium 2004; Pico et al. 2008; Cerami et al. 2010; Liberzon et al. 2011). Most clusters were conserved between ChIP+KO+ATAC and ATAC-only networks, and within clusters, TFs shared features. For example, several clusters contained TFs defined in the de novo Th17 cores. RORC was a member of a Th17-promoting TF–TF module including RORA, NR1D1, and VAX2 (Fig. 5C, light-blue square; Supplemental Figs. S14–S17); functional annotations for this cluster include “IL23 signaling” and “rheumatoid arthritis,” which are consistent with prior knowledge. Th17-promoting TFs HIF1A, HIF3A, DPF1, SP9, and SCRT1 cluster with five to six other TFs

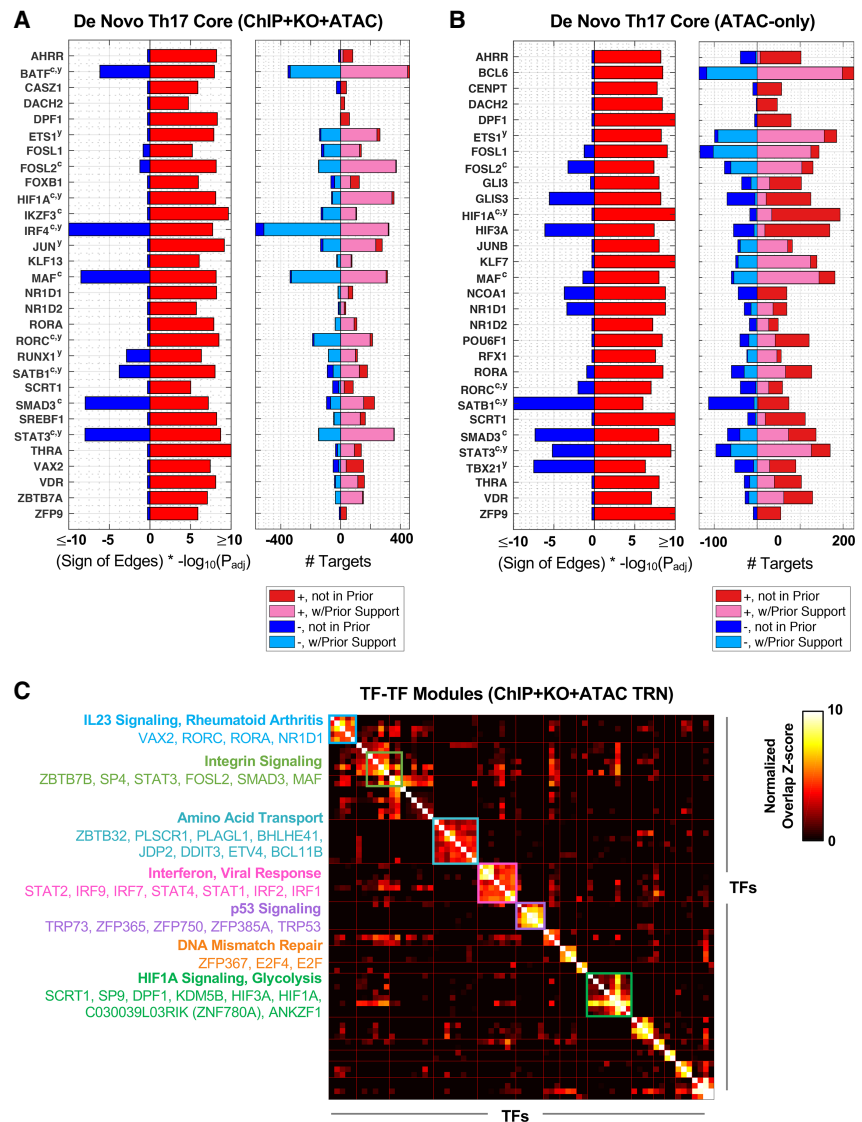


Figure 5. Analysis of the Th17 TRNs expands the “core” Th17 TRNs and predicts multivariate regulation of Th17 gene pathways. De novo Th17 core TFs in the ChIP + KO + ATAC TRN (A) and ATAC-only TRN (B). The Top 30 most significant “core” TFs are displayed. Significance was based on enrichment of a TF’s (1) positive gene targets in up-regulated Th17 genes or (2) negative targets in down-regulated Th17 genes. (Left) Significance and direction of regulation; (right) number, sign, and prior support of TF target edges. Superscripts “c” and “y” indicate TF Th17 association from Ciofani et al. (2012) and Yosef et al. (2013), respectively. (C) Top 15 TF–TF modules for ChIP + KO + ATAC TRN. TFs were clustered into modules based on shared positive target genes between TFs (Methods). Gene-set enrichment was used to annotate clusters, and TF members are listed.

(green square), and enrichments for this cluster include “hypoxia,” “HIF1A transcription factor network,” and “glycolysis.”

Other clusters contained TFs that promote the expression of genes repressed at 48 h in Th17 cells. One such cluster contained Th1 TFs (IRF1, STAT1, and STAT2) with additional interferon response factors and STATs (Fig. 5C, hot-pink box). As expected, this “interferon cluster” has enrichments for “response to interferon gamma,” “type 1 interferon pathway,” “response to virus.” Although TF gene expression for this cluster is highest in Th1 relative to other Th populations at 48 h, gene expression is at its highest at the 1-h Th17 time point, suggesting an interferon-like response for Th17 cells very early in the Th17 polarization time

course (Supplemental Figs. S14, S15). This result is consistent with predictions from another Th17 TRN (Yosef et al. 2013), in which authors also predict roles for IRF1, IRF2, IRF9, STAT1, and STAT2 within the first 4 h of Th17 polarization. Both findings are consistent with potential plasticity, observed in vivo, in Th17 cell programs that are homeostatic or pathogenic, with expression of Th1-like features in the latter.

Gene-set enrichment provides functional predictions for other TF–TF modules, including “amino acid transport,” “integrin signaling,” “DNA mismatch repair,” “p53 signaling,” and others (Fig. 5C; Supplemental Figs. S14–S17). These predictions provide further confirmation of TRN quality, as many modules have predicted function in processes for which individual TFs are already implicated (e.g., HIF1A and HIF3A in the HIF1A/hypoxia module, TRP53 and TRP73 in the “p53 signaling” module). TF–TF modules and functional annotations are largely conserved between KO + ChIP + ATAC and ATAC-only TRNs, the latter prior network being much more economically feasible than the first. More fundamentally, these predictions suggest how altering sets of TFs might influence Th17 pathways and responses.

New phenotypes are associated with TFs in the Th17 TRN

Th17 cells contribute to the pathogenesis of multiple autoimmune diseases (Stadhouders et al. 2018). We previously tested whether genes coregulated by the “Th17 core” (RORC, STAT3, BATEF, IRF4, and MAF) were enriched for gene sets from GWAS of nine autoimmune diseases and three “negative controls” (Alzheimer’s, schizophrenia, and type 2 diabetes) (Ciofani et al. 2012). Consistent with the known role for Th17 in autoimmune disease, genes from the autoimmune-disease sets were enriched (Ciofani et al. 2012). Since then, Th17 cells have also been implicated in obesity-related diseases (Harley et al. 2014; Endo et al. 2017) and psychiatric disorders (Debnath and Berk 2014; Choi et al. 2016). In parallel, the number of genome-wide association studies grew exponentially (MacArthur et al. 2016), and as demonstrated above, our network model improved in both comprehensiveness and accuracy.

We performed an extensive, unbiased GWAS analysis of our “final” updated (KO + ChIP + ATAC) Th17 TRN, including any phenotype with five or more associated genes; 991 phenotypes met this criterion. Not only did we dramatically expand the phenotypes considered, we more broadly queried the Th17 TRN. For each of the 605 TFs individually, we tested for TF–target genes enrichment in each of the GWAS gene sets (Supplemental Table S6). Despite the large number of TF–phenotype associations tested, eight reached significance (FDR = 10%) (Fig. 6A). STAT3 targets were significantly enriched for genes associated with inflammatory bowel disease (IBD), as well as the two IBD-sub-

types, Crohn’s disease and ulcerative colitis. Both genetic (Cho 2008) and functional studies (Xavier and Podolsky 2007) support a role for STAT3 in IBD; indeed, STAT3 is a proposed IBD therapeutic target (Lee et al. 2015; Nguyen et al. 2015). Our analysis also newly implicates FOXB1 in regulation of IBD genes. We compared the centrality of STAT3 and FOXB1 in the Th17 TRN (Supplemental Fig. S18A) to their centrality in the subnetwork limited to the 54 IBD genes in the Th17 TRN (Fig. 6B, left). We examined both degree and betweenness centrality. For each TF, betweenness is the fraction of shortest paths connecting TFs to target genes in the network that contain the TF. Whereas degree is a local measure (TF’s direct effect on gene expression), betweenness is a more global measure of TF importance, as it can also capture TFs that regulate a large number of genes through control of other TFs. Although STAT3 had the sixth-highest degree in the full Th17 TRN (Supplemental Fig. S18A), it has the highest-degree TF in the IBD subnetwork (Fig. 6B). Relative degree more than doubles for both STAT3 and FOXB1 in the IBD subnetwork, and betweenness centrality increased, too (Fig. 6B). The IBD genes regulated by STAT3 and FOXB1 include a number of Th17 genes: *Rorc*, *Il23r*, *Tnfsf15* (Fig. 6B, right).

NFKB2 and ETS1 are also associated with immune phenotypes (Fig. 6A). NFKB2’s targets are enriched in the phenotype “chronic inflammatory diseases (ankylosing spondylitis, Crohn’s disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy)” (Supplemental Fig. S18B). Mutations in *NFKB2* have been previously associated with common variable

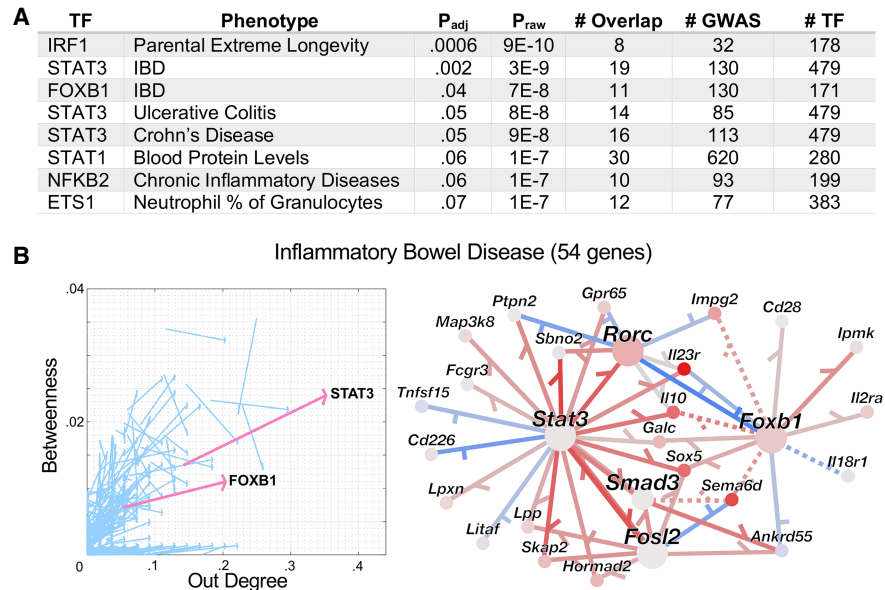


Figure 6. The Th17 TRNs implicate phenotypes and putative regulators in Th17 cells. (A) TFs whose target genes are enriched in GWAS phenotype genes (FDR = 10%). (IBD) Inflammatory bowel disease; (chronic inflammatory diseases) chronic inflammatory diseases (ankylosing spondylitis, Crohn’s disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis; pleiotropy). # GWAS, # TF, and # Overlap correspond to the number of genes associated with the phenotype, regulated by the TF in the Th17 TRN (KO + ChIP + ATAC), and the overlap between those two sets, respectively. Further details are contained in the Methods. (B) STAT3 and FOXB1 are central regulators of IBD genes. (Left) Each arrow corresponds to a single TF. Arrow source is TF’s centrality (out degree, betweenness) in the full Th17 TRN, and arrowhead is TF centrality for the IBD subnetwork (in which target genes are limited to the 54 shared between the Th17 TRN and IBD GWAS set). STAT3 and FOXB1 (pink arrows) both show significant increase in degree centrality for IBD genes (FDR = 10%). (Right) The subnetwork connecting STAT3 and FOXB1 to their target genes in the IBD set. Node color indicates log₂(fold-change) in Th17 48-h condition relative to other Th timepoints (red indicates increased; blue, decreased), whereas red/blue edges indicate positive/negative regulation. Solid edges have support in the ChIP + KO + ATAC prior, whereas dotted edges do not.

immunodeficiency (CVID) (Chen et al. 2013; Lindsley et al. 2014; Liu et al. 2014), a heterogeneous disorder in which 25% of patients suffer autoimmune disorders, including thrombocytopenic purpura, autoimmune hemolytic anemia, rheumatoid arthritis, and autoimmune enteropathy (which can be classified as Crohn's disease) (Cunningham-Rundles 2008; Lopez-Herrera et al. 2012). Thus, *NFKB2* was previously genetically associated with pleiotropic autoimmune diseases, in the context of CVID. (We note that our set of GWAS phenotypes did not include CVID.) ETS1's targets are associated with "neutrophil percentage of granulocytes" (Supplemental Fig. S18C). ETS1 is known to repress the Th17 program (Moisan et al. 2007). *Ets1* expression decreases over the course of both Th0 and Th17 polarization and, of the 48-h Th polarization conditions, has highest expression in Treg. Mutations in *ETS1* have been associated with systemic lupus erythematosus (SLE) (Leng et al. 2011), an autoimmune disease in which the role of neutrophils has become increasingly appreciated (Smith and Kaplan 2015). Thus, a predicted role in neutrophil regulation could be consistent with the known role of ETS1 in SLE.

Discussion

Th17 cells protect mucosa from bacteria and fungi but can also drive autoimmune and inflammatory disease (Khader et al. 2009; Littman and Rudensky 2010; Stadhouders et al. 2018). These diverse roles require coordination of thousands of genes. TF regulation of gene expression provides a map for immuno-engineering Th17 behavior in disease. Researchers in academia and industry have used our first genome-scale Th17 TRN (Ciofani et al. 2012) to develop hypotheses in the context of autoimmunity (Isono et al. 2014; Yang et al. 2014; Patel and Kuchroo 2015). Here, we provide an important update to our knowledge of Th17 transcriptional regulation, enabled by technical advances in genomic measurement and computational advances in TRN inference. KO data for 20 TFs and TF ChIP data for nine TFs were central to the original Th17 TRN, providing excellent coverage of TF-gene targets for TFs in that set. However, technical limitations and cost precluded application of these tools to the hundreds of TFs expressed over the course of Th17 differentiation, all of which could play important roles in Th17 gene expression regulation.

In combination with large-scale efforts to learn TF DNA-binding motifs (Badis et al. 2009; Jolma et al. 2013; Weirauch et al. 2014; Najafabadi et al. 2015), the advent of ATAC-seq represents an opportunity to overcome limitations of sequential TF ChIP experiments, expanding the number of TFs with chromatin binding profiles by over an order of magnitude. In addition, although TF KO and ChIP data were pragmatically limited to 48-h Th17 conditions, standard ATAC-seq protocols require two orders of magnitude fewer cells than TF ChIP. Here, we obtained (indirect) TF binding profiles from multiple differentiation time points. Yet TF binding profiles derived from motif analysis of ATAC-seq are noisy. Here, we provide a single, integrated method to infer regulatory roles for TFs genome-wide. At its core, gene expression is modeled as a function of TF activities, in which prior information (e.g., from ATAC-seq) can be used to (1) improve TF activity estimates for some TFs and (2) favor TF-gene interactions that also have prior support. We rigorously test the performance of our method in terms of precision recall and gene expression prediction. Our methods have two very desirable features (1) they prune initial noisy prior networks (by over an order of magnitude in this study) while (2) also learning new TF-gene interactions for TFs with and without prior information.

Our final Th17 TRN is built integrating our best knowledge (KO and ChIP-seq of key Th17 TFs with ATAC-seq and a rich gene expression data set). Our de novo Th17 core includes the original core (RORC, STAT3, BATAF, IRF4, MAF) and dozens of additional TFs. The TF-TF module analyses predict gene pathway regulation by multiple TFs. We also exhaustively test for the association of TFs with nearly 1000 GWAS phenotypes, uncovering known associations between STAT3 and IBD, as well as several novel TF associations with immune phenotypes. Notably, these TFs were not themselves members of the gene sets for the phenotypes they are predicted to regulate. Thus, application of our TRN methods might provide new links between TF regulators and disease-associated genetic polymorphisms. The resulting Th17 TRN provides an important update to our knowledge of transcriptional regulation in Th17 cells and can be used to query key regulators of pathways and disease genes.

Of perhaps greater importance, the TRN experimental design and computational methods proposed are generalizable, designed for regimes in which prior knowledge of transcriptional regulators and/or sample material is scarce (e.g., cells directly from humans and animal models). Given the rigorous testing and case study presented here, we have high expectations for their successful application in other systems. Indeed, we have already applied our methods to a new physiological setting, constructing and experimentally validating TRNs for innate lymphoid cells of the intestine (Pokrovskii et al. 2018). Our methods are widely applicable. Prior information can be derived from diverse sources: chromatin state data, systems genetics, and literature-curated databases.

This work also highlights avenues for future improvement of TRN inference methods. We tested two methods for TF activity estimation: (1) based on TF mRNA levels and (2) based on prior knowledge of TF-gene interactions. Although prior-based TFA improved TRN inference in *Bacillus subtilis* and yeast (Arrieta-Ortiz et al. 2015; Tchourine et al. 2018), neither method consistently outperformed the other in this study. As a result, final TRNs were built using both TFA methods. There are multiple dimensions along which TFA estimation could be improved. The simplicity of the linear framework proposed for prior-based estimation has limitations in the context of complex mammalian transcriptional regulation, and a more sophisticated mathematical model for TFA estimation could be of value. TFA estimation would also improve from better prediction of TF binding events. Here, we limited our approach to a simple TF motif analysis of accessible chromatin, yet several more sophisticated methods exist and merit testing (Pique-Regi et al. 2011; Sherwood et al. 2014; Chen et al. 2017; Lamparter et al. 2017). Another limitation of our method is the mapping of putative TF binding events to gene loci. In our analysis, 3D distance between potential regulatory regions and gene loci is approximated by linear distance, a shortcoming that chromatin capture data (e.g., Hi-C [Lieberman-Aiden et al. 2009] and other 3D-chromatin techniques [Zhang et al. 2012; Beagrie et al. 2017]) would mitigate. Thus, the Th17 genomics data set (Ciofani et al. 2012), augmented by our new ATAC-seq and RNA-seq experiments, provides a fertile testing ground for the development of future TRN inference methods and innovation.

Methods

ATAC-seq

CD4⁺ T cells were sorted and polarized according the method previously described (Ciofani et al. 2012), and ATAC-seq samples were

prepared as described previously (Buenrostro et al. 2013). Paired-end 50-bp sequences were generated from samples on an Illumina HiSeq 2500. Sequences were mapped to the murine genome (mm10) with Bowtie 2 (2.2.3) (Langmead and Salzberg 2012), filtered based on mapping score (MAPQ>30, SAMtools [0.1.19] [Li et al. 2009]), and duplicates removed (Picard; <http://broadinstitute.github.io/picard>). The ATACseqQC package (Ou et al. 2018) was used to evaluate ATAC-seq fragment-length distributions and signal at TSS for each sample (Supplemental Fig. S19). For each sample individually, we ran PeakDEck (parameters `-bin 75, -STEP 25, -back 10000, -npBack100000`) (McCarthy and O'Callaghan 2014) and filtered peaks with a $P_{\text{raw}} < 10^{-4}$. To enable quantitative comparison of accessibility across samples, we generated a reference set of accessible regions, taking the union (BEDTools; Quinlan and Hall 2010) of peaks detected in individual samples. The reference set of ATAC-seq peaks contained 63,049 potential regulatory loci, ranging from 75 to 3725 bp (median, 275 bp). Reads per reference peak were counted with HTSeq-count (Anders et al. 2015). ATAC-seq data were robustly normalized using DESeq2 (Love et al. 2014) for PCA and clustering (Fig. 1A; Supplemental Fig. S2). The 33 ATAC-seq experiments are available from NCBI's Gene Expression Omnibus (GEO) Database (GSE113721).

RNA-seq

The 18 new samples composing "late Th17" time points were generated as follows: Naive CD4⁺ T cells were primed on an anti-CD3 (Bio X Cell BE0001-1) and anti-CD28 (Bio X Cell BE0015-1) coated plate (without any additional cytokine) for 12–14 h (overnight). Cells were then polarized with one of two cytokine cocktails: (1) Th17N, TGF- β (0.3 ng/mL, PeproTech 100-21-10)+IL6 (20 ng/mL, eBioscience 34-8061-82); or (2) Th17P, IL6 (20 ng/mL) + IL1b (20 ng/mL, PeproTech 211-11b) + IL23 (20 ng/mL, R&D Systems 1887-ML-010). Cells were harvested for RNA-seq at 60 and 108 h after TCR stimulation. Cells were lysed and snap-frozen in TRIzol then thawed for chloroform extraction, using a 1:1 ratio of 70% ethanol to aqueous phase. Samples were loaded onto a Qiagen RNeasy column according to the manufacturer's instructions. rRNA was depleted with a Ribo-Zero gold kit; libraries were then prepared using the Illumina TruSeq stranded total RNA library prep and sequenced on an Illumina HiSeq 2500. The remaining 81 new CD4⁺ T cell samples were (1) naive or polarized and (2) processed as previously described (Ciofani et al. 2012). The 99 RNA-seq experiments are available from GEO (GSE113720). Publicly available RNA-seq data were downloaded from GEO: GSE40918 (156 samples), GSE70108 (four samples), and GSE92992 (eight samples). Sequences were mapped to mm10 (STAR aligner) (Dobin et al. 2013). Reads per gene were counted (using HTSeq-count [Anders et al. 2015] with parameters `--stranded=no --mode=union`) and robustly normalized (DESeq2) (Love et al. 2014). Supplemental Note 2 details treatment of batch effects.

TRN inference

Selection of target genes

We built gene expression models for 3578 target genes, composed of the union of (1) genes differentially expressed between Th17 and Th0 at 48 h (FDR=10%, $\log_2|\text{FC}| > \log_2(1.5)$) and (2) the 2100 genes in the original Th17 TRN (Supplemental Table S2; Ciofani et al. 2012).

Selection of potential regulators

We initially generated a custom list of potential mouse protein TFs, combining (1) mouse and human TFs from TFClass (Wingender

et al. 2014) and genes with the GO annotation "transcription factor activity." (Human TFs were mapped to mouse using the MGI database.) From our list of potential mouse protein TFs (2093 genes), we generated a list of 869 potential TF regulators, limited to TFs with differential gene expression in at least one pairwise comparison between Th17, Th0, Th1, Treg, or Th2 at 48 h (FDR=10%, $\log_2|\text{FC}| > \log_2(1.5)$). This initial list was used for all analyses comparing mLASSO-StARS and BBSR-BIC (Figs. 2, 3B; Supplemental Figs. S3, S6). However, given recent efforts in TF annotation (Lambert et al. 2018), we generated a new list of mouse TFs for subsequent TRN analyses. Lambert et al. (2018) manually curated lists of (1) likely TFs and (2) "likely non-TFs." We converted both lists to mouse. To gain mouse TFs without human orthologs, we integrated with mouse TFs from AnimalTFDB (Zhang et al. 2015) but removed any mouse TFs (70) that were "likely non-TFs." Our final mouse TF list contained 1577 TFs, 715 of which served as potential regulators (differentially expressed as described above). Both candidate TF lists are available (Supplemental Table S3).

Generation of prior matrices

ATAC-seq peaks were associated with putative TF binding events and target genes to generate a "prior" network, $P \in \mathbb{R}^{|\text{genes}| \times |\text{TFs}|}$, of TF-gene interactions. We used a compendium of human and mouse TF motifs. Human and/or mouse TF binding motifs (PWMs) were downloaded from the Cis-BP motif collection version 1.02 (Weirauch et al. 2014; <http://cisbp.cibr.utoronto.ca>) and the ENCODE motif collection (Kheradpour and Kellis 2014; <http://compbio.mit.edu/encode-motifs>). Transfac version 2014.2 motifs (Wingender 2008), referenced in the human Cis-BP collection, were reformatted with the MEME Suite tool `transfac2meme` version 4.10.1. Human ENCODE motifs were added to the Cis-BP motif collection if the TF PWM had $R^2 < 0.95$ with a Cis-BP entry for that TF. The combined human ENCODE and Cis-BP set were mapped to mouse orthologs. We scanned peaks for individual motif occurrences with FIMO (parameters `--thresh .00001, --max-stored-scores 500000`, and a first-order-Markov background model) (Grant et al. 2011). We found inclusion of human TF orthologs from the ENCODE motif collection slightly increased precision recall relative to mouse Cis-BP alone (Supplemental Fig. S20). TF motif occurrences with raw P -value $< 10^{-5}$ were included in downstream analysis. Putative binding events were associated with a target gene, if the peak fell within ± 10 kb of gene body. We tested several peak-gene association rules based on distance from gene body or TSS, and TRN inference was robust to that choice (Supplemental Fig. S21). We generated two ATAC-seq priors: (1) A(Th17), for which only peaks from Th17 48 h wild-type conditions were included, and (2) A(Th), for which all Th samples were included. For the resulting prior matrix of TF-gene interactions, entries were one if a TF motif was found proximal to the gene and zero otherwise. Similar methods were used to derive priors from ChIP-seq, TRRUST, ENCODE DHS, and combined sources (Supplemental Note 3).

Inference framework

We used the Inferelator model for TRN inference (Bonneau et al. 2006). At steady state (consideration of time-series is discussed in Supplemental Note 4, Supplemental Fig. S23), gene expression is modeled as a sparse, multivariate linear combination of TFAs:

$$x_{ij} = \sum_{k \in \text{TFs}} b_{ik} a_{kj}, \quad (1)$$

where x_{ij} corresponds to the expression level of gene i in condition j , a_{jk} is the activity of TF k in condition j , and b_{ik} describes the effect

of TF k on gene i . A TF's mRNA expression can serve as a proxy for protein TFA. More recently (Arrieta-Ortiz et al. 2015), TFA has been estimated based on partial prior knowledge of a TF's gene targets:

$$X = PA, \quad (2)$$

where $X \in \mathbb{R}^{|\text{genes}| \times |\text{samples}|}$ is the expression matrix for genes in the prior, $P \in \mathbb{R}^{|\text{genes}| \times |\text{TFS}|}$ is the prior matrix of known TF-gene interactions, and $A \in \mathbb{R}^{|\text{TFS}| \times |\text{samples}|}$ contains the unknown TFAs. Equation 2 has no unique solution, but the least-squares solution has worked well in simpler organisms (Arrieta-Ortiz et al. 2015; Tchourine et al. 2018). Given first-time application to a mammalian setting, we tested both methods of TFA estimation: (1) TF mRNA levels and (2) prior-based (Equation 2). Note that all expressed genes (24,007) with edges in the prior were used to solve Equation 2. As described in results, we solved for the interaction terms $\{b_{ik}\}$ in Equation 1 using the current Inferelator (BBSR-BIC) (Arrieta-Ortiz et al. 2015), as well as a new method (mLASSO-StARS; detailed below).

Model-building with LASSO and StARS

We constructed sparse models of gene expression using a modified LASSO framework:

$$\hat{B} = \underset{B}{\text{argmin}} |X - BA|_2^2 + |\Lambda \circ B|_1, \quad (3)$$

where X and A are defined as above, $B \in \mathbb{R}^{|\text{genes}| \times |\text{TFS}|}$ is the matrix of inferred TF-gene interaction coefficients, $\Lambda \in \mathbb{R}^{|\text{genes}| \times |\text{TFS}|}$ is a matrix of nonnegative penalties, and \circ represents an entry-wise matrix product (Studham et al. 2014; Gustafsson et al. 2015). Matrix representation of the LASSO penalty enables incorporation of prior information. Specifically, a smaller penalty, Λ_{ik} , is used if there is evidence for the TF-gene interaction in the prior matrix. Similar to the G-prior in the current Inferelator BBSR and older Inferelator modified elastic net framework (Greenfield et al. 2013), this procedure encourages selection of interactions supported by the prior if there is also support in the gene expression data. For this study, the entries of the Λ matrices were limited to two values: the nonnegative value λ , for TF-gene interactions without evidence in the prior, and $\text{bias} * \lambda$, where $\text{bias} \in [0,1]$, for TF-gene interactions with support in the prior.

We hypothesized that a data-driven approach to model selection might perform better in a complex, mammalian setting than a theoretical one (e.g., BIC used in Inferelator-BBSR-BIC). Specifically, we chose to test StARS (Liu et al. 2010), anticipating that the resulting networks would be larger than those built using our BBSR-BIC method. We hypothesized that a larger model might be needed to describe a mammalian TRN. StARS was designed to ensure that the inferred network of interactions includes the true set of network interactions with high probability. In contrast, another popular data-driven λ selection method, stability selection, seeks to limit false-positive rate (Meinshausen and Bühlmann 2010), which in a biological setting might be overly conservative (Liu et al. 2010). Thus, StARS seemed ideally suited to our objective.

In brief, StARS rests on the definition of edge instabilities. For a fixed value of λ , instabilities are estimated via subsampling and can be interpreted as twice the Bernoulli variance of a subsampled edge or the fraction of times subsample edge predictions disagree (Liu et al. 2010). This definition is used to select the smallest λ value corresponding to an acceptable average edge instability; authors heuristically recommend an average instability cutoff = 0.05.

Importantly, given our application of StARS in the new setting of TRN inference and a modified LASSO objective function, we used out-of-sample gene expression prediction and precision

recall of GS interactions to guide selection of an appropriate instability cutoff rather than relying on the recommended heuristic (Supplemental Note 5; Supplemental Figs. S24–S29). Following our previous work in ecological network inference (Kurtz et al. 2015), TF-gene interactions were ranked according to nonzero subsamples per edge. To better and more efficiently prioritize high-confidence edges for TRN inference (see Supplemental Note 5; Supplemental Figs. S26–S28), we developed the following edge confidence score:

$$\text{Confidence}(i,k) = \text{Nonzero Subsamples} + |pcorr(i,k)|, \quad (4)$$

where i and k correspond to gene i and TF k , and $pcorr(i,k)$ is the partial correlation between gene i and TF k , for a set model size. For TF-gene interactions with the same number of nonzero subsamples, the TF-gene interaction with higher absolute partial correlation will be higher confidence.

mLASSO-StARS was implemented in MATLAB R2016b, and code relies on the Glmnet for MATLAB package (http://www.stanford.edu/~hastie/glmnet_matlab/) to solve Equation 3. Computational speed-ups using bStARS (Müller et al. 2016) are discussed in Supplemental Note 6 and Supplemental Fig. S30. Code is available in Supplemental Materials and from https://github.com/emiraldi/infTRN_lassoStARS.git.

Prior reinforcement

We tested several levels of prior reinforcement (none, moderate, and high) for BBSR-BIC and mLASSO-StARS. For BBSR-BIC, these corresponded to G-prior weights of 1, 1.1, and 1.5, for mLASSO-StARS, $\text{bias} = 1, 0.5, 0.25$, respectively. These prior-reinforcement parameters resulted in commensurate levels of TRN prior-edge incorporation between methods (Supplemental Fig. S22).

Gene expression prediction

We generated three leave-out (test) data sets (Fig. 3A; Supplemental Table S4). For each leave-out prediction challenge, the training set included all samples excluding test. For each training set, we performed model selection and parameter estimation independently of the test set. Both BBSR-BIC and the mLASSO-StARS methods provide confidence estimates for predicted TF-gene interactions, and we built TRN models of various sizes as a function of edge confidence cutoffs for each of the training sets. For parameter estimation, training TFA matrices were mean-centered and variance-normalized according to the training-set means $\bar{a}_{\text{train}} \in \mathbb{R}^{|\text{TFS}|}$ and standard deviations $\bar{\sigma}_{\text{train}}^a \in \mathbb{R}^{|\text{TFS}|}$. Target gene expression vectors were mean-centered according to the training-set mean $\bar{x}_{\text{train}} \in \mathbb{R}^{|\text{genes}|}$. Then, for each confidence-level cutoff, we regressed the vector of normalized training gene expression data onto the reduced set of normalized training TFA estimates to arrive at a set of multivariate linear coefficients $B_{\text{train}} \in \mathbb{R}^{|\text{genes}| \times |\text{TFS}|}$. Sum of squared error of prediction was calculated as follows:

$$SSE_{\text{pred}} = \sum_{\substack{i \in |\text{genes}| \\ j \in |\text{test}|}} \left(x_{ij} - \sum_{k \in |\text{TFS}|} b_{ik, \text{train}} \left(\frac{a_{kj} - \bar{a}_{k, \text{train}}}{\bar{\sigma}_{k, \text{train}}^a} \right) - \bar{x}_{i, \text{train}} \right)^2. \quad (5a)$$

The “null” model SSE was calculated relative to the mean of training data:

$$SSE_{\text{null}} = \sum_{\substack{i \in |\text{genes}| \\ j \in |\text{test}|}} (x_{ij} - \bar{x}_{i, \text{train}})^2. \quad (5b)$$

We then calculated R^2_{pred} , a normalized measure of predictive performance:

$$R^2_{pred} = 1 - \frac{SSE_{pred}}{SSE_{null}}, R^2_{pred} \in (-\infty, 1]. \quad (5c)$$

For gene expression prediction with prior-based TFA, the mRNA of target genes with edges in the prior contribute to TFA estimation (Equation 2), and then their gene expression patterns are predicted as a function of TFA. This circularity did not lead to overfitting and inflated R^2_{pred} values (Supplemental Note 7; Supplemental Figs. S31, S32).

Final TRNs

To generate “final” TRNs, we used mLASSO-StARS with the following parameters: moderate prior reinforcement (bias=0.05) and λ corresponding to average network instability=0.05 to rank edges by confidence. To ensure that our final Th17 TRN was as complete and accurate as possible, our final network edge inclusion criteria was context-specific, guided by the most rigorous tools at hand, precision recall and out-of-sample gene expression prediction (see supporting results and discussion in Supplemental Note 5). We included the highest confidence edges until we reached a model size of average 15 TFs/gene (3578 genes \times 15 TFs/gene = 53,670 TF–gene interactions). Given the complementary performance of TF–mRNA and prior-based TFA, we combined resulting TRNs by taking the maximum edge confidence to preserve the individual strengths of each (Kittler et al. 1996; Castro et al. 2019). See Supplemental Note 8 and Supplemental Figures S33 and S34 for performance comparison of max- to rank-combine (Marbach et al. 2012) relative to individual TRNs as well as performance combining TRNs from different priors.

De novo Th17 core

Th17 TFs were limited to TFs specifically promoting Th17 gene expression patterns. TFs were included in the core if they met one of two criteria: The TF promotes Th17 gene expression through (1) activation (the TF’s positive edges are enriched in up-regulated Th17 genes at an FDR=1%) or (2) repression of non-Th17 genes (TF’s negative edges are enriched in down-regulated Th17 genes at an FDR=1%).

Gold standards

For the GSs from our laboratory, we used recommended cutoffs of 0.75, 0.75, and 1.5 for KO, ChIP, and KO + ChIP networks, respectively (Ciofani et al. 2012). For the six additional TF KO experiments, we downloaded networks without filtering (Yosef et al. 2013). For both GSs, gene symbols were mapped from mm9 to mm10, and only genes mapping to both genome builds were considered in precision recall analysis. Random AUPR was calculated as the ratio of total GS edges to the number of possible edges between target genes and TFs in the GS.

Network visualization and availability

Networks were visualized using `jp_gene_viz`, a newly designed interactive interface, based on iPython. Software is available at https://github.com/simonsfoundation/jp_gene_viz. All 36 LASSO-StARS Th17 TRNs (from Supplemental Fig. S8), GSs, and final, combined TRNs are available in a Jupyter-notebook binder: https://mybinder.org/v2/gh/simonsfoundation/Th17_TRN_Networks/master. Both `jp_gene_viz` codebase and TRN notebooks are also included in Supplemental Materials.

TF–TF module analysis

We calculated the number of shared target genes between each pair of TFs, analyzing positive and negative target edges separately. (Edges with $|\text{partial correlation}| < 0.01$ were excluded from analysis, as were TFs with fewer than 20 gene targets.) TFs vary greatly by number of target genes (Supplemental Figs. S10, S11), so we devised an overlap normalization scheme that controlled for the variable number of targets per TF (Supplemental Note 9; Supplemental Fig. S35).

GWAS analysis

The NHGRI-EBI GWAS Catalog v1.0.2 (MacArthur et al. 2016) was downloaded on August 4, 2018. SNPs were mapped to the nearest gene within ± 1 Mbp using the catalog’s “mapped gene(s).” Phenotype-associated gene sets were converted to mouse gene symbols. Sets containing five or more genes (991 sets) were retained. For each TF in the Th17 TRN (KO + ChIP + ATAC prior) with five or more targets (605 TFs), overlap with the GWAS gene sets was calculated and significance estimated using the hypergeometric CDF. Benjamini–Hochberg correction was applied to control for multiple hypothesis testing. Network statistics were calculated in MATLAB R2016b, normalizing degree by total target genes and fraction of shortest paths (betweenness) by total number of paths between TFs and target genes (some of which were also TFs).

Data access

The data sets from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE113723.

Acknowledgments

We thank the Flatiron Institute Scientific Computing Core (I. Fisk) for enabling the computational aspects of this work and the New York University Langone Medical Center Genomics Core (A. Heguy and P. Zappile) for help with sequencing. We thank T. Aijo, M. Weirauch, and R. Taylor for advice on the manuscript and G. Atluri for helpful discussions of the TF–TF module analysis. This work was supported by the Cincinnati Children’s Research Foundation (Trustee Award Grant to E.R.M.), the Simons Foundation (E.R.M., A.W., N.D., N.C., R.B.), U.S. National Institutes of Health (5T32AI100853 to M.P.; R01-DK103358-01 to R.B. and D.R.L.; R01-GM112192-01 to R.B., and T32 CA009161 [Levy] to J.A.H.), the Howard Hughes Medical Institute (D.R.L.), the Colton Center for Autoimmunity (D.R.L.), Crohn’s and Colitis Foundation of America (fellowship to M.C.), Damon Runyon Cancer Research Foundation (Dale and Betty Frey Fellowship to J.A.H.), and the Laura and Isaac Perlmutter Cancer Center (P30CA016087 to A. Heguy).

References

- Åijö T, Bonneau R. 2016. Biophysically motivated regulatory network inference: progress and prospects. *Hum Hered* **81**: 62–77. doi:10.1159/000446614
- Anders S, Pyl PT, Huber W. 2015. HTSeq: a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169. doi:10.1093/bioinformatics/btu638
- Arrieta-Ortiz ML, Hafemeister C, Bate AR, Chu T, Greenfield A, Shuster B, Barry SN, Gallitto M, Liu B, Kacmarczyk T, et al. 2015. An experimentally supported model of the *Bacillus subtilis* global transcriptional regulatory network. *Mol Syst Biol* **11**: 839. doi:10.15252/msb.20156236
- Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzger G, Vedenko A, Chen X. 2009. Diversity and complexity in

- DNA recognition by transcription factors. *Science* **324**: 1720–1723. doi:10.1126/science.1162327
- Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837. doi:10.1016/j.cell.2007.05.009
- Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, Xie SQ, Barbieri M, de Santiago I, Lavitas L-M, Branco MR, et al. 2017. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**: 519–524. doi:10.1038/nature21411
- Blatti C, Kazemian M, Wolfe S, Brodsky M, Sinha S. 2015. Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. *Nucleic Acids Res* **43**: 3998–4012. doi:10.1093/nar/gkv195
- Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol* **7**: R36. doi:10.1186/gb-2006-7-5-r36
- Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, et al. 2007. A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131**: 1354–1365. doi:10.1016/j.cell.2007.10.053
- Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311–322. doi:10.1016/j.cell.2007.12.014
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218. doi:10.1038/nmeth.2688
- Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**: 486–490. doi:10.1038/nature14590
- Castro DM, de Veaux NR, Miraldi ER, Bonneau R. 2019. Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS Comput Biol* **15**: e1006591. doi:10.1371/journal.pcbi.1006591
- Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, Schultz N, Bader GD, Sander C. 2010. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**: D685–D690. doi:10.1093/nar/gkq1039
- Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. 2014. A review on the computational approaches for gene regulatory network construction. *Comput Biol Med* **48**: 55–65. doi:10.1016/j.compbiomed.2014.02.011
- Chen K, Coonrod EM, Kumánovics A, Franks ZF, Durtschi JD, Margraf RL, Wu W, Heikal NM, Augustine NH, Ridge PG, et al. 2013. Germline mutations in *NFKB2* implicate the noncanonical NF- κ B pathway in the pathogenesis of common variable immunodeficiency. *Am J Hum Genet* **93**: 812–824. doi:10.1016/j.ajhg.2013.09.009
- Chen X, Yu B, Carriero N, Silva C, Bonneau R. 2017. Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Res* **45**: 4315–4329. doi:10.1093/nar/gkx174
- Cho JH. 2008. The genetics and immunopathogenesis of inflammatory bowel disease. *Nat Rev Immunol* **8**: 458–466. doi:10.1038/nri2340
- Choi GB, Yim YS, Wong H, Kim S, Kim H, Kim S V, Hoeffler CA, Littman DR, Huh JR. 2016. The maternal interleukin-17a pathway in mice promotes autism-like phenotypes in offspring. *Science* **351**: 933–939. doi:10.1126/science.1240314
- Christie D, Zhu J. 2014. Transcriptional regulatory networks for CD4 T cell differentiation. *Curr Top Microbiol Immunol* **381**: 125–172. doi:10.1007/82_2014_372
- Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, Agarwal A, Huang W, Parkurst CN, Muratet M, et al. 2012. A validated regulatory network for Th17 cell specification. *Cell* **151**: 289–303. doi:10.1016/j.cell.2012.09.016
- Cunningham-Rundles C. 2008. Autoimmune manifestations in common variable immunodeficiency. *J Clin Immunol* **28**: 42–45. doi:10.1007/s10875-008-9182-7
- Debnath M, Berk M. 2014. Th17 pathway-mediated immunopathogenesis of schizophrenia: mechanisms and implications. *Schizophr Bull* **40**: 1412–1421. doi:10.1093/schbul/sbu049
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. doi:10.1093/bioinformatics/bts635
- Duren Z, Chen X, Jiang R, Wang Y, Wong WH. 2017. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci* **114**: E4914–E4923. doi:10.1073/pnas.1704553114
- Endo Y, Yokote K, Nakayama T. 2017. The obesity-related pathology and Th17 cells. *Cell Mol Life Sci* **74**: 1231–1245. doi:10.1007/s00018-016-2399-3
- Fu Y, Jarboe LR, Dickerson JA. 2011. Reconstructing genome-wide regulatory network of *E. coli* using transcriptome data and predicted transcription factor activities. *BMC Bioinformatics* **12**: 233. doi:10.1186/1471-2105-12-233
- Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: D258–D261. doi:10.1093/nar/gkh036
- Giresi PG, Kim J, McDaniel RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**: 877–885. doi:10.1101/gr.5533506
- Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018. doi:10.1093/bioinformatics/btr064
- Greenfield A, Hafemeister C, Bonneau R. 2013. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* **29**: 1060–1067. doi:10.1093/bioinformatics/btt099
- Gustafsson M, Gawel DR, Alfredsson L, Baranzini S, Björkander J, Blomgran R, Hellberg S, Eklund D, Ernerudh J, Kockum I, et al. 2015. A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. *Sci Transl Med* **7**: 313ra178. doi:10.1126/scitranslmed.aad2722
- Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, Kim H, Cho A, Kim E, Lee T, et al. 2015. TRRUST: a reference database of human transcriptional regulatory interactions. *Sci Rep* **5**: 11432. doi:10.1038/srep11432
- Harley ITW, Stankiewicz TE, Giles DA, Softic S, Flick LM, Cappelletti M, Sheridan R, Xanthakos SA, Steinbrecher KA, Sartor RB, et al. 2014. IL-17 signaling accelerates the progression of nonalcoholic fatty liver disease in mice. *Hepatology* **59**: 1830–1839. doi:10.1002/hep.26746
- Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. 2009. Gene regulatory network inference: data integration in dynamic models: a review. *Biosystems* **96**: 86–103. doi:10.1016/j.biosystems.2008.12.004
- Heng TSP, Painter MW, Elpek K, Lukacs-Kornek V, Mauermann N, Turley SJ, Koller D, Kim FS, Wagers AJ, Asinowski N. 2008. The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* **9**: 1091. doi:10.1038/ni1008-1091
- Isono F, Fujita-Sato S, Ito S. 2014. Inhibiting ROR γ t/Th17 axis for autoimmune disorders. *Drug Discov Today* **19**: 1205–1211. doi:10.1016/j.drudis.2014.04.012
- Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpää MJ, et al. 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **20**: 861–873. doi:10.1101/gr.100552.109
- Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327–339. doi:10.1016/j.cell.2012.12.009
- Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27–30. doi:10.1093/nar/28.1.27
- Karwacz K, Miraldi ER, Pokrovskii M, Madi A, Yosef N, Wortman I, Chen X, Watters A, Carriero N, Awasthi A, et al. 2017. Critical role of IRF1 and BATF in forming chromatin landscape during type 1 regulatory cell differentiation. *Nat Immunol* **18**: 412–421. doi:10.1038/ni.3683
- Khader SA, Gaffen SL, Kolls JK. 2009. Th17 cells at the crossroads of innate and adaptive immunity against infectious diseases at the mucosa. *Mucosal Immunol* **2**: 403–411. doi:10.1038/mi.2009.100
- Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**: 2976–2987. doi:10.1093/nar/gkt1249
- Kittler J, Hater M, Duin RPW. 1996. Combining classifiers. *Proc Int Conf Pattern Recognit* **2**: 897–901. doi:10.1109/ICPR.1996.547205
- Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* **11**: e1004226. doi:10.1371/journal.pcbi.1004226
- Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The human transcription factors. *Cell* **172**: 650–665. doi:10.1016/j.cell.2018.01.029
- Lamparter D, Marbach D, Rueedi R, Bergmann S, Kutalik Z. 2017. Genome-wide association between transcription factor expression and chromatin accessibility reveals regulators of chromatin accessibility. *PLoS Comput Biol* **13**: e1005311. doi:10.1371/journal.pcbi.1005311
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357. doi:10.1038/nmeth.1923
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**: 799–804. doi:10.1126/science.1075090
- Lee SY, Lee SH, Yang EJ, Kim EK, Kim JK, Shin DY, Cho ML. 2015. Metformin ameliorates inflammatory bowel disease by suppression of the stat3

- signaling pathway and regulation of the between Th17/Treg balance. *PLoS One* **10**: e0135858. doi:10.1371/journal.pone.0135858
- Leng RX, Pan HF, Chen GM, Feng CC, Fan YG, Ye DQ, Li XP. 2011. The dual nature of Ets-1: focus to the pathogenesis of systemic lupus erythematosus. *Autoimmun Rev* **10**: 439–443. doi:10.1016/j.autrev.2011.01.007
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. doi:10.1093/bioinformatics/btp352
- Li P, Spolski R, Liao W, Leonard WJ. 2014. Complex interactions of transcription factors in mediating cytokine biology in T cells. *Immunol Rev* **261**: 141–156. doi:10.1111/immr.12199
- Liao JC, Boscolo R, Yang Y-L, Tran LM, Sabatti C, Roychowdhury VP. 2003. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci* **100**: 15522–15527. doi:10.1073/pnas.2136632100
- Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**: 1739–1740. doi:10.1093/bioinformatics/btr260
- Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293. doi:10.1126/science.1181369
- Lindsley AW, Qian Y, Valencia CA, Shah K, Zhang K, Assa'ad A. 2014. Combined immune deficiency in a patient with a novel *NFKB2* mutation. *J Clin Immunol* **34**: 910–915. doi:10.1007/s10875-014-0095-3
- Littman DR, Rudensky AY. 2010. Th17 and regulatory T cells in mediating and restraining inflammation. *Cell* **140**: 845–858. doi:10.1016/j.cell.2010.02.021
- Liu H, Roeder K, Wasserman L. 2010. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in neural information processing systems 23 (NIPS 2010)* (ed. Lafferty JD, et al.), pp. 1432–1440, Vancouver.
- Liu Y, Hanson S, Gurugama P, Jones A, Clark B, Ibrahim MAA. 2014. Novel *NFKB2* mutation in early-onset CVID. *J Clin Immunol* **34**: 686–690. doi:10.1007/s10875-014-0064-x
- Lopez-Herrera G, Tampella G, Pan-Hammarström Q, Herholz P, Trujillo-Vargas CM, Phadwal K, Simon AK, Moutschen M, Etzioni A, Mory A, et al. 2012. Deleterious mutations in *LRBA* are associated with a syndrome of immune deficiency and autoimmunity. *Am J Hum Genet* **90**: 986–1001. doi:10.1016/j.ajhg.2012.04.015
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550. doi:10.1186/s13059-014-0550-8
- MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J. 2016. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**: D896–D901. doi:10.1093/nar/gkw1133
- Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. 2010. DREAM3: network inference using dynamic context likelihood of relatedness and the Inferelator. *PLoS One* **5**: e9803. doi:10.1371/journal.pone.0009803
- Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Aderhold A, Bonneau R, Chen Y. 2012. Wisdom of crowds for robust gene network inference. *Nat Methods* **9**: 796. doi:10.1038/nmeth.2016
- McCarthy MT, O'Callaghan CA. 2014. PeakKDEck: a kernel density estimator-based peak calling program for DNase-seq data. *Bioinformatics* **30**: 1302–1304. doi:10.1093/bioinformatics/btt774
- Meinshausen N, Bühlmann P. 2010. Stability selection. *J R Stat Soc Ser B Stat Methodol* **72**: 417–473. doi:10.1111/j.1467-9868.2010.00740.x
- Moisan J, Grenningloh R, Bettelli E, Outka M, Ho I-C. 2007. Ets-1 is a negative regulator of Th17 differentiation. *J Exp Med* **204**: 2825–2835. doi:10.1084/jem.20070994
- Müller CL, Bonneau R, Kurtz Z. 2016. Generalized stability approach for regularized graphical models. arXiv:1605.07072 [stat.ME].
- Najafabadi HS, Mnaimeh S, Schmitz FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM. 2015. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* **33**: 555–562. doi:10.1038/nbt.3128
- Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatiyannopoulos JA. 2012. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**: 1274–1286. doi:10.1016/j.cell.2012.04.040
- Nguyen PM, Putoczki TL, Ernst M. 2015. STAT3-Activating Cytokines: a therapeutic opportunity for inflammatory bowel disease? *J Interf Cytokine Res* **35**: 340–350. doi:10.1089/jir.2014.0225
- Ou J, Liu H, Yu J, Kelliher MA, Castilla LH, Lawson ND, Zhu LJ. 2018. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**: 169. doi:10.1186/s12864-018-4559-3
- Ouyang Z, Zhou Q, Wong WH. 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci* **106**: 21521–21526. doi:10.1073/pnas.0904863106
- Patel DD, Kuchroo VK. 2015. Th17 cell pathway in human immunity: lessons from genetics and therapeutic interventions. *Immunity* **43**: 1040–1051. doi:10.1016/j.immuni.2015.12.003
- Pico AR, Kelder T, Van Iersel MP, Hanspers K, Conklin BR, Evelo C. 2008. WikiPathways: pathway editing for the people. *PLoS Biol* **6**: e184. doi:10.1371/journal.pbio.0060184
- Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21**: 447–455. doi:10.1101/gr.112623.110
- Pokrovskii M, Hall JA, Ochayon DE, Yi R, Chaimowitz NS, Seelamneni H, Carriero N, Watters A, Waggoner SN, Littman DR, et al. 2018. Transcriptional regulatory networks that promote and restrict identities and functions of intestinal innate lymphoid cells. bioRxiv doi:10.1101/465435.
- Qin J, Hu Y, Xu F, Yalamanchili HK, Wang J. 2014. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* **67**: 294–303. doi:10.1016/j.ymeth.2014.03.006
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. doi:10.1093/bioinformatics/btq033
- Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A. 2017. Dynamic gene regulatory networks of human myeloid differentiation. *Cell Syst* **4**: 416–429.e3. doi:10.1016/j.cels.2017.03.005
- Rendeiro AF, Schmidl C, Strefford JC, Walewska R, Davis Z, Farlik M, Osier D, Bock C. 2016. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun* **7**: 11938. doi:10.1038/ncomms11938
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**: 651–657. doi:10.1038/nmeth1068
- Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, Van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32**: 171–178. doi:10.1038/nbt.2798
- Siahpirani AF, Roy S. 2016. A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res* **45**: e21. doi:10.1093/nar/gkw963
- Smith CK, Kaplan MJ. 2015. The role of neutrophils in the pathogenesis of systemic lupus erythematosus. *Curr Opin Rheumatol* **27**: 448–453. doi:10.1097/BOR.0000000000000197
- Stadhouders R, Lubbers E, Hendriks RW. 2018. A cellular and molecular view of T helper 17 cell plasticity in autoimmunity. *J Autoimmun* **87**: 1–15. doi:10.1016/j.jaut.2017.12.007
- Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelting-Sun S, Lee K, et al. 2014. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**: 365–370. doi:10.1038/nature13972
- Studham ME, Tjärnberg A, Nordling TEM, Nelander S, Sonnhammer ELL. 2014. Functional association networks as priors for gene regulatory network inference. *Bioinformatics* **30**: 130–138. doi:10.1093/bioinformatics/btu285
- Tchourine K, Vogel C, Bonneau R. 2018. Condition-specific modeling of biophysical parameters advances inference of regulatory networks. *Cell Rep* **23**: 376–388. doi:10.1016/j.celrep.2018.03.048
- Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431–1443. doi:10.1016/j.cell.2014.08.009
- Wilkins O, Hafemeister C, Plessis A, Holloway-Phillips M-M, Pham GM, Nicotra AB, Gregorio GB, Jagadish SVK, Septiningsih EM, Bonneau R, et al. 2016. EGRINs (environmental gene regulatory influence networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *Plant Cell* **28**: 2365–2384. doi:10.1105/tpc.16.00158
- Wingender E. 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform* **9**: 326–332. doi:10.1093/bib/bbn016
- Wingender E, Schoeps T, Haubrock M, Dönitz J. 2014. TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res* **43**: D97–D102. doi:10.1093/nar/gku1064
- Xavier RJ, Podolsky DK. 2007. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**: 427–434. doi:10.1038/nature06005

- Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RDG, Chenoweth JG, Tesar PJ, Furey TS. 2007. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* **3**: e136. doi:10.1371/journal.pgen.0030136
- Yang J, Sundrud MS, Skepner J, Yamagata T. 2014. Targeting Th17 cells in autoimmune diseases. *Trends Pharmacol Sci* **35**: 493–500. doi:10.1016/j.tips.2014.07.006
- Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, Wu C, Karwacz K, Xiao S, Jorgolli M, et al. 2013. Dynamic regulatory network controlling T_H17 cell differentiation. *Nature* **496**: 461–468. doi:10.1038/nature11981
- Zhang J, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, Goh Y, Fullwood MJ, Sung W-K, Ruan X. 2012. ChIA-PET analysis of transcriptional chromatin interactions. *Methods* **58**: 289–299. doi:10.1016/j.ymeth.2012.08.009
- Zhang HM, Liu T, Liu CJ, Song S, Zhang X, Liu W, Jia H, Xue Y, Guo AY. 2015. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* **43**: D76–D81. doi:10.1093/nar/gku887

Received April 8, 2018; accepted in revised form January 15, 2019.