GRaNIE and GRaNPA: Inference and evaluation of enhancer-mediated gene regulatory networks applied to study macrophages

Authors

Aryan Kamal*[1,2], Christian Arnold*[1], Annique Claringbould[1], Rim Moussa[1], Neha Daga[1], Daria Nogina[1], Maksim Kholmatov[1], Nila Servaas[1], Sophia Mueller-Dott[1], Armando Reyes-Palomares[1,3], Giovanni Palla[1,4], Olga Sigalova[1,2], Daria Bunina[1], Caroline Pabst[5,6], Judith B. Zaugg[1,6]

[1]European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany.
[2]Faculty of Biosciences, Collaboration for Joint PhD Degree between EMBL and Heidelberg University, Heidelberg, Germany
[3]Current address: Department of Biochemistry and Molecular Biology, Complutense University of Madrid, 28040, Madrid, Spain
[4]Current address: Helmholtz Center Munich
[5]Department of Medicine V, Hematology, Oncology and Rheumatology, University Hospital Heidelberg, Heidelberg, Germany.
[6]Molecular Medicine Partnership Unit, University of Heidelberg, Heidelberg, Germany.

Abstract

Among the biggest challenges in the post-GWAS (genome-wide association studies) era is the interpretation of disease-associated genetic variants in non-coding genomic regions. Enhancers have emerged as key players in mediating the effect of genetic variants on complex traits and diseases. Their activity is regulated by a combination of transcription factors (TFs), epigenetic changes and genetic variants. Several approaches exist to link enhancers to their target genes, and others that infer TF-gene connections. However, we currently lack a framework that systematically integrates enhancers into TF-gene regulatory networks. Furthermore, we lack an unbiased way of assessing whether inferred regulatory interactions are biologically meaningful. Here we present two methods, implemented as user-friendly R-packages, for building and evaluating enhancer-mediated gene regulatory networks (eGRNs) called GRaNIE (Gene Regulatory Network Inference including Enhancers - https://git.embl.de/grp-zaugg/GRaNIE) and GRaNPA (Gene Regulatory Network Performance Analysis - https://git.embl.de/grp-zaugg/GRaNPA), respectively. GRaNIE jointly infers TF-enhancer, enhancer-gene and TF-gene interactions by integrating open chromatin data such as ATAC-Seq or H3K27ac with RNA-seq across a set of samples (e.g. individuals), and optionally also Hi-C data. GRaNPA is a general framework for evaluating the biological relevance of TF-gene GRNs by assessing their performance for predicting cell-type specific differential expression. We demonstrate the power of our tool-suite by investigating gene regulatory mechanisms in macrophages that underlie their response to infection, and their involvement in common genetic diseases including autoimmune diseases.

Introduction

Enhancers are genomic locations that play an important role in cell-type specific gene regulation, and an increasing number of diseases are linked to impaired enhancer function [1, 2]. This is particularly obvious for common genetic traits and diseases for which genome-wide association studies (GWAS) have linked over 200,000 genetic variants with over 40,000 traits and diseases, yet the vast majority of the disease-associated genetic variants lie in non-coding regions far from promoters [1], thus likely affecting enhancers and having a regulatory role.

Among the biggest challenges in the post-GWAS era is the interpretation of these disease-associated genetic variants in non-coding genomic regions. For most disease-associated genetic variants in regulatory regions it is still unclear what genes they target. This is further complicated by the cell- and condition-specific activity of gene regulatory elements, which is likely conferred by transcription factors (TFs) that regulate the specific elements. A recent study points at the importance of studying TFs for understanding genetic variants associated with auto-immune diseases [3], which is corroborated by the fact that trans-eQTLs, which are likely to act via TFs, tend to be more enriched in disease-associated variants than cis-eQTLs [4]. On the other hand, it is crucial to include putative enhancers for predicting the function of TFs in cell-fate determination [5, 6]. Enhancer-mediated gene regulatory networks reconstructed from single cell RNA and ATAC-seq profiling in the fly brain have led to a better understanding of the regulatory diversity across the different neuronal cell types [6]. Our work has shown the power of using enhancer-based gene regulatory networks to understand disease mechanisms in pulmonary arterial hypertension [7]. Thus, for interpreting disease-associated genetic variants, or enhancers in general, it will be crucial to jointly investigate TF activity, enhancers and gene expression in a cell-type specific manner.

Several approaches have been proposed to infer bi-partite TF-gene networks for example based on co-expression in bulk [8, 9], or single cell data [10–12], based on partial information decomposition [13], time-course data [14], or data curation [15–18]. At the same time, methods for inferring enhancer-gene links have been proposed e.g. using co-variation of peaks [19–21], or targeted perturbations of enhancers followed by sequencing [19, 20]. However, we currently lack tools to jointly infer TF-enhancer and enhancer-gene links, both of which are needed for inferring the context-specific interaction between TFs and regulatory elements.

An important step in the regulatory network reconstruction is to evaluate their biological significance. Common approaches for assessing regulatory interactions include benchmarking them against networks from simulated data or against known biological networks [22–24]. The drawback of simulated networks is that they are based on many assumptions of the structure of a “true” biological network. Benchmarking against true biological networks typically suffer from a strong literature bias [25], low complexity of these “true” circuits and a limited range of connections and cell types. In general, each network inference method will have its own bias and shortcomings, and performance will depend on

the benchmarking dataset [22–24]. Thus, there is a strong need for an unbiased approach to assess the biological relevance of inferred regulatory interactions.

Here we present a tool-suite for building and evaluating enhancer-inclusive gene regulatory networks (eGRNs) called GRaNIE (Gene Regulatory Network Inference including Enhancers - https://git.embl.de/grp-zaugg/GRaNIE) and GRaNPA (Gene Regulatory Network Performance Analysis - https://git.embl.de/grp-zaugg/GRaNPA), respectively. GRaNIE jointly infers TF-enhancer and enhancer-gene interactions based on bulk RNA-seq expression and chromatin accessibility (ATAC-seq) or ChIP-seq on active histone marks (e.g. H3K27ac) data. GRaNPA assesses the biological relevance (of any GRN) using a machine learning framework to predict cell-type specific differential expression. The eGRNs are based on integrating expression and chromatin accessibility/activity patterns across a cohort of individuals. We demonstrate the power of eGRNs in macrophages. Using GRaNIE followed by GRaNPA, we gain biological insights into the TFs that drive the expression programs in response to infection with different pathogens and regulatory elements that overlap with common disease variants in autoimmune diseases. Finally, we provide a comprehensive resource of cell-type specific gene regulatory networks for three other cell types (https://apps.embl.de/grn/).

Results

Overview and conceptual description of the GRaNIE algorithm

To interpret genetic and epigenetic variation in regulatory (enhancer and promoter) regions in a context-specific manner, we devised GRaNIE, an approach for jointly inferring TF-enhancer/promoter and enhancer/promoter-gene interactions from the same data, that we have applied in a recent study to investigate the disease mechanisms of pulmonary arterial hypertension [7]. In the following, we refer to regulatory regions (enhancers and promoters) as peaks (as in, signal obtained from ATAC-seq or ChIP-seq data). Briefly, GRaNIE first identifies TF-peak and peak-gene links, and then integrates them into an eGRN.

The TF-peak links are based on statistically significant co-variation of TF expression and peak accessibility across samples (e.g. individuals, minimum number ~10-15), taking into account predicted TF binding sites. To obtain them, GRaNIE calculates all pairwise correlations between TF expression levels (RNA-seq) and peak signal (ATAC/ChIP-seq), stratified by whether or not the peak contains a predicted binding site of the TF. For each TF it then uses the distribution of all peaks that do not contain its predicted binding site as background to calculate an empirical false discovery rate (FDR) for assessing the significance of TF-enhancer links (**Fig. S1**). In our previous work we have demonstrated that negative TF-peak correlations are indicative of a TF to act as transcriptional repressor while positive correlations are indicative of activators [26]. As a quality control (QC) in GRaNIE, we recommend to compare the number of TF-peak links to what is expected by chance by running GRaNIE in permutation mode (permutation of sample labels).

Similarly, peak-gene links are based on significant co-variation of enhancer accessibility and gene expression across the same samples (**Fig. 1A, Fig. S1**). To obtain them, GRaNIE calculates the correlation between the expression of a gene and peak signal of all peaks within a defined distance of its transcription start site (TSS, default is 250kb). GRaNIE also allows the use of chromatin conformation data, such as Hi-C, to define which peak-gene pairs will be tested (**Fig. S1**). Since chromatin accessibility in active regulatory elements, such as enhancers and promoters, is associated with active gene regulation and transcription we only expect positive correlations for functional peak-gene links. We also expect a positive correlation between accessibility and expression for silencer elements based on the fact that they are mostly found in repressive chromatin regions .[27]. Negative correlations have no clear biological meaning and therefore may indicate remaining batch effects or random noise. We therefore implemented the assessment of positive versus negative peak-gene correlations as a quality control (QC) metric in GRaNIE to judge the signal-to-noise ratio in the data, and by default only retain positive enhancer-gene links.

GRaNIE then combines these two types of edges to create a tripartite TF-enhancer-gene eGRN, based on three user-defined thresholds: false discovery rate (FDR) of TF-enhancer edges, FDR of enhancer-gene edges, and maximum distance between enhancer and transcription start site (TSS). The latter can optionally be replaced by chromatin domain data, such as Hi-C data, to test all enhancer-gene pairs within a 3D proximity.

The whole framework is implemented in a user-friendly R-package that can be installed from git (https://git.embl.de/grp-zaugg/GRaNIE). In addition to the eGRN reconstruction function, GRaNIE comprises a set of easy-to-use functions for generating and visualising network statistics, identifying network communities, performing subnetwork-specific Gene Ontology (GO) enrichment analysis, and quality control (QC) plots. An extensive documentation is available at (https://grp-zaugg.embl-community.io/GRaNIE).

Application of GRaNIE for generating cell-type specific eGRN in macrophages

Macrophages are large white blood cells of the innate immune system that can be found in essentially all tissues and play a role in inflammatory disorders. Genetic variants associated with several auto-immune and other diseases are enriched for enhancers active in macrophages [27, 28]. Given that inflammatory conditions are underlying many common diseases, macrophages present an important cell type within which disease-associated variants manifest their effect. Thus, macrophages present an ideal cell type for applying and testing the eGRN framework.

We obtained joint RNA- and ATAC-seq data for (induced pluripotent stem cell) iPSC-derived macrophages from 31-45 individuals in four conditions (naïve, interferon gamma (IFN-g)-primed, infected (with Salmonella), and IFN-g-primed-infected) [27](**Fig. 1B**). For each of these, we applied GRaNIE, using default parameters (see below; 0.2 for TF-peak FDR, 0.1 for peak-gene FDR, and 250kb for peak-gene distance) to reconstruct an eGRN using TF binding sites predictions based on HOCOMOCO v11 and PWMScan as described in [26].

For all eGRNs we observed much fewer significant connections when running GRaNIE in permutation mode (**Fig. 1C and Fig. S2A-C**), signifying that they pass the QC for TF-peak links. Similarly, for the peak-gene links, we find that all eGRNs show more signal for positive than for negative correlations, and that the signal-to-noise ratio decreases with peak-gene distance until no signal is left for random peak-gene pairs (**Fig. 1D, Fig. S3A-C**), thus passing the peak-gene QC.

The resulting eGRNs comprised 78-126 TFs, 1802-6742 enhancers and 1793-3869 genes (**Fig. 1E, Fig. S4A-C; Table 1**). We observed a slightly larger number of TFs classified as activators than as repressors (1.5-2 fold), yet activators tend to be connected with more enhancers resulting in over 10-fold more enhancers being linked to an activator than a repressor (**Fig. 1C, Fig. S2A-C; Table 1**). Notably, in all networks, only about 20-30% of the enhancers are linked to their closest gene TSS (**Fig. 1F**), an observation that is consistent with previous observations in pulmonary arterial endothelial cells [7] and iPSC derived cardiomyocytes [29]. Stratifying the enhancer-gene links based on their distance to the gene TSS revealed a steep drop in number of connections for distances greater than 10kb (likely reflecting extended promoter regions) with a slower decline in numbers of significant connections for distances greater than 10kb (**Fig. 1G**). This trend is particularly strong for highly significant enhancer-gene links (p-value<0.05) and is in stark contrast to chromatin capture data such as from Hi-C, where typically a much stronger decay with distance is observed [30], thus indicating that our approach is able to capture functional interactions independent of distance.

Most genes are linked to one or two peaks (**Fig. 1H**). The majority of TFs are connected to very few genes, yet a handful of TFs are connected to over 50 genes, as exemplified for the eGRN from infected macrophages (**Fig. 1I**), which is in line with the typically scale-free structure of gene-regulatory networks [31]. The most connected TFs in the infected eGRN include IRF8 known for its role in macrophage biology [32], as well as TFs that have not extensively been linked to macrophages, such as NRF1.

To dive into the biological processes captured by eGRNs, we implemented functionalities for identifying subnetworks, or communities, and perform GO term enrichment on them in GRaNIE. For identifying subnetworks, or communities, GRaNIE by default uses Louvain clustering, implemented in the igraph package in R [33]. Applied to the infected macrophage eGRN, this revealed 64 communities, of which seven had more than 60 genes (C1-C7), while the majority (44) had only 2-3 members (**Fig S5A-C**). The largest community (C1) is dominated by genes targeted by IRF8 and enriched in processes related to differentiation of immune cells, host defence, defence response to virus, and TNF-mediated signalling (**Fig. 1J**). The second largest community (C2) is dominated by the TFs MBD2, NRF1 and ETV1, and enriched in functions related to phagocytosis and interferon-gamma response. The MECP2-dominated module C3 is enriched in processes related to chemotaxis and cytokine production. The KLF6-driven module C4 is enriched in viral defence response and MAPK signalling. The smaller communities C5 and C6 are regulated by TEAD1,2,4/GATA4 and TYY1/KAISO respectively and enriched in BMP signalling, which is relevant for inflammation phenotypes. Overall, we found that each community is enriched for a set of distinct biological processes.

In summary, we applied GRaNIE to generate four eGRNs from naive, IFN-primed, Salmonella-infected, and IFN-primed-infected macrophages. GO enrichment and network community functions implemented within GRaNIE identified the expected macrophage-biology related processes (including immune response, response to virus and chemotaxis), and revealed specific network communities, each defined by a small set of TFs, and enriched for specific macrophage-related biological processes.

Conceptual description of GRaNPA, an approach for evaluating the biological relevance of GRNs

To evaluate the biological relevance of the GRNs we propose a test based on the premise that cell-type specific GRNs should capture gene expression regulation in a cell type specific manner. We devised a machine learning approach, GRaNPA (Gene Regulatory Network Performance Analysis), that evaluates how well the bipartite TF-gene connections of an eGRN can predict cell-type specific differential gene expression.

GRaNPA requires differential RNA expression data for the cell type for which the eGRN was constructed, that is independent from the data used to generate the eGRN. It then trains a random forest regression model to predict a differential expression value per gene, based on the TF-target gene connections from the GRN in a 10-fold cross validation setting (see methods; **Fig. 2A**). To assess the specificity of the network, it also trains a separate model based on randomized edges of similar structure and compares the obtained predictions (R^2). A performance (R^2) greater than 0.05 for the random networks indicates that even unspecific TF-gene connections can predict differential expression. To assess overfitting, GRaNPA trains the same random network on completely random differential expression data (uniform distribution between -10 to 10; see methods). A performance greater than 0.05 for a random network to predict random values indicates that the model is overfitting and can predict anything. Notably, GRaNPA can be applied to assess any GRN that contains TF-gene connections and may be used to benchmark GRNs constructed using various methods (see below).

In short, our strategy is based on the following steps:

1. Obtain differential expression data for cell types matching the GRNs
2. For each cell type, train a random forest regression model (10-fold cross validation) to predict a differential expression value per gene ($diff.\ exp.\ (gene_i)$) based on TF-gene links from the eGRN $edge(gene_i, tf_j)$, by learning a coefficient (β_j) for each TF (tf_j) (see eq. 1).

$$diff.\ exp.\ (gene_i)\ \sim \beta_1 * edge(gene_i,\ tf_1) \ldots\ \beta_n * edge(gene_i,\ tf_n) \quad \text{eq.1}$$

Where:

$$edge(gene_i,\ tf_1)\ =\ edge\ between\ gene_i\ and\ TF_j = \{0\ if\ no\ edge;\ != 0\ otherwise\}$$

3. Compare the performance of models learned on real TF-gene links, random TF-gene links, and TF-gene links from other cell types (cross-validation R^2).

In addition to the predictive evaluation of GRNs, GRaNPA also estimates the importance of each TF towards the prediction of the differential expression data, which is directly stored in the GRaNPA object. The calculation of TF importance is based on the in-built *importance* function in the package *ranger* that quantifies importance of features in random forest models based on how much their exclusion affects prediction accuracy. GRaNPA is implemented as a user-friendly R-package (at https://git.embl.de/grp-zaugg/GRaNPA) and documentation is available at (https://grp-zaugg.embl-community.io/GRaNPA/).

Biological evaluation of the macrophage eGRNs

To evaluate the macrophage eGRNs, we obtained RNA-seq data for naive, primed, naive-infected, and primed-infected from [27], and calculated the pairwise differential expression (Methods) only using the data that was not used in the network reconstruction.

We first used GRaNPA to find a set of reasonable default parameters (TF-peak FDR and peak-gene FDR) based on the naive macrophage eGRN using the differential expression between naive vs salmonella infected macrophages after 5h from [27]. We constructed a set of eGRNs varying one of these parameters while fixing the other, and keeping the maximum distance at 250kb. For the peak-gene FDR, we observe very little impact on predictive power and an almost linear increase in the number of genes present in the network with an increasingly lenient FDR threshold (**Fig. S6A top**). In contrast, the TF-peak FDR cut-off has a strong effect on the predictive power (**Fig. S6A bottom**). This evaluation revealed a set of reasonable parameters (considering performance and number of genes): peak-gene FDR=0.1, TF-peak FDR=0.2 when using a maximal distance to TSS of 250kb. Retesting the same parameters in the infected and primed macrophage eGRNs confirmed that this is a reasonable set of parameters (**Fig. S6B-C**).

Using these parameters we evaluated all four macrophage eGRNs and found that three of them perform well in terms of R^2 (>0.15) as well as classification of direction of differential expression values, as measured by area under the receiver operating characteristic curve (AUROC), while the performance of the corresponding randomised networks was <0.05 R^2 for two of them (**Fig. 2B-C** and **Fig. S8A-C** and **Fig. S9A-D**). The eGRN for primed-infected macrophages was unable to predict any differential expression and was therefore excluded from further analyses. The significant difference between the randomised and the actual networks shows that the eGRNs indeed capture biologically relevant links between TFs and genes. Notably, when comparing the actual with the predicted fold-change for large differential expression effects (absolute log2 fold-changes > 1), the eGRNs are able to predict the direction of effect with 0.71 AUROC (**Fig. S9A-D**).

eGRNs built from single cell-types show cell-type specific predictions

We next assessed the cell-type specificity of the predictions in eGRNs. To this end, we obtained datasets in different cell types with matched RNA and chromatin accessibility data for naive CD4+ T-cells [34] and from acute myeloid leukaemia (AML) [35] and He et al *in revision*. Running GRaNIE using the same parameters as described above, we obtained eGRNs for T-cells and for AML.

To assess their cell-type specific prediction power, we run GRaNPA on the infected macrophage, the T-cell, and the AML eGRNs, comparing their performance to predict differential expression between naive T-cells in a resting condition vs lipopolysaccharide (LPS) stimulation (data from [36]), between two subtypes of AML: GPR56-high vs GPR56-low (data from [35]), and between naive vs salmonella-infected macrophages [27]. We found that the eGRN that matches the respective cell type led to the best prediction (**Fig. 2D;** R^2 0.08 for T-cells, 0.27 for AML, and 0.14 for macrophages). While T-cells and macrophages were only predictive in their own cell type, we observed that the AML eGRN was to a smaller extent also predictive for the macrophage response (R^2~0.1). Since AML cells and macrophages are both from the myeloid lineage this could indicate some shared regulatory architecture between these cells.

We hypothesised that the cell-type specific predictions may be caused by the fact that distinct TFs are driving gene expression patterns in distinct cell types. To test this, we made use of the TF-importance estimate implemented in GRaNPA. We found that the most important TFs are mostly unique for each cell type (**Fig. 2E**). One exception to this is IRF8, which was important for predicting both the differential expression between AML subtypes using the AML eGRN, and the macrophage expression response to infection using the macrophage eGRN (**Fig. 2E**). However, we found that the IRF8 regulons (defined as all genes connected to a TF) in AML and macrophages had only 22 genes (and no single enhancer) in common, while each included hundreds of non-overlapping genes (**Fig. 2F**), suggesting a highly cell-type specific regulon composition of IRF8. These two IRF8 regulons are enriched for very distinct GO terms: in AML IRF8 regulates genes involved in “neutrophil degranulation”, “positive regulation of blood vessel endothelial migration”, whereas in the macrophage eGRN, IRF8 is connected to genes involved in “response to virus”, “blood coagulation”, “TNF-mediated signaling” (**Fig. 2G**). These observations indicate that the cell-type specificity of the predictions is not only dependent on distinct sets of TFs driving the response, but also that the same TF can have vastly different genes that it targets depending on the cell type.

The eGRNs connect TFs to genes through active regulatory regions, comprising both promoters and enhancers. The predictive evaluation setup allowed us to compare the relative importance of promoter (i.e. <10kb from TSS) and enhancer links (> 10kb from TSS) in different eGRNs. To do so, we divided the gene-enhancer pairs into 10 groups based on their distance to TSS and ran GRaNPA for each group separately. The promoter-only eGRNs from infected and primed macrophages showed limited or no predictive power (**Fig. 2H-I and Fig. S10**). This highlights the importance of enhancers and is in line with a recent study that demonstrated the importance of considering enhancers for predicting the cell-fate potential of TFs [5]. In contrast, for the eGRN from naive macrophages we found a similar prediction performance for the enhancer-only and promoter-only networks, suggesting that for some conditions promoters may carry most information.

Application of GRaNPA to compare eGRNs with other GRN methods

Notably, GRaNPA is applicable to assess any type of bipartite TF-gene network and can be used more generally to assess the utility of a GRN for understanding cellular response to

specific perturbations. Here we used it to evaluate the performance of several previously published GRNs that draw links between TFs and genes based on different approaches: DoRothEA, which uses a data-driven approach including co-expression to draw TF-gene links [15, 37, 38], ChEA3, which uses ChIP-seq experiments from ENCODE to draw TF-gene links [17], and TRRUST, which is a curation of TF-gene links published in pubmed indexed articles [16]. Our eGRNs and several versions of DoRothEA (A-E reflect the confidence of the TF-gene links) are predictive for at least one of the differential expression data, while ChEA3 and TRRUST were not predictive in any setting (**Fig. 3A**). We also tried different versions of ChEA3 as well as Regnetwork [18], again none of which was predictive. The predictiveness of GRNs was not affected by their size (**Fig. 3A**). However, DoRothEA is built based on expression data across many different conditions and thus by definition is not cell-type specific. Thus, the GRaNIE approach is the only GRN tool that builds predictive, cell-type specific GRNs (for macrophages, AML and T-cells; **Fig. 3B**). Furthermore, it is the only tool that generates predictive GRNs and that includes enhancers.

Macrophage eGRNs reveal distinct set of TFs driving response to different types of infection

We next highlight the biological insights that GRaNIE and GRaNPA can provide. Specifically, we employed them for studying the response of macrophages to different types of bacterial infections. We obtained data from previously published studies that measured the expression response of macrophages infected with *Mycobacterium Tuberculosis* (MTB) [39], *Listeria monocytogenes* [40], *Salmonella Typhimurium* [27, 40]), and stimulation with IFN-g [27]. The studies are quite diverse in their set-up (**Table 2**).

To understand how macrophages respond to these distinct perturbations, we employed GRaNPA on the three eGRNs that passed our assessment above (naive, infected, primed) and evaluate them in each infection condition. Overall, the naive and the naive-infected eGRNs were able to predict the differential expression response to most pathogens (R^2>0.075 and R^2<0.05 for the respective random networks), while the primed network failed the specificity control for some conditions (**Fig. 4A** and **Fig. S11A**). Since naive and infected eGRNs both showed high predictive power in these settings, we decided to assess the pathogen response using the union of these two eGRNs (**Fig. 4B**).

One of the well-understood responses of macrophages upon infection is the activation of the NFKB family of inducible transcription factors [41]. In line with this, we find the NFKB1/NFKB2 regulons in the union macrophage eGRN strongly up-regulated across all infections and upon IFN-g stimulation (**Fig. 4C**).

To further investigate regulatory mechanisms underlying the infection responses, we used GRaNPA to determine the importance for each TF for predicting a specific differential expression response. Comparing TF importance across two independent datasets from Salmonella infected macrophages revealed very similar profiles despite differences in cell types (**Fig. 4D**; iPSC-derived vs monocyte-derived macrophages) and protocols (5h and 2h post infection), thus highlighting the robustness of the GRaNPA approach as well as the biological congruence between the experiments.

In contrast, when comparing across infection settings, TF-importance profiles were quite variable (**Fig. 4D**). GO enrichment analysis of the regulons of the top 5 most important TFs across all conditions (as implemented in GRaNIE) in the naive-union eGRN revealed that each TF is associated with very specific terms, likely reflecting different defence mechanisms triggered by the pathogens (**Fig. 4E**). Macrophages use several strategies to fight infections, including phagocytosis followed by several degradation mechanisms, starvation of pathogens, as well as recruiting other players in the immune system [42]. For example, we find the NFKB1/2-regulon strongly enriched for genes involved in chemokine signaling and taxis and IRF8 for response to virus and bacteria (**Fig. 4E**). Among TFs that are less well known for their role in macrophages we find PURA, whose regulon is enriched for IFNg response and chemotaxis, and the cell cycle TF E2F7, which was mostly important for the response to Tuberculosis 18 and 36h post infection.

The most important TF for predicting the response to Salmonella infection was IRF8, followed by MBD2 (**Fig. 4D**). IRF8 is a known pro-inflammatory interferon response factor, associated with the polarization of macrophages into the M1, pro-inflammatory type [43], which we confirmed in our data using gene-set enrichment analysis of the IRF8 regulon (**Fig. S12**). Less is known about MBD2 in macrophages, although it has been linked to intestinal inflammation in mice [44] and MBD2 deficiency has been associated with repressing the M2 macrophage programme in pulmonary fibrosis [45]. When we assessed the expression change upon infection with Salmonella for the genes in the IRF8 and MBD2 regulons, we observed that the IRF8 regulon is generally up-regulated upon infection, whereas the MBD2 regulon is mostly down-regulated (**Fig. 4F**). Notably, the stimulation with IFN-g only activated the pro-inflammatory IRF8 programme, whereas the MBD2 regulon was not affected (**Fig. 4F**). We observed similar expression dynamics of the IRF8 and MBD2 regulons in Listeria (**Fig. S13**) even if IRF8 was a less important TF for predicting the response (**Fig. 4D**).

ZBT14 and MBD2 are not as known for their role in macrophage biology, yet GRaNPA finds them important for predicting macrophage response to all of the studied pathogens. In our data, the MBD2 and ZBT14 regulons are significantly overlapping (**Fig. 4F**, p=3.3e-13, hypergeometric test) and genes jointly regulated by them are significantly enriched in response to metal ions (**Fig. 4G**). The use of zinc and copper ions in macrophage defence strategies is well documented [46, 47] and given that the ZBT14 and MBD2 are only important for predicting response to pathogens, and not for predicting response to IFN-g stimulation, we speculate that ZBT14 and MBD2 may jointly induce a macrophage-intrinsic mechanism to counteract toxic metal ions, potentially aimed at overcoming the toxic effects of its own weapons.

Overall, these results highlight the use of cell-type specific eGRNs for investigating the specific biological functions that are regulated by a TF in a specific cell type.

Macrophage-specific eGRNs are enriched in finemapped GWAS variants and immune-related traits

The strength of the eGRN framework is that we can specifically investigate the role of gene regulatory elements, such as enhancers, which are enriched for disease-associated genetic variants [1]. We therefore sought to explore the macrophage eGRNs to learn about gene regulatory mechanisms underlying associations of genetic variants with common complex traits and diseases.

First, we tested whether the enhancer regions specific to the three macrophage eGRNs that GRaNPA identified as predictive in at least one infection setting (naive, primed, naive-infected) were enriched in heritability for 442 GWAS traits (**STable 7**). We applied stratified linkage disequilibrium score regression (S-LDSC [48]) and compared the eGRN peaks to all peaks identified in macrophages (Methods). Notable enrichments include HbA1c measurement, a measure for diabetes severity, large artery stroke and heart failure for the naive eGRN; adolescent idiopathic scoliosis and nonischemic cardiomyopathy for the primed eGRN and rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) for the infected eGRN (**Fig. 5A**). SLE and RA are both driven by activated macrophages [49, 50] as a result of known (for SLE) or hypothesised (RA) upregulation of IFN-g signalling [51–53]. Interestingly, we find enriched heritability for these traits in the peaks for the IFN-g primed eGRN, but not for either of the naive or infected eGRNs. Given this association, we also assessed the heritability enrichment of the regulatory elements and genes connected to the TFs that are particularly important for predicting the response of macrophages to IFN-g (NFkB1/2, RELB, IRF8). Inflammatory bowel disease (specifically ulcerative colitis) comes out as the top enriched trait (**Fig. S15**), which is in line with the known role of IFN-g in this disease [54]. Literature evidence for other traits is summarised in **STable 8**.

Next, we zoomed in to a specific set of fine-mapped GWAS variants associated with autoimmune diseases with a known link to macrophages. Across all three macrophage eGRNs we found 11 fine-mapped variants that were located in the regulatory regions (2 in the naive, 2 in infected, and 8 in the primed eGRN). The infected and primed eGRNs were significantly enriched for containing fine-mapped variants (**Fig. 5B**). Investigating the TFs regulating the fine-mapped auto-immune disease enhancers, we find the known immune response TFs NFKB1/2, RELB and IRF8, but also MBD2, ZBT14 and PURA (**Fig. 5C**), which we identified as important TFs for predicting response to infection.

One of the fine-mapped SNPs in SLE is rs9893132, located in an enhancer regulated by SP1 and NFKb2 that is linked to the non-coding RNA MALAT1 and the gene cathepsin W (CTSW) in the primed macrophage eGRN. Both genes are over 50kb from the fine-mapped SNP and there are several other genes in the locus that are not linked to the enhancer in question (**Fig. 5D**). MALAT1 has been implicated in SLE through several studies [55], suggesting that the fine-mapped SNP may target MALAT1. While CTSW expression in lymphocytes has been linked to autoimmune diseases (e.g. [56]), its role in macrophages is much less studied. Yet CTSW knockdown in macrophages reportedly increased Mycobacterium Tuberculosis survival in macrophages [57], suggesting it does play an

important function in the pro-inflammatory macrophage response. Overall, the macrophage eGRNs provide the putative target genes of 11 fine-mapped GWAS loci, often linking to genes that are over 50kb away from the SNP (**Table 3**).

Discussion

Phenotypic variation across individuals has two major sources. On the one hand it is genetic variation that can lead to variation in molecular phenotypes that impact on complex traits. On the other hand, phenotypic variation can arise from variation in the environment that can be long-lived (epigenetics) or short-lived (signalling). Thus, to understand mechanisms underlying phenotypic variation (including disease phenotypes), it is crucial to study the interplay between genetic variants, epigenetic marks and extrinsic cellular signalling. Here we present GRaNIE and GRaNPA, a tool-suite that provides a framework for jointly analysing these three layers and investigating their biological relevance.

GRaNIE is a flexible and user-friendly R-package for building enhancer-based gene regulatory networks, while GRaNPA is a separate R-package for evaluating the biological relevance (i.e. predictive power) of TF-gene networks. GRaNIE requires RNA-Seq and open chromatin data such as ATAC-Seq or ChIP-Seq for histone modifications (e.g, H3K27ac) across a cohort of individuals (>15), along with TFBS data (that can either be obtained from the package or provided by the user) to generate cell-type specific eGRNs. It provides a range of quality control plots and functionalities for downstream analyses, such as identification of communities within the network, and GO enrichment analyses. A dedicated website accompanies the package and is automatically updated whenever a new package version becomes available.

GRaNPA requires a bipartite TF-gene network and gene-specific differential expression values as input, and assesses the predictive power of the network. As a downstream functionality it provides quantification of TF importance for driving a specific differential expression response. Notably, the purpose of GRaNPA predictions is to assess the TF-gene connectivities, and while including additional gene-specific features may improve prediction performance, as shown in [58], it would not help in the assessment of GRNs. An additional advantage of GRaNPA is we can directly compare different GRNs and get a quantitative assessment of their performance for a perturbation of interest, without the need for a "ground-truth" network to evaluate their edges and connectivities.

Compared to most of the available GRNs, GRaNIE is an enhancer-based regulatory network that only captures TF-gene links mediated by enhancers. This has several advantages: first, we showed that at least for macrophages, parts of their expression response to infection could only be predicted when using enhancer connections. Second, including enhancers in TF-driven GRNs, allows the investigation of mechanisms underlying GWAS traits that are driven by specific TFs, and facilitates interpretation of (finemapped) trait-associated SNPs. Third, since enhancers tend to be highly cell-type specific [59], eGRNs are likely more cell type specific than TF-gene networks. Finally, by requiring a correlation between the expression level of a TF and the accessibility of the peak, in addition to the motif presence, the GRaNIE approach circumvents the inherent limitation of TF binding site predictions,

which cannot distinguish between TFs with similar binding motifs [60]. This will exclude many TF-enhancer links that have the TF motif yet show no correlation with the TF expression level and are thus likely false positive binding sites in the given cell type.

eGRNs consist of TFs and their respective downstream enhancer and gene targets, which means we can zoom in to network communities that capture specific pathways or functions. For example, we showed that when we divide the network into communities, each community is enriched in distinct TF-driven biological processes. The modularity of the eGRN also showed that NFKB1, NFKB2, RELB and IRF8, the TFs important for predicting the macrophage response to IFNg priming, and their connected enhancers and genes were specifically enriched for heritability of GWAS traits that are commonly linked to IFNg signalling. In contrast to other approaches for interpreting trait-associated variants that are solely based on epigenetics, such as the activity by contact model [61], or purely based on genetics, such as eQTLs [4] or variable chromatin domains [62], eGRNs by GRaNIE capture TF-enhancer-gene links based on variation due to genetic, epigenetic, or TF-activity differences across individuals, thus integrating these three layers in one framework. In sum, eGRNs can be used to identify the target genes of individual TFs (regulons), to pinpoint the cell-type specific regulatory regions that connect to a TF, and to investigate genetic variation in the tripartite TF-enhancer-gene graphs.

Comparing eGRNs across cell types revealed that for some TFs (e.g. IRF8) the regulons are highly cell-type specific. One explanation for this cell-type specific regulons is that TFs may have different co-binding TF partners depending on the cell type. In our previous work we found that TFs regulate distinct biological processes depending on their co-binding partners [63, 64]. An alternative explanation for cell-type specific regulons is that different cell types may differ in their chromatin potential [65].

Among the notable observations from applying GRaNIE and GRaNPA to study the gene expression response in macrophages was that some TFs, including MBD2, were specifically important only for predicting response to bacterial infection, and not for IFN-g stimulation. Interestingly, IRF transcriptional programs are generally more related to a virus response, so it could be that MBD2 is needed in the bacterial specific response, indicating that we can potentially use these networks to identify TFs important for responses to different types of pathogens. Another observation is that three of important TFs for predicting the response to infection and but not to IFN-g are known to bind methylated DNA: MBD2 [66], MECP2 [67], NRF1 [68]. Recent reports provide evidence for a pathogen-induced global DNA methylation alteration [69] downstream of NFkB-signalling, and it was shown that MBD2 inhibits IFNy by selectively binding to methylated regions in the Stat1 promoter in other cell types [70]. Our results are consistent with a pathogen-response mechanism that is partially mediated by DNA methylation, which may modulate the impact of DNA-methylation sensitive TFs and demonstrates the level of novel biological insights that can be gained with GRaNIE and GRaNPA.

Acknowledgements

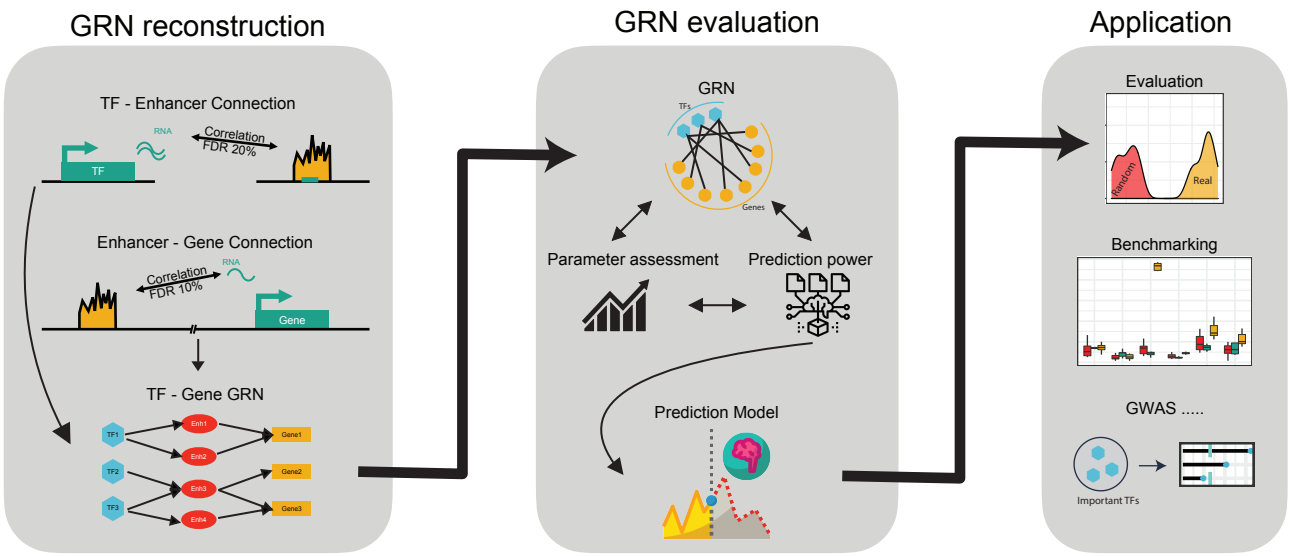
We thank Anna Mathoudaki for proofreading and for naming GRaNIE and GRaNPA, Martina Muckenthaler, Oriana Marques and Christina Mertens for discussions about macrophage biology. M.K. is funded by the European Union's H2020 MSCA-ITN Project ENHPATHY, grant agreement number 860002. A.C. is funded by the European Union's Horizon 2020 MSCA, grant agreement number 847543. ARP has been granted by the Regional Programme of Research and Technological Innovation for Young Doctors UCM-CAM (PR65/19- 22460).

Declarations

We declare that none of the authors have any competing interests.

Figures

Graphical abstract


GRN reconstruction
TF - Enhancer Connection
RNA
Correlation
FDR 20%
TF
Enhancer - Gene Connection
Correlation
FDR 10%
RNA
Gene
TF - Gene GRN
TF1
TF2
TF3
Enh1
Enh2
Enh3
Enh4
Gene1
Gene2
Gene3
GRN evaluation
GRN
TFs
Genes
Parameter assessment
Prediction power
Prediction Model
Application
Evaluation
Random
Real
Benchmarking
GWAS .....
Important TFs

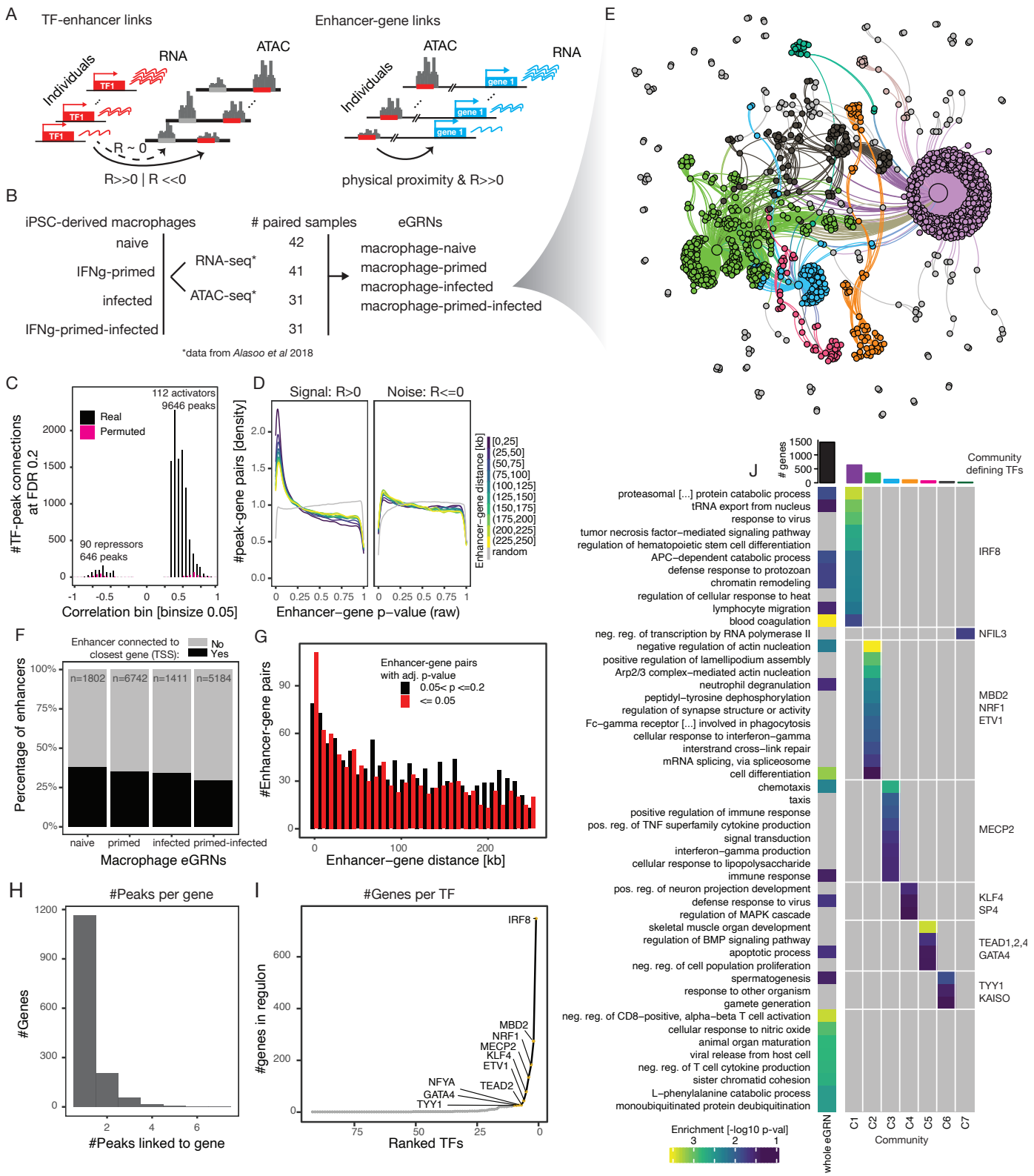

A
TF-enhancer links
Enhancer-gene links
Individuals
RNA
ATAC
TF1
R ~ 0
R>>0 | R <<0
gene 1
physical proximity & R>>0
B
iPSC-derived macrophages
paired samples
eGRNs
naive
IFNg-primed
infected
IFNg-primed-infected
RNA-seq*
ATAC-seq*
42
41
31
31
macrophage-naive
macrophage-primed
macrophage-infected
macrophage-primed-infected
*data from Alasoo et al 2018
C
112 activators
9646 peaks
Real
Permuted
90 repressors
646 peaks
#TF-peak connections at FDR 0.2
Correlation bin [binsize 0.05]
D
Signal: R>0
Noise: R<=0
#peak-gene pairs [density]
Enhancer–gene p–value (raw)
Enhancer–gene distance [kb]
[0,25]
(25,50]
(50,75]
(75,100]
(100,125]
(125,150]
(150,175]
(175,200]
(200,225]
(225,250]
random
E
F
Enhancer connected to closest gene (TSS):
No
Yes
n=1802
n=6742
n=1411
n=5184
Percentage of enhancers
naive
primed
infected
primed-infected
Macrophage eGRNs
G
Enhancer-gene pairs with adj. p-value
0.05< p <=0.2
<= 0.05
#Enhancer-gene pairs
Enhancer–gene distance [kb]
H
#Peaks per gene
#Genes
#Peaks linked to gene
I
#Genes per TF
IRF8
MBD2
NRF1
MECP2
KLF4
ETV1
NFYA
GATA4
TYY1
TEAD2
#genes in regulon
Ranked TFs
J
genes
Community defining TFs
proteasomal [...] protein catabolic process
tRNA export from nucleus
response to virus
tumor necrosis factor–mediated signaling pathway
regulation of hematopoietic stem cell differentiation
APC–dependent catabolic process
defense response to protozoan
chromatin remodeling
regulation of cellular response to heat
lymphocyte migration
blood coagulation
neg. reg. of transcription by RNA polymerase II
negative regulation of actin nucleation
positive regulation of lamellipodium assembly
Arp2/3 complex–mediated actin nucleation
neutrophil degranulation
peptidyl–tyrosine dephosphorylation
regulation of synapse structure or activity
Fc–gamma receptor [...] involved in phagocytosis
cellular response to interferon–gamma
interstrand cross–link repair
mRNA splicing, via spliceosome
cell differentiation
chemotaxis
taxis
positive regulation of immune response
pos. reg. of TNF superfamily cytokine production
signal transduction
interferon–gamma production
cellular response to lipopolysaccharide
immune response
pos. reg. of neuron projection development
defense response to virus
regulation of MAPK cascade
skeletal muscle organ development
regulation of BMP signaling pathway
apoptotic process
neg. reg. of cell population proliferation
spermatogenesis
response to other organism
gamete generation
neg. reg. of CD8–positive, alpha–beta T cell activation
cellular response to nitric oxide
animal organ maturation
viral release from host cell
neg. reg. of T cell cytokine production
sister chromatid cohesion
L–phenylalanine catabolic process
monoubiquitinated protein deubiquitination
IRF8
NFIL3
MBD2
NRF1
ETV1
MECP2
KLF4
SP4
TEAD1,2,4
GATA4
TYY1
KAISO
Enrichment [-log10 p-val]
whole eGRN
C1
C2
C3
C4
C5
C6
C7
Community

Figure 1: Overview and application of GRaNIE. A) Schematic of the eGRN construction by GRaNIE, including the TF to enhancer (left) and enhancer to gene (right) links (detailed workflow in **SFig. 1**). **B**) Datasets used for macrophage eGRN construction and evaluation. **C-J**): QC and summary from the eGRN from infected macrophages (**Fig S2-S3** for the other eGRNs). **C**) Distribution of significant TF-peak correlation coefficients are shown as histogram for real (black) and permuted (pink) networks. **D**) Distributions of p-values are shown for positive (left) and negative (right) peak-gene correlation. Peak-gene pairs are stratified by their distance (heat colors) and permuted gene-peak pairs (randomized network) are shown in grey. **E**) Forced-directed visualization of the infected eGRN (**SFig. 4** for the other eGRNs). The colors correspond to the communities (see J). **F**) Fraction of peak connected to the closest gene (black) vs other (grey) genes. **G**) Number of significant (FDR < 0.2; black) and highly significant (FDR < 0.05; red) peak-gene connections stratified by their distance to the TSS. **H**) Indegree: number of peaks connected to each gene **I**) Outdegree: number of genes connected to each TF. **J**) GO enrichment analysis for the whole network (left) and the 7 biggest network communities (see **STable 5** for other eGRNs).

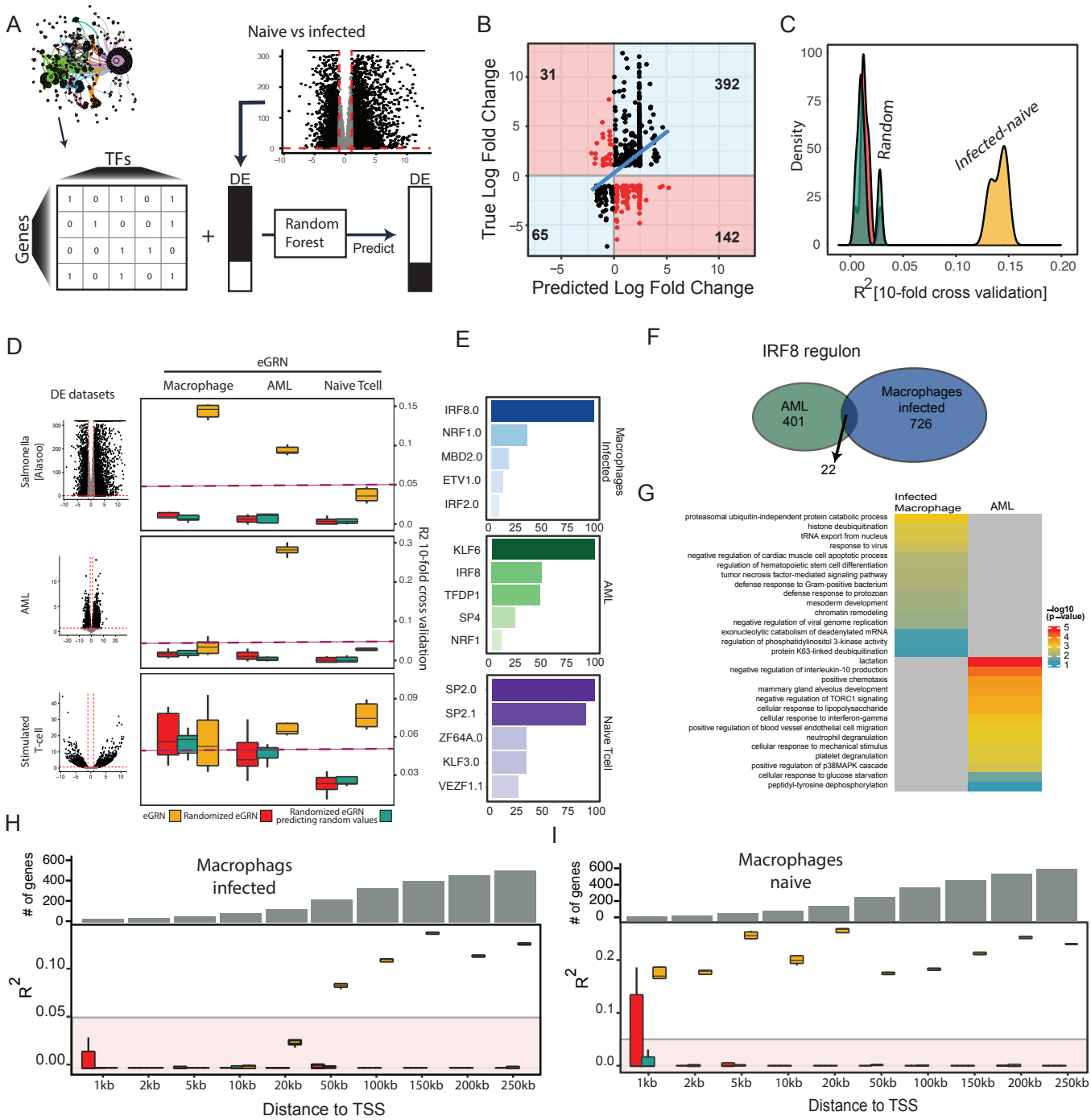

A
Naive vs infected
TFs
Genes
DE
Random Forest
Predict
B
True Log Fold Change
Predicted Log Fold Change
31
392
65
142
C
Density
Random
Infected-naive
R2[10-fold cross validation]
D
eGRN
Macrophage
AML
Naive Tcell
DE datasets
Salmonella [Alasoo]
AML
Stimulated T-cell
R2 10-fold cross validation
eGRN
Randomized eGRN
Randomized eGRN predicting random values
E
IRF8.0
NRF1.0
MBD2.0
ETV1.0
IRF2.0
Macrophages Infected
KLF6
IRF8
TFDP1
SP4
NRF1
AML
SP2.0
SP2.1
ZF64A.0
KLF3.0
VEZF1.1
Naive Tcell
F
IRF8 regulon
AML 401
Macrophages infected 726
22
G
Infected Macrophage
AML
proteasomal ubiquitin-independent protein catabolic process
histone deubiquitination
tRNA export from nucleus
response to virus
negative regulation of cardiac muscle cell apoptotic process
regulation of hematopoietic stem cell differentiation
tumor necrosis factor-mediated signaling pathway
defense response to Gram-positive bacterium
defense response to protozoan
mesoderm development
chromatin remodeling
negative regulation of viral genome replication
exonucleolytic catabolism of deadenylated mRNA
regulation of phosphatidylinositol 3-kinase activity
protein K63-linked deubiquitination
lactation
negative regulation of interleukin-10 production
positive chemotaxis
mammary gland alveolus development
negative regulation of TORC1 signaling
cellular response to lipopolysaccharide
cellular response to interferon-gamma
positive regulation of blood vessel endothelial cell migration
neutrophil degranulation
cellular response to mechanical stimulus
platelet degranulation
positive regulation of p38MAPK cascade
cellular response to glucose starvation
peptidyl-tyrosine dephosphorylation
-log10 (p-value)
H
Macrophags infected
# of genes
R2
Distance to TSS
I
Macrophages naive
# of genes
R2
Distance to TSS

Figure 2: Overview and application of GRaNPA. A) Schematic of the GRN evaluation approach GRaNPA. **B**) The output of GRaNPA is shown as true vs predicted log2 fold changes for the macrophage expression response to salmonella infection. Predictions are based on the infected eGRN (output for other macrophage eGRNs are shown in **Fig. S6**). **C**) Distribution of R2 for 10 random forest runs for the infected eGRN predicting differential expression upon salmonella infection is shown as densityplot. **D**) GRaNPA evaluation of eGRNs for infected macrophages (left), AML (middle) and naive T-Cells (right) on differential expression from Salmonella infected macrophages (top), two subtypes of AML (middle), and resting vs stimulated T-cells (bottom). **E**) Top 5 most important TFs for each of the eGRNs in D) based on prediction in the same cell-type. **F**) Intersection of IRF8 regulon in AML and infected macrophage eGRNs. **G**) GO enrichment analysis for IRF8 regulon in F. **H-I**) GRaNPA evaluation of subnetworks stratified by gene-enhancer distance for eGRNs from infected (**H**) and naive (**I**) macrophages using differential expression between infection and naive.

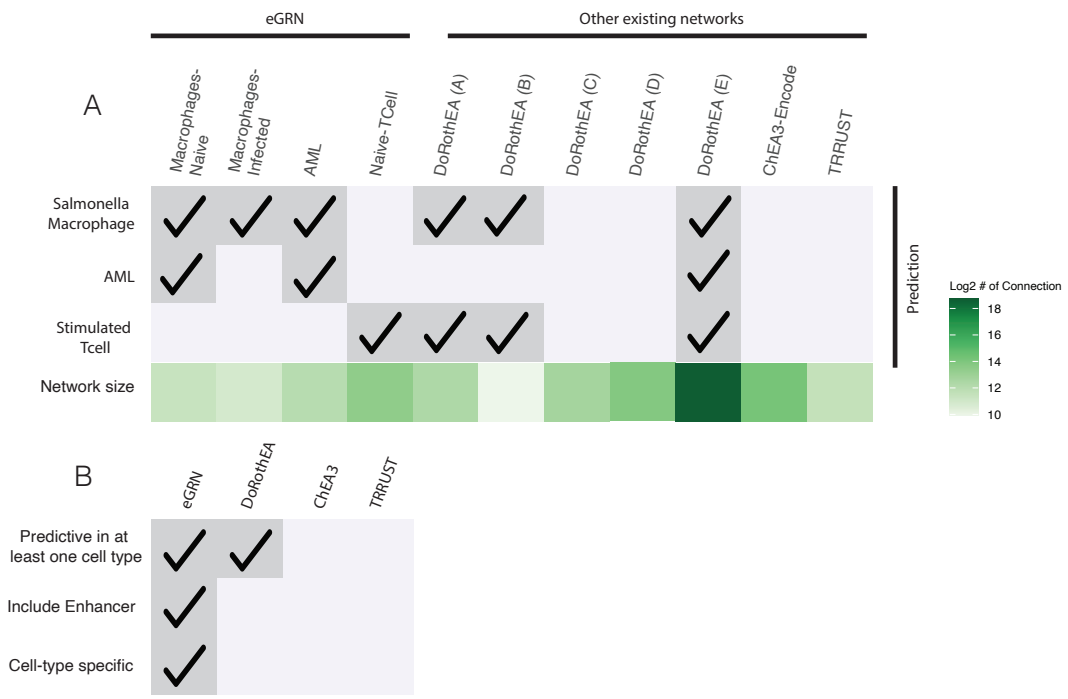

eGRN
Other existing networks
A
Macrophages-Naive
Macrophages-Infected
AML
Naive-TCell
DoRothEA (A)
DoRothEA (B)
DoRothEA (C)
DoRothEA (D)
DoRothEA (E)
ChEA3-Encode
TRRUST
Salmonella Macrophage
AML
Stimulated Tcell
Network size
Prediction
Log2 # of Connection
18
16
14
12
10
B
eGRN
DoRothEA
ChEA3
TRRUST
Predictive in at least one cell type
Include Enhancer
Cell-type specific

Figure 3: Comparison of eGRNs and other GRN approaches. A) Comparison of three eGRNs (macrophage naive, infected, AML and naive T-cells) with publicly available GRNs: DoRothEA (A)-(E), ChEA3, and TRRUST. GRNs are evaluated by GRaNIE for their performance in predicting the differential expression of naive vs Salmonella-infected macrophages, between subtypes of AML, and resting vs LPS-stimulated T-cells. For DoRothEA, (A)-(E) indicate decreasing confidence in the edges with (E) being only computationally predicted. Check marks signify the GRN passed GRaNPA assessment (prediction >0.075). **B**) Summarized evaluation of the four types of GRNs. “Predictive in at least one cell type” refers to the analysis on top of whether the GRN was able to predict differential expression at all. “Include Enhancer” refers to whether or not the GRN algorithm includes enhancers. “Cell-type specific” refers to whether or not the prediction of the GRN is specific for only one cell type.

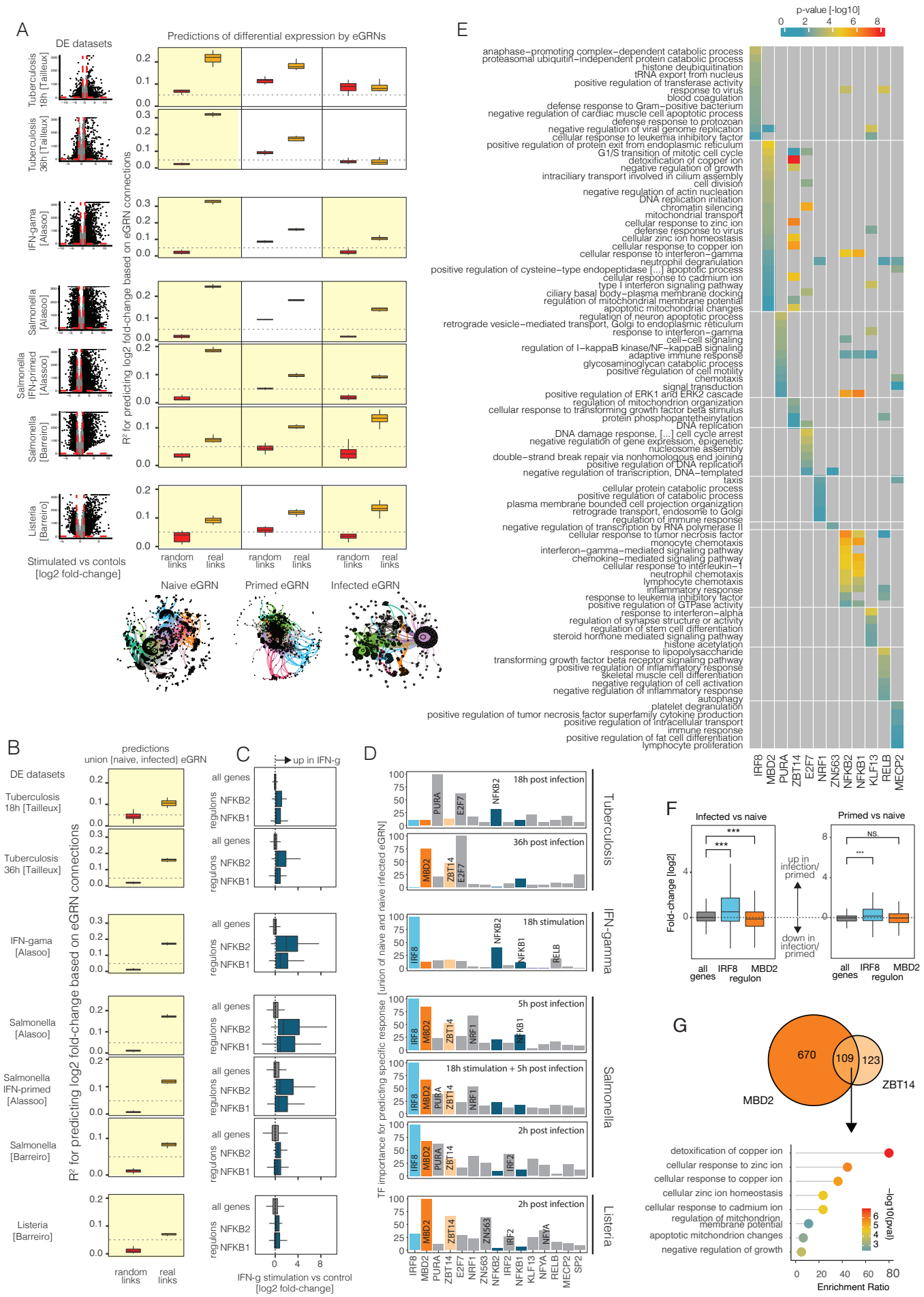

A
DE datasets
Predictions of differential expression by eGRNs
Tuberculosis 18h [Tailleux]
Tuberculosis 36h [Tailleux]
IFN-gama [Alasoo]
Salmonella [Alasoo]
Salmonella IFN-primed [Alasoo]
Salmonella [Barreiro]
Listeria [Barreiro]
R² for predicting log2 fold-change based on eGRN connections
random links
real links
Stimulated vs contols [log2 fold-change]
Naive eGRN
Primed eGRN
Infected eGRN
B
predictions union [naive, infected] eGRN
C
up in IFN-g
all genes
regulons
NFKB2
NFKB1
IFN-g stimulation vs control [log2 fold-change]
D
18h post infection
36h post infection
18h stimulation
5h post infection
18h stimulation + 5h post infection
2h post infection
Tuberculosis
IFN-gamma
Salmonella
Listeria
TF importance for predicting specific response [union of naive and naive infected eGRN]
E
p-value [-log10]
IRF8
MBD2
PURA
ZBT14
E2F7
NRF1
ZN563
NFKB2
NFKB1
KLF13
RELB
MECP2
F
Infected vs naive
Primed vs naive
Fold-change [log2]
up in infection/ primed
down in infection/ primed
all genes
IRF8 regulon
MBD2 regulon
G
670
109
123
MBD2
ZBT14
detoxification of copper ion
cellular response to zinc ion
cellular response to copper ion
cellular zinc ion homeostasis
cellular response to cadmium ion
regulation of mitchondrion membrane potential
apoptotic mitchondrion changes
negative regulation of growth
-log10(adj)
Enrichment Ratio

Figure 4: Application of GRaNIE and GRaNPA to investigate macrophage biology. A) GRaNPA evaluation of naive (left), primed (middle) and infected (right) eGRNs across seven infection settings of macrophages. Yellow shading indicates eGRN is predictive (R2>0.75) and randomized eGRN is not predictive (R2<0.05). **B**) Evaluation of union-naive (infected + naive) eGRNs. **C**) Distribution of log2 fold-change values for genes in the NFKB1 and NFKB2 regulons from the naive union eGRN (dark blue) are shown for each infection setting, alongside the response of all genes (grey). **D**) TF importance profiles for the union-naive eGRN are shown for each of the seven infection settings from B. The top 5 most predictive TFs in any of the settings are displayed. TFs with an importance score >25 are labeled. **E**) Distribution of log2 fold-change between IFN-g stimulation vs naive macrophages are shown for genes in the IRF8 (blue) and MBD2 (orange) regulons alongside the response of all genes (grey). **F**) The overlap between the MBD2 and ZBT14 regulons are shown as Venn Diagram (top). Enriched GO terms for the genes in the intersection are shown (bottom).

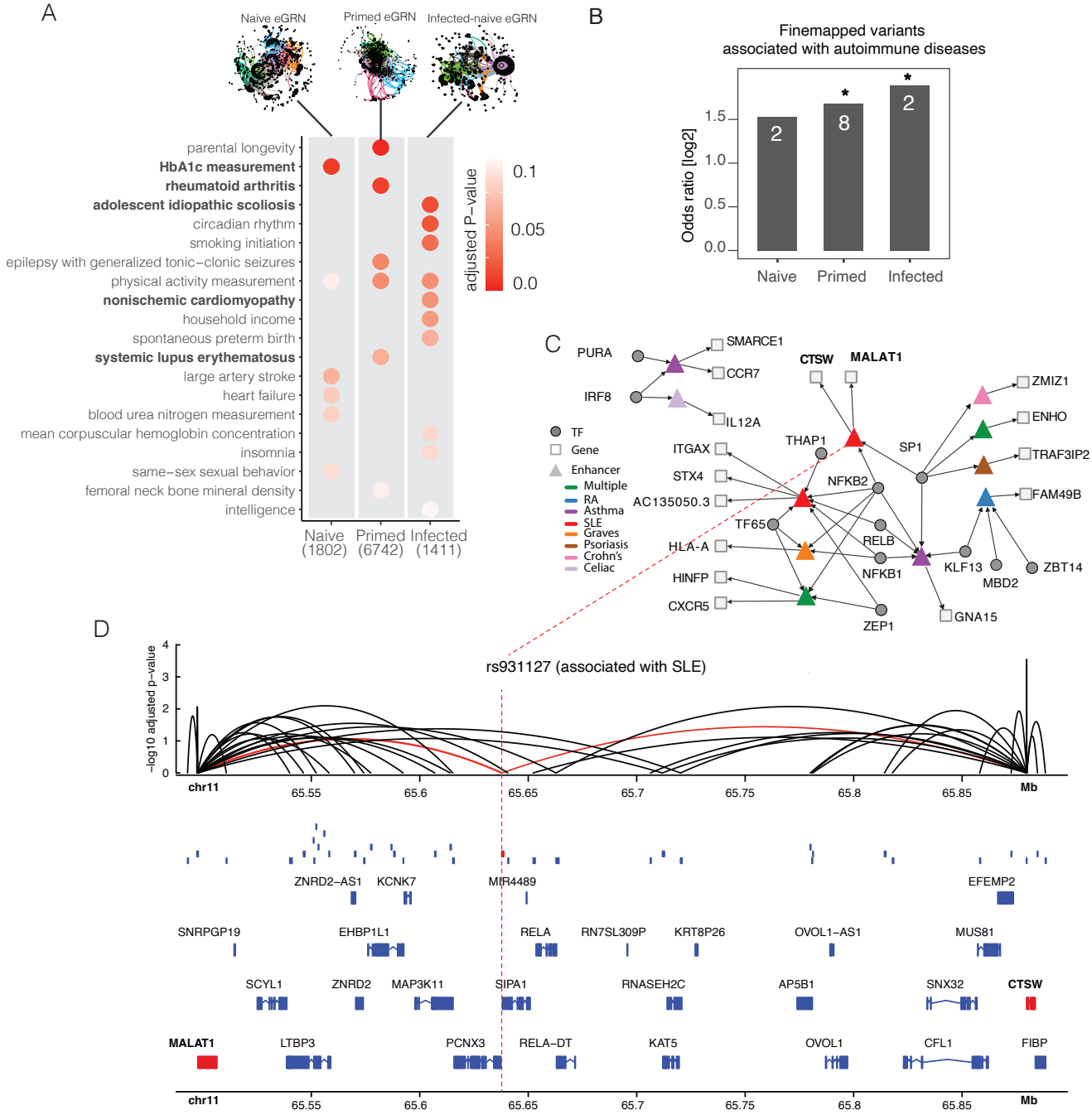

A
Naive eGRN
Primed eGRN
Infected-naive eGRN
parental longevity
HbA1c measurement
rheumatoid arthritis
adolescent idiopathic scoliosis
circadian rhythm
smoking initiation
epilepsy with generalized tonic−clonic seizures
physical activity measurement
nonischemic cardiomyopathy
household income
spontaneous preterm birth
systemic lupus erythematosus
large artery stroke
heart failure
blood urea nitrogen measurement
mean corpuscular hemoglobin concentration
insomnia
same−sex sexual behavior
femoral neck bone mineral density
intelligence
Naive (1802)
Primed (6742)
Infected (1411)
adjusted P−value
0.1
0.05
0.0
B
Finemapped variants associated with autoimmune diseases
Odds ratio [log2]
1.5
1.0
0.5
0.0
2
8
2
Naive
Primed
Infected
C
PURA
IRF8
SMARCE1
CCR7
IL12A
CTSW
MALAT1
ZMIZ1
ENHO
TRAF3IP2
FAM49B
ZBT14
MBD2
KLF13
GNA15
SP1
THAP1
NFKB2
RELB
NFKB1
ZEP1
TF65
ITGAX
STX4
AC135050.3
HLA-A
HINFP
CXCR5
TF
Gene
Enhancer
Multiple
RA
Asthma
SLE
Graves
Psoriasis
Crohn's
Celiac
D
rs931127 (associated with SLE)
−log10 adjusted p−value
chr11
65.55
65.6
65.65
65.7
65.75
65.8
65.85
Mb
ZNRD2−AS1
KCNK7
MIR4489
EFEMP2
SNRPGP19
EHBP1L1
RELA
RN7SL309P
KRT8P26
OVOL1−AS1
MUS81
SCYL1
ZNRD2
MAP3K11
SIPA1
RNASEH2C
AP5B1
SNX32
CTSW
MALAT1
LTBP3
PCNX3
RELA−DT
KAT5
OVOL1
CFL1
FIBP

Figure 5: Application of GRaNIE for investigating trait-associated SNPs. A) Heritability enrichment is shown for the naive, primed and infected macrophage eGRNs. **B**) The odds ratio of the enrichment in fine-mapped GWAS variants within the naive, primed, and infected eGRNs is shown; *: p-value <0.05; n: number of SNPs. **C**) The tripartite TF-enhancer-gene network involving all fine-mapped GWAS variants for autoimmune diseases, is shown. **D**) The genomic context of the fine-mapped, SLE-associated variant rs931127 in a ATAC-seq peak (red box) is shown, including gene tracks, other peaks present in the infected macrophage eGRN (blue boxes), and peak-gene links from the infected macrophage eGRN (arcs). Genes targeted by the peak overlapping with rs931127 (red) are colored in red.

Tables

Table 1: Summary of the eGRNs described in the main text

eGRNs	*# TFs (activators, repressors)*	*# enhan-cers*	*# genes*	*# connections TF-enhancer-genes*
Macrophage naïve	114 (52, 28)	1802	1793	3209
Macrophage interferon gamma (IFNg)-primed	126 (65, 31)	6742	3869	22082
Macrophage naïve-infected (with Salmonella)	92 (35, 26)	1411	1454	2128
Macrophage IFNg-primed- infected	78 (30, 22)	5184	2732	14697
AML	53 (30, 4)	2896	2525	5466
naive CD4+ T-cells	139 (30 ,16)	4153	4433	15868

Table 1: Network statistics for the various eGRNs based on default parameters (TF-peak FDR=0.2, peak-gene FDR=0.1, peak-gene distance <=250kb, activator/repressor stringency threshold based on the 10th percentile)

Table 2: Differential expression experiments for the different infection settings

Cell types	Treatment	set-up	reference
Monocyte-derived macrophages from healthy donors	Listeria monocytogenes	Uninfected vs 2 hours post infection	[40]
Monocyte-derived macrophages from healthy donors	Salmonella Typhimurium	Uninfected vs 2 hours post infection	[40]
Monocyte-derived macrophages from healthy donors	Mycobacterium Tuberculosis strain resistant to BDQ treatment	Uninfected vs 18h post infection	[39]
Monocyte-derived macrophages from four healthy donors	Mycobacterium Tuberculosis strain resistant to BDQ treatment	Uninfected vs 36h post infection	[39]
iPSC-derived macrophages	Salmonella typhimurium	Uninfected vs 5h post infection	[27]
iPSC-derived macrophages	Salmonella typhimurium	18h IFN-primed vs 18h IFN-primed + 5h post infection	[27]
iPSC-derived macrophages	Interferon-gamma stimulation	Naive vs 18h INF-g treatment	[27]

Table 3: Finemapped GWAS variants projected onto macrophage eGRNs

peak	TF	gene.name	FM.gwas	rsid	Disease	Network
chr8:129939305-129940672	KLF13.0.D,MBD2.0.B,ZBT14.0.C	FAM49B	chr8:129939865	rs11785995	Rheumatoid arthritis	control_naive
chr17:40598595-40599644	PURA.0.D	SMARCE1	chr17:40598769	rs9893132	Asthma	control_naive
chr19:3135922-3136231	KLF13.0.D,NFKB1.1.B,NFKB2.0.B,RELB.0.C,SP1.0.A,SP1.1.A	GNA15	chr19:3136093	rs117552144	Asthma	control_primed
chr16:31265059-31265802	NFKB1.1.B,NFKB2.0.B,RELB.0.C,TF65.0.A,THAP1.0.C,ZEP1.0.D	STX4,AC135050.3,ITGAX	chr16:31265490	rs1143679	Systemic lupus erythematosus	control_primed
chr6:30006131-30007036	NFKB1.1.B,NFKB2.0.B,TF65.0.A	HLA-A	chr6:30006148	rs4313034	Graves' disease	control_primed
chr11:65637802-65638765	NFKB2.0.B,SP1.0.A,SP1.1.A	MALAT1,CTSW	chr11:65637829	rs931127	Systemic lupus erythematosus	control_primed
chr11:118883323-118883647	NFKB2.0.B,TF65.0.A,ZEP1.0.D	CXCR5,HINFP	chr11:118883644	rs630923	Multiple sclerosis,Inflammatory bowel disease,Crohn's disease	control_primed
chr6:111605185-111606373	SP1.0.A	TRAF3IP2	chr6:111605706	rs7769061	Psoriasis	control_primed
chr9:34709959-34710335	SP1.0.A,SP1.1.A	ENHO	chr9:34710263	rs2812378	Rheumatoid arthritis,Celiac disease	control_primed
chr10:79285352-79285717	SP1.1.A	ZMIZ1	chr10:79285450	rs1250569	Crohn's disease	control_primed
chr10:79285352-79285717	SP1.1.A	ZMIZ1	chr10:79285523	rs1250568	Celiac disease	control_primed
chr17:40598595-40599644	IRF8.0.B	CCR7	chr17:40598769	rs9893132	Asthma	infected_naive
chr3:159929439-159930124	IRF8.0.B	IL12A	chr3:159929885	rs17753641	Celiac disease	infected_naive

Methods

Datasets used in this study:

Expression and chromatin accessibility data for iPSC-derived macrophages
We used a publicly available dataset (ERP020977) for naive and primed macrophages (iPSC-derived) in two conditions, uninfected and 5-h infected with Salmonella from [27]. In total we obtained 304 RNAseq profiles from 86 different individuals, of which 145 also had ATACseq data available. The samples are split into four groups: primed, primed-infected, naive and naive-infected for which 41 (43), 31 (55), 42 (42), 31 (55) paired RNA/ATAC (only RNAseq) samples were available respectively. The paired samples were used to reconstruct the GRNs with GRaNIE (see below). The unpaired RNA-seq data were used for evaluation of the eGRNs with GRaNPA (see below).

Expression data for macrophages infected with Listeria & Salmonella (GEO accession number: GSE73502): Pai et al generated expression data on cultured monocytes obtained from PBMCs of healthy donors [40]. Matured macrophages were divided into three groups: control and infected by the Listeria and Salmonella bacteria, respectively. Each group consists of 57 samples. Due to the original experiment design, only samples collected 2 hours after infection were used in the analysis.

Expression data for macrophages infected with Tuberculosis (GEO accession numbers: GSE133145, GSE143731): Giraud-Gatineau et al collected two datasets on the effect of bedaquiline (GSE133145) and five other drugs (GSE143731) treatment for *Mycobacterium tuberculosis* infection in Monocyte-derived macrophages from healthy donors [39]. The GSE133145 series consists of 16 control and 16 *M. tuberculosis*-infected samples, which are later divided into four groups: untreated / DMSO treatment (control) / two variants of bedaquiline treatment (0.5 or 5 microg/mL). Differential expression analysis revealed that differences caused by treatment are not substantial, so we considered the treatment as a controlling variable. The GSE143731 series consists of 28 control and 28 *M. tuberculosis*-infected samples, which are later divided into groups corresponding to the treatment with isoniazid (INH), rifampicin (RIF), ethambutol, pyrazinamide (PZA) or amikacin (AMK), and control group. We considered treatment as a controlling variable for the differential expression analysis.

Expression and H3K27ac data for CD4+ T-cells
Paired RNAseq and ChIPseq H3K27ac data for naive CD4-positive T-cell GRN was obtained for 138 healthy individuals samples from The International Human Epigenome Consortium [34].

Expression data for resting vs LPS-stimulated CD4+ T-cells (GSE118165)
RNAseq was obtained from [36], which measured expression in resting and stimulated subsets of CD4-positive T-cells provided. We used only the naive T-cells for differential expression analysis.

Expression and chromatin accessibility for AML
We obtained RNA-seq data for 23 AML patients from [78]. ATAC-seq data for the same patients was obtained from He et al (*in revision*).

Differential expression analysis (see also Table 2)

Differential expression analysis was performed for all genes using DESeq2 [79], typically using the contrast between treatment and no treatment or disease and control. Dataset-specific details of the differential expression analysis datasets are described below.

iPSC-derived macrophages infected with Salmonella from [27]: Differential expression was calculated using only the RNA-seq data that was not used for eGRN reconstruction. We quantified differential expression for the following contrasts: naive vs infected, naive vs IFN-g primed, IFN-g primed vs IFN-g primed-infected. The formula used in DESeq2 was "~condition".

Macrophages infected with Listeria and Salmonella from [40]: We analysed the differential expression between control samples, listeria-infected samples and salmonella-infected samples separately. No samples were removed. Information on the donor was used as a covariate, using the design formula: "~patient + condition".

Macrophages infected with Tuberculosis from [39]: We calculated differential expression between monocyte-derived macrophages from healthy donors infected with tuberculosis versus control samples. Datasets GSE133145 and GSE143731 were analyzed separately, but with a common design formula. Although there were also multiple treatments, the expression variance was almost exclusively driven by the difference in disease status. We therefore added the treatment as a covariate to the design formula ("~patient + treatment + condition"), but only investigated differential expression between infected macrophages and controls. One control and one infected sample from the GSE143731 series were removed from the analysis as outliers.

CD4+ naive T-cells resting vs LPS-stimulated
We quantified differential expression between CD4-positive naive T-cells in resting vs stimulated condition [36]. The design formula used in DESeq is "~condition".

AML subtypes
Differential expression was calculated using data from [78] and comparing samples with high leukemia stem cell burden (GPR56-high) vs low leukemia stem cell burden (GPR56-low samples) based on immunophenotyping as defined in [78]. The design formula was: "~GPR56status".

GRaNIE: Construction of eGRNs

The following is needed as input for GRaNIE:

1. Raw or pre-normalized chromatin accessibility data (e.g., ATAC-seq, DNase-Seq or histone modification ChIP-seq data such as H3K27ac)
2. Raw or pre-normalized RNA-seq counts

3. Pre-compiled lists of TFBS predictions per TF (we provide predictions for human and mouse TFBS that were derived as described in Berest, Arnold et al. 2019)
4. TAD domains (optional)

For all datasets in this study, we used the same default parameters when constructing the eGRNs as described below.

GRaNIE is conceptually based on the procedure described in [7] and has the following main steps:

(1) Process chromatin accessibility and RNA-seq data

Both ATAC-seq and RNA-seq may be raw counts or normalized counts. If raw counts are provided for RNA-seq, by default we quantile normalized the RNA-Seq count data in order to minimize the effects of outlier values that may otherwise have a large influence on the resulting correlations. For chromatin accessibility data, we employ a size factor normalization as implemented in DESeq2 [79]. However, the user can define which type of normalization shall be used for either data. Additional filters for excluding particular chromosomes (e.g., sex chromosomes) or genes / peaks with low counts can optionally be used. The latter is implemented by removing genes / peaks if the average counts across all samples are below a specified threshold (5 by default).

(2) Overlap TF binding sites with ATAC-Seq peaks

Based on the provided list of putative TFBS per TF (see [26] for details), we overlap all TFBS from all TF with the open chromatin peaks and record for each peak and TF whether at least one putative TFBS is located within the peak. This binary TF-peak binding matrix is used in subsequent steps.

(3) Identify statistically significant TF - peak connections

To identify statistically significant TF-peak connections we implement a cell-type specific data-driven approach. In brief, we first calculate the Pearson correlation coefficients between the expression level of each TF and the open chromatin signal of each peak across all samples.

We then use an empirical FDR procedure to identify statistically significant TF-peak connections. For this, for each TF, we split the peaks into two sets: A foreground set containing the peaks with a putative TFBS and a corresponding background set consisting of peaks without putative TFBS based on the TF-peak binding matrix calculated above. We then discretize the TF-peak correlation *r* into 40 bins in steps of 0.05 ranging from -1, -0.95, …, 0, …, 1 and calculate a bin-specific FDR value using two different directions (*positive*: left to right from -1 to 1, *negative*: right to left from 1 to -1). For each bin (correlation threshold) *k*, we calculate the empirical FDR according to the formula $efdr_k = \frac{nfp_k}{nfp_k + ntp_k}$, with nfp_k and ntp_k denoting the total number of TF-peaks in the background and foreground, respectively, for which $r \geq k$ (direction positive) *and* $r < k$ (direction negative). To make the numbers from foreground and background compatible, we normalize nfp_k beforehand by their ratio (i.e., $nfp_k = nfp_k \times \frac{ntp}{nfp}$, with ntp and nfp denoting the total number of TF-peaks in the foreground and background, respectively.

(4) *Activator-repressor TF classification (optional)*
Optionally, the TF classification as described in [26] can be run and is fully integrated in GRaNIE. It produces a classification of TFs into putative activators, repressors, or undetermined. Briefly, it compares the distribution of correlations for peaks with putative binding sites (foreground) against all other peaks (i.e., background) and classifies TFs depending on whether the correlations of putative targets are significantly more positive than (activator), more negative than (repressor), or indistinguishable from (undetermined) the background.

(5) *Identify statistically significant peak-gene connections*
Next, we add peak-gene connections to our network. We identify highly correlated peak-gene pairs based on their Pearson correlation and the associated p-value (using cor.test in R) between the normalized RNA-seq for the expression of a gene and the corresponding open chromatin peak.
GRaNIE offers two options to decide which peak-gene pairs to test for correlation: in absence of additional topologically associating domain (TADs) data from Hi-C or similar approaches it uses a local neighborhood-based approach with a custom neighborhood size (default: 250 kb up- and downstream of the peak) to select peak-gene pairs to test. In the presence of TAD data, only peak-gene pairs within a TAD are tested. The user has furthermore the choice to specify whether overlapping TADs should be merged or not. We offer two options of where in the gene the overlap with the extended peak may occur: at the 5' end of the gene (the default) or anywhere in the gene.
GRaNIE also records additional properties for each peak-gene pair such as their distance as well as gene type & status as annotated by Gencode. By default, only protein-coding and lincRNA genes are kept in the eGRN, but this can be customized to include other gene types.

Quality controls
We implemented a range of quality controls for the different steps to assess signal vs noise in our eGRN. The package optionally offers PCA plots for both the RNA-seq as well as open chromatin data, and upon availability of additional metadata that can be provided, the PCA data can also be colored accordingly. This facilitates the detection of batch effects and outlier samples that may introduce unwanted variation.
For assessing TF-enhancer links, as described in the main text, we recommend a permutation-based approach in addition to the real data for comparison. We provide heatmaps and boxplots that compare the connectivity for the real and random eGRNs.
For assessing enhancer-gene links we base our quality controls on the assumption that peak accessibility and gene expression is positively correlated. Thus assessing the amount of signal (i.e. small p-values) of positive correlations vs negative correlations serves as a proxy for signal to noise. We then assess the signal to noise in our real network vs a set of random peak-gene links. Specifically, we expect that positive correlations outnumber negative in the real network, and not in the randomized network. We also expect that with increasing distance the fraction of pos vs neg correlations gets smaller.

Lastly, we also perform the same procedure for randomized peak-gene combinations to empirically model the background distribution. For this, after selecting all peak-gene pairs, the peak IDs are shuffled and the same procedure as described above is run. We include a number of QC plots as part of the output.

(6) Filter GRN connections and calculate peak-gene FDR

Lastly, we offer a variety of options to combine and filter TF-peaks and peak-genes to derive the full GRN for subsequent analyses. For example, both types of connections can be filtered by their FDRs or by their correlation, peak-gene links can further be filtered by their distance, gene type, and other criteria. By default, we retain only peak-gene pairs that are positively correlated. After applying all filters for the peak-gene links, multiple testing adjustment is performed using Benjamini-Hochberg. The default thresholds for TF-peak and peak-gene links are FDR < 0.2 and FDR < 0.1, respectively.

Downstream analyses implemented in GRaNIE

The following functionalities are available within the GRaNIE package:

- Descriptive statistics pertaining to the structure of the graph, such as the number of nodes and edges and their types, the distribution of node degrees, and the top nodes with regard to degree centrality and eigenvector centrality.
- Community identification, for which the package supports multiple algorithms (louvain, walktrap, leading eigenvector, fast greedy, and optimal). The communities can be supplemented with descriptive statistics similar to those previously described, but specific to each individual community.
- Enrichment analyses in three different flavours; a general enrichment analysis for the whole network, a community-based enrichment analysis, or a TF-based enrichment analysis.

GRaNPA: Prediction model

For the prediction model, first we need to compute the differential expression resulting from a perturbation between two different conditions, for which we used DESeq2. We define DE(j) as the differential expression value for j^{th} Gene.

Second, we need to define the GRN as a mathematical function:

$$grn(gene, tf) = \begin{cases} 1 & tf \text{ and gene connected in the GRN} \\ 0 & otherwise \end{cases}$$

Based on this, we construct matrix *X*, a relation matrix between genes and TFs where each row is a gene and each column is a TFs. There is a 1 in the cell for a gene and TF if and only if they are connected through the GRN.

Finally, we use Random Forest (RF) Regression based on the following formula to predict a differential expression value for each gene *i* based on its connected TFs X[*i,tf*]:

$$DE(i) = \sum_{tf \in TFs} \beta tf \times X[i, tf]$$

RF has been implemented using the “ranger” package in R. To avoid overfitting we use 10-fold cross validation. We measure performance by assessing the cross-validation R2.

GRaNPA: Construction of the randomized network to assess edge-specificity

To assess the edge-specificity of our network during the random forest regression we constructed a randomized control network based on the structure of the actual GRN. It has the same number of edges with the same degree distribution for TFs with only the gene labels permutated so the connection between TFs and genes are randomized. The differential expression values are unchanged. In case we use any weighing method for the edges, the same method will be applied for each edge of the randomized version to recalculate weight if needed.

GRaNPA: Random signal generation to assess overfitting

To control for over-fitting in the random forest regression, we used the real network structure and assigned randomly generated values to genes instead of their differential expression value. The random values are chosen from a uniform distribution with minimum and maximum of -5 and 5 respectively. Any prediction value for this network is a consequence of overfitting. The reason we used a uniform distribution is that we want to check overfitting and we don't want the distribution to be like any differential expression distribution.

Parameter optimization using GRaNPA

We optimized the parameters for the GRN reconstruction part based on the performance of the network on predicting the selected differential expression response. These parameters include peak-gene FDR and TF-peak FDR. In each optimization process, we fixed all parameters except one and chose the value with highest performance of the respective network.

Benchmarking against other networks / tools

Dorothea: Dorothea [11] is a resource containing TF-target interactions. The connections are tagged by a confidence level based on the number of supporting evidence. The confidence levels are from A considered to be highly reliable to B-D reserved for curated and/or ChIP-seq interactions and the lowest confidence level E, for connection with only computational support.

TRRUST: TRRUST [16] is a database of transcriptional regulatory networks for humans and mice. It has been constructed using text-mining followed by manual curation. TRRUST v2 regulatory network for humans contains 795 TFs, 2067 genes and 8427 regulatory links. We did not use any weighing for the edges as it was not provided by TRRUST v2.

ChEA3: ChEA3 [17] are TF-target gene libraries containing targets determined by ChIP-seq experiments from ENCODE, ReMap and publications. It also contains co-expression connections based on RNA-seq data from resources like GREx and ARCHs4.

Assessing the cell-type specificity of network models

In order to assess cell-type specificity of network models, we check their performance on different DE data from other cell-types. In this analysis, we used naive-macrophages, naive CD4 positive T-cell and AML and evaluated their performance for DE data from GPR56-positive vs GPR56-negative AML, resting vs stimulated CD4 positive T-cell (in 10 different subcell-types) and naive-macrophages vs 5-h infected with Salmonella. We filtered genes using 0.1 adjusted p-value and 1 log fold change thresholds.

Calculating variable importance measures

We used the permutation approach to measure variable importance using the “ranger” package in R. This accuracy-based approach uses the out-of-bag sample to calculate the importance of a specific variable. The importance is based on the difference in the prediction accuracy of out-of-bag sample and the prediction accuracy of out-of-bag sample while its variables have been randomly shuffled while all other variables kept the same.

Visualization (ShinyApp)

We provide a web application based on ShinyApp for interactive visualization of the eGRNs for different cell types(https://apps.embl.de/grn/).

Gene set enrichment analysis

Pre-ranked gene set enrichment analysis (GSEA, ranking based on log2 fold change) was performed using the Bioconductor/R package “Fast Gene Set Enrichment Analysis” [80]. The M1 macrophage signature gene set was obtained from [81].

GWAS enrichment

We tested if the macrophage GRNs were enriched for genetic heritability of GWAS traits using stratified linkage disequilibrium score regression (S-LDSC) [48]. 806 GWAS summary statistics that included participants of European descent were downloaded from the GWAS catalog [82], harmonized and converted to the LDSC format as described on the LDSC github repository (https://github.com/bulik/ldsc). We removed summary statistics with fewer than 10,000 individuals and fewer than 100,000 SNPs, because those studies were likely underpowered, leaving 442 traits. We created peak sets for each of the three macrophage GRNs and the union of the three networks by extracting the enhancer regions that were present in the GRNs after filtering for peak-gene distance, TF-peak FDR and gene differential expression effect size. We used 54 sets of general genomic features (downloaded from https://alkesgroup.broadinstitute.org/LDSCORE/, following Finucane et al. 2015) and a peak set based on all enhancers identified in the macrophage ATAC-seq as background regions. Adding these regions as a background ensures that the identified enrichments are not due to the general enrichment of heritability in (macrophage) enhancers, but specifically in the enhancers that are part of the GRNs. We calculated the heritability enrichment p-values and corrected them for multiple testing within each trait.

GO enrichment of GRNs

The general enrichment analysis was run using topGO (v2.42.0), with the foreground being the genes in the filtered GRN, and the background being the genes within a predefined 250kb neighborhood of the peaks in the GRN. In more specific enrichment analyses such as those for the top transcription factors, which are ranked by their predictive capacity, the foreground is selected based on the genes a given TF is connected to within the network. Similarly, in community-based enrichment analyses, the foreground is simply the genes that are classified to a given community. To calculate the enrichment, a Fisher test is used alongside the weight01 algorithm, which is a mixture of the “elim” and the “weight” algorithms introduced by [83], to account for the GO hierarchy. Additionally, terms with less than 4 significantly annotated genes were omitted from the results in the figures. For the sake of better visualizing the enriched terms, the figures are limited to the top ten enriched terms per category. The full list of enriched terms can be found in supplementary tables 5 and 6.

Bibliography

1. Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment strategies. Trends Mol Med. 2021;27:1060–73.

2. Karnuta JM, Scacheri PC. Enhancers: bridging the gap between gene control and human disease. Hum Mol Genet. 2018;27:R219–27.

3. Freimer JW, Shaked O, Naqvi S, Sinnott-Armstrong N, Kathiria A, Chen AF, et al. Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. BioRxiv. 2021.

4. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10.

5. Xu Q, Georgiou G, Frölich S, van der Sande M, Veenstra GJC, Zhou H, et al. ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. Nucleic Acids Res. 2021.

6. Janssens J, Aibar S, Taskiran II, Ismail JN, Spanier KI, Bravo Gonzalez-Blas C, et al. Decoding gene regulation in the fly brain. BioRxiv. 2021.

7. Reyes-Palomares A, Gu M, Grubert F, Berest I, Sa S, Kasowski M, et al. Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension. Nat Commun. 2020;11:1673.

8. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS ONE. 2010;5.

9. Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, Brent MR. Mapping functional transcription factor networks from gene expression data. Genome Res. 2013;23:1319–28.

10. Kamimoto K, Hoffmann CM, Morris SA. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. BioRxiv. 2020.

11. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.

12. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35:2159–61.

13. Chan TE, Stumpf MPH, Babtie AC. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Syst. 2017;5:251-267.e3.

14. Huynh-Thu VA, Geurts P. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. Sci Rep. 2018;8:3384.

15. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;29:1363–75.

16. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded

reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46:D380–6.

17. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 2019;47:W212–24.

18. Liu Z-P, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database (Oxford). 2015;2015.

19. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.

20. Schraivogel D, Gschwind AR, Milbank JH, Leonce DR, Jakob P, Mathur L, et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat Methods. 2020;17:629–35.

21. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell. 2018;71:858-871.e8.

22. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics. 2018;19:232.

23. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17:147–54.

24. Stone M, Li J, Grace McCalla S, Fotuhi Siahpirani A, Periyasamy V, Shin J, Roy S. Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. bioRxiv 2021.06.01.446671;

25. Weidemüller P, Kholmatov M, Petsalaki E, Zaugg JB. Transcription factors: Bridge between cell signaling and gene regulation. Proteomics. 2021;:e2000034.

26. Berest I, Arnold C, Reyes-Palomares A, Palla G, Rasmussen KD, Giles H, et al. Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. Cell Rep. 2019;29:3147-3159.e12.

27. Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet. 2018;50:424–31.

28. Novikova G, Kapoor M, Tcw J, Abud EM, Efthymiou AG, Chen SX, et al. Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. Nat Commun. 2021;12:1610.

29. Bunina D, Germain P-L, Lopez Tobon A, Fernandez-Novel Marx N, Arnold C, Ó hEachteirn A, et al. Pathological LSD1 mutations cause HDAC-mediated aberrant gene repression during early cell differentiation. BioRxiv. 2021.

30. Carty M, Zamparo L, Sahin M, González A, Pelossof R, Elemento O, et al. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun. 2017;8:15454.

31. Ouma WZ, Pogacar K, Grotewold E. Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. PLoS Comput Biol. 2018;14:e1006098.

32. Langlais D, Barreiro LB, Gros P. The macrophage IRF8/IRF1 regulome is required for protection against infections and is associated with chronic inflammation. J Exp Med. 2016;213:585–603.

33. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008:P10008.

34. Bujold D, Morais DA de L, Gauthier C, Côté C, Caron M, Kwan T, et al. The international human epigenome consortium data portal. Cell Syst. 2016;3:496-499.e2.

35. Garg S, Reyes-Palomares A, He L, Bergeron A, Lavallée V-P, Lemieux S, et al. Hepatic leukemia factor is a novel leukemic stem cell regulator in DNMT3A, NPM1, and FLT3-ITD triple-mutated AML. Blood. 2019;134:263–76.

36. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Müller F, Nguyen V, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat Genet. 2019;51:1494–505.

37. Holland CH, Szalai B, Saez-Rodriguez J. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. Biochim Biophys Acta Gene Regul Mech. 2020;1863:194431.

38. Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. 2020;21:36.

39. Giraud-Gatineau A, Coya JM, Maure A, Biton A, Thomson M, Bernard EM, et al. The antibiotic bedaquiline activates host macrophage innate immune resistance to bacterial infection. eLife. 2020;9.

40. Pai AA, Baharian G, Pagé Sabourin A, Brinkworth JF, Nédélec Y, Foley JW, et al. Widespread shortening of 3' untranslated regions and increased exon inclusion are evolutionarily conserved features of innate immune responses to infection. PLoS Genet. 2016;12:e1006338.

41. Medzhitov R, Horng T. Transcriptional control of the inflammatory response. Nat Rev Immunol. 2009;9:692–703.

42. Leseigneur C, Lê-Bury P, Pizarro-Cerdá J, Dussurget O. Emerging evasion mechanisms of macrophage defenses by pathogenic bacteria. Front Cell Infect Microbiol. 2020;10:577559.

43. Chistiakov DA, Myasoedova VA, Revin VV, Orekhov AN, Bobryshev YV. The impact of interferon-regulatory factors to macrophage differentiation and polarization into M1 and M2. Immunobiology. 2018;223:101–11.

44. Jones G-R, Brown SL, Phythian-Adams AT, Ivens AC, Cook PC, MacDonald AS. The Methyl-CpG-Binding Protein Mbd2 Regulates Susceptibility to Experimental Colitis via Control of CD11c+ Cells and Colonic Epithelium. Front Immunol. 2020;11:183.

45. Wang Y, Zhang L, Wu G-R, Zhou Q, Yue H, Rao L-Z, et al. MBD2 serves as a viable target against pulmonary fibrosis by inhibiting macrophage M2 program. Sci Adv. 2021;7.

46. Stafford SL, Bokil NJ, Achard MES, Kapetanovic R, Schembri MA, McEwan AG, et al. Metal ions in macrophage antimicrobial pathways: emerging roles for zinc and copper. Biosci Rep. 2013;33.

47. Festa RA, Thiele DJ. Copper at the front line of the host-pathogen battle. PLoS Pathog. 2012;8:e1002887.

48. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228–35.

49. Udalova IA, Mantovani A, Feldmann M. Macrophage heterogeneity in the context of rheumatoid arthritis. Nat Rev Rheumatol. 2016;12:472–85.

50. Ma C, Xia Y, Yang Q, Zhao Y. The contribution of macrophages to systemic lupus erythematosus. Clin Immunol. 2019;207:1–9.

51. Harigai M, Kawamoto M, Hara M, Kubota T, Kamatani N, Miyasaka N. Excessive production of IFN-gamma in patients with systemic lupus erythematosus and its contribution to induction of B lymphocyte stimulator/B cell-activating factor/TNF ligand superfamily-13B. J Immunol. 2008;181:2211–9.

52. Kato M. New insights into IFN-γ in rheumatoid arthritis: role in the era of JAK inhibitors. Immunol Med. 2020;43:72–8.

53. Rönnblom L, Leonard D. Interferon pathway in SLE: one key to unlocking the mystery of the disease. Lupus Sci Med. 2019;6:e000270.

54. Andreou N-P, Legaki E, Gazouli M. Inflammatory bowel disease pathobiology: the role of the interferon signature. Ann Gastroenterol. 2020;33:125–33.

55. Gatchalian J, Liao J, Maxwell MB, Hargreaves DC. Control of Stimulus-Dependent Responses in Macrophages by SWI/SNF Chromatin Remodeling Complexes. Trends Immunol. 2020;41:126–40.

56. Buhling F, Kellner U, Guenther D, Kahl S, Brömme D, Weber E, et al. Characterization of novel anti-cathepsin W antibodies and cellular distribution of cathepsin W in the gastrointestinal tract. Biol Chem. 2002;383:1285–9.

57. Pires D, Marques J, Pombo JP, Carmo N, Bettencourt P, Neyrolles O, et al. Role of Cathepsins in Mycobacterium tuberculosis Survival in Human Macrophages. Sci Rep. 2016;6:32247.

58. Sigalova OM, Shaeiri A, Forneris M, Furlong EE, Zaugg JB. Predictive features of gene expression variation reveal mechanistic link with differential expression. Mol Syst Biol. 2020;16:e9539.

59. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A,

et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

60. Zeitlinger J. Seven myths of how transcription factors read the cis-regulatory code. Current Opinion in Systems Biology. 2020;23:22–31.

61. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593:238–43.

62. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, et al. Population variation and genetic control of modular chromatin architecture in humans. Cell. 2015;162:1039–50.

63. Ibarra IL, Hollmann NM, Klaus B, Augsten S, Velten B, Hennig J, et al. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. Nat Commun. 2020;11:124.

64. Bunina D, Abazova N, Diaz N, Noh K-M, Krijgsveld J, Zaugg JB. Genomic Rewiring of SOX2 Chromatin Interaction Network during Differentiation of ESCs to Postmitotic Neurons. Cell Syst. 2020;10:480-494.e8.

65. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell. 2020;183:1103-1116.e20.

66. Hainer SJ, McCannell KN, Yu J, Ee L-S, Zhu LJ, Rando OJ, et al. DNA methylation directs genomic localization of Mbd2 and Mbd3 in embryonic stem cells. eLife. 2016;5.

67. Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P, Klein F, et al. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell. 1992;69:905–14.

68. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015;528:575–9.

69. Qin W, Scicluna BP, van der Poll T. The role of host cell DNA methylation in the immune response to bacterial infection. Front Immunol. 2021;12:696280.

70. Yue T, Sun F, Wang F, Yang C, Luo J, Rong S, et al. MBD2 acts as a repressor to maintain the homeostasis of the Th1 program in type 1 diabetes by regulating the STAT1-IFN-γ axis. Cell Death Differ. 2021.

71. Fadini GP, Cappellari R, Mazzucato M, Agostini C, Vigili de Kreutzenberg S, Avogaro A. Monocyte-macrophage polarization balance in pre-diabetic individuals. Acta Diabetol. 2013;50:977–82.

72. Samaan MC, Missiuna P, Peterson D, Thabane L. Understanding the role of the immune system in adolescent idiopathic scoliosis: Immunometabolic CONnections to Scoliosis (ICONS) study protocol. BMJ Open. 2016;6:e011812.

73. Keller M, Mazuch J, Abraham U, Eom GD, Herzog ED, Volk H-D, et al. A circadian clock in macrophages controls inflammatory immune responses. Proc Natl Acad Sci USA. 2009;106:21407–12.

74. Rana A, Musto AE. The role of inflammation in the development of epilepsy. J

Neuroinflammation. 2018;15:144.

75. Liao X, Shen Y, Zhang R, Sugi K, Vasudevan NT, Alaiti MA, et al. Distinct roles of resident and nonresident macrophages in nonischemic cardiomyopathy. Proc Natl Acad Sci USA. 2018;115:E4661–9.

76. Chiba T, Umegaki K. Pivotal roles of monocytes/macrophages in stroke. Mediators Inflamm. 2013;2013:759103.

77. Boyle AK, Rinaldi SF, Norman JE, Stock SJ. Preterm birth: Inflammation, fetal injury and treatment strategies. J Reprod Immunol. 2017;119:62–6.

78. Garg S, Reyes-Palomares A, He L, Bergeron A, Lavallée V-P, Lemieux S, et al. Hepatic leukemia factor is a novel leukemic stem cell regulator in DNMT3A, NPM1, and FLT3-ITD triple-mutated AML. Blood, The Journal of the American Society of Hematology. 2019;134:263–76.

79. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

80. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. BioRxiv. 2016.

81. Orecchioni M, Ghosheh Y, Pramod AB, Ley K. Macrophage Polarization: Different Gene Signatures in M1(LPS+) vs. Classically and M2(LPS-) vs. Alternatively Activated Macrophages. Front Immunol. 2019;10:1084.

82. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12.

83. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22:1600–7.