Mapping disease regulatory circuits at cell-type resolution from single-cell multiomics data

Xi Chen[1], Yuan Wang[2], Antonio Cappuccio[3], Wan-Sze Cheng[3], Frederique Ruf Zamojski[3], Venugopalan Nair[3], Clare M Miller[3], Aliza Rubenstein[3], German Nudelman[3], Alicja Tadych[2], Chandra Theesfeld[2], Alexandria Vornholt[3], Mary-Catherine George[3], Felicia Ruffin[4], Michael Dagher[4], Daniel Chawla[5], Alessandra Soares-Schanoski[3], Rachel R. Spurbeck[6], Lishomwa C. Ndhlovu[7], Robert Sebra[11], Steven Kleinstein[5,8], Andrew G. Letizia[9], Irene Ramos[3,10], Vance G. Fowler Jr[4], Christopher W. Woods[4], Elena Zaslavsky[3,*], Olga G. Troyanskaya[1,2,*], Stuart C. Sealfon[3,*,]

1. Center for Computational Biology, Flatiron Institute, New York, NY, USA
2. Department of Computer Science and Lewis-Sigler Institute of integrative genomics, Princeton University, Princeton, NJ, USA
3. Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
4. Division of Infectious Diseases, Department of Medicine, Duke University School of Medicine, Durham, NC, USA
5. Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
6. Battelle Memorial Institute, Columbus, OH, USA
7. Division of Infectious Diseases, Department of Medicine, Weill Cornell Medicine, New York, NY, USA
8. Department of Pathology and Department of Immunobiology, Yale School of Medicine, New Haven, CT, USA
9. Naval Medical Research Center, Silver Spring, MD, USA
10. Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
11. Department of Genetics and Genomic Sciences, the Icahn School of Medicine at Mount Sinai

* Corresponding authors

Lead contact: stuart.sealfon@mssm.edu (S.C.S.)

Abstract

Resolving chromatin remodeling-linked gene expression changes at cell type resolution is important for understanding disease states. We describe MAGICAL, a hierarchical Bayesian approach that leverages paired scRNA-seq and scATAC-seq data from different conditions to map disease-associated transcription factors, chromatin sites, and genes as regulatory circuits. By simultaneously modeling signal variation across cells and conditions in both omics data types, MAGICAL achieved high accuracy on circuit inference. We applied MAGICAL to study *Staphylococcus aureus* sepsis from peripheral blood mononuclear single-cell data that we generated from infected subjects with bloodstream infection and from uninfected controls. MAGICAL identified sepsis-associated regulatory circuits predominantly in CD14 monocytes, known to be activated by bacterial sepsis. We addressed the challenging problem of distinguishing host regulatory circuit responses to methicillin-resistant- (MRSA) and methicillin-susceptible *Staphylococcus aureus* (MSSA) infections. While differential expression analysis failed to show predictive value, MAGICAL identified epigenetic circuit biomarkers that distinguished MRSA from MSSA.

Introduction

Gene expression can be modulated through the interplay of proximal and distal regulatory domains brought together in 3D space[1]. Chromatin regulatory domains, transcription factors, and downstream target genes form regulatory circuits[2]. Within circuits, the binding of transcription factors to chromatin regions and the 3D looping between these regions and gene promoters represent the mechanisms governing how transcription factors transform regulatory signals into changes in RNA transcription[3,4]. In disease, these circuits could be dysregulated in a cell type specific manner and may not be observed from bulk samples[5]. Therefore, identifying the impact of disease on regulatory circuits requires a framework for mapping regulatory domains with chromatin accessibility changes to altered gene expression in the context of cell-type resolution[6]. Single-cell data (scRNA-seq and scATAC-seq) characterizing disease states have

improved the identification of differential chromatin sites and/or differentially expressed genes within individual cell types[5,7,8]. Yet, advances in single-cell assay technology have outpaced the development of methods to maximize the value of multiomics datasets for studying disease-associated regulation, especially for the regulatory interactions that are not directly measured by the omics data. Recent computational approaches[9-12] to support the multiomics data analysis demonstrate the promise of this area but still lack the capacity to resolve regulation changes within individual cell types, which precludes elucidating regulatory circuits affected by the disease or showing different responses in varying disease states.

To address these, we developed MAGICAL (Multiome Accessibility Gene Integration Calling And Looping), a method that models coordinated chromatin accessibility and gene expression variation to identify circuits (both the units and their interactions) that differ between conditions. MAGICAL analyzes scRNA-seq and scATAC-seq data using a hierarchical Bayesian framework. To accurately detect differences in regulatory circuit activity between conditions, MAGICAL introduces hidden variables for explicitly modeling the transcriptomic and epigenetic signal variations between conditions and optimization against the noise in both scRNA-seq and scATAC-seq datasets. Because regulatory circuits are cell-type specific[13], MAGICAL reconstructs them at cell-type resolution. Systematic benchmarking against multiple public datasets supported the accuracy of MAGICAL-identified regulatory circuits. The circuit chromatin sites and genes that have coordinated changes between conditions are more robustly associated with the disease condition than the sites or genes obtained from one type of omics data identified through differential analysis.

Staphylococcus aureus (*S. aureus*), a bacterium often resistant to common antibiotics, is a major cause of severe infection and mortality[14,15]. Using single-cell multiomics data generated from PBMC samples of *S. aureus* infected subjects and healthy controls, MAGICAL identified host response regulatory circuits that are modulated during *S. aureus* bloodstream infection, and circuits that discriminate the responses to methicillin-resistant (MRSA) and methicillin-susceptible *S. aureus* (MSSA). Genes in the host circuits accurately predicted *S. aureus* infection in multiple validation datasets.

Moreover, in contrast to conventional differential analysis that failed to identify specific genes for robust antibiotic-sensitivity prediction, MAGICAL-identified circuit genes can differentiate MRSA from MSSA. Therefore, MAGICAL can be used for multiomics data-based gene signature development, providing a bioinformatic solution that can improve disease diagnosis.

Results

MAGICAL framework

MAGICAL identifies disease-associated regulatory circuits by comparing single-cell multiomics data (scRNA-seq and scATAC-seq) from disease and control samples (**Fig.1a**). The framework incorporates transcription factor (TF) motifs and chromatin topologically associated domain (TAD) boundaries as prior information to infer regulatory circuits comprising chromatin regulatory sites, modulatory TFs, and downstream target genes for each cell type (see **Methods** for details). In brief, to build candidate disease-modulated circuits, differentially accessible sites (DAS) within each cell type are first associated with TFs by motif sequence matching and then linked to differentially expressed genes (DEG) in that cell type by genomic localization within the same TAD. Next, MAGICAL uses a Bayesian framework to iteratively model chromatin accessibility and gene expression variation across cells and samples in each cell type and to estimate the confidence of TF-peak and peak-gene linkages for each candidate circuit (**Fig.1b**).

To accurately identify varying circuits between different conditions, the MAGICAL framework introduces variables to explicitly model signal and noise in chromatin accessibility and gene expression data in each cell type (**Fig.1b;** see **Methods** for details). The circuit TF-peak binding variable and a hidden TF activity variable are jointly estimated to fit to the chromatin accessibility variation across cells from the conditions being compared. These two variables are then used together with the circuit peak-gene

looping variable to fit the gene expression variation. Using Gibbs sampling, MAGICAL iteratively optimizes the states of TF-peak-gene linkages by controlling the variances of fitting residuals (data noise) in both modalities. Finally, high-confidence circuits fitting the signal variation in both data types are selected.

TF activity represents the regulatory capacity of a particular TF protein[16,17], which is distinct from TF expression. Since this type of activity (protein level) is not measured by either assay when the multiomics data are generated, we assume that, in each cell type, the hidden TF activities follow the same distribution in cells from the same sample, regardless of if the cells are from the scATAC-seq assay or the scRNA-seq assay or both. MAGICAL learns the activity distribution for each TF, samples a TF activity for each cell, and optimizes circuit variables against data noise in both modalities (**Fig. S1**). This procedure eliminates the limitation of existing correlation-based approaches that require a one-to-one, cell-level pairing of RNA-seq and ATAC-seq data. It makes MAGICAL a general tool that can analyze single-cell true multiome or sample-paired multiomics datasets.

We validated MAGICAL in multiple ways, demonstrating that it infers regulatory circuits accurately. We evaluated MAGICAL-inferred linkages between chromatin sites and genes using experimental 3D chromatin interactions. The resulting circuit genes, peaks and their regulatory TFs were respectively evaluated in multiple independent studies. And finally, as one example of utility, we used circuit genes as features to classify disease states, to provide a rapid solution to challenging diagnostic problems (**Fig. 1c**).

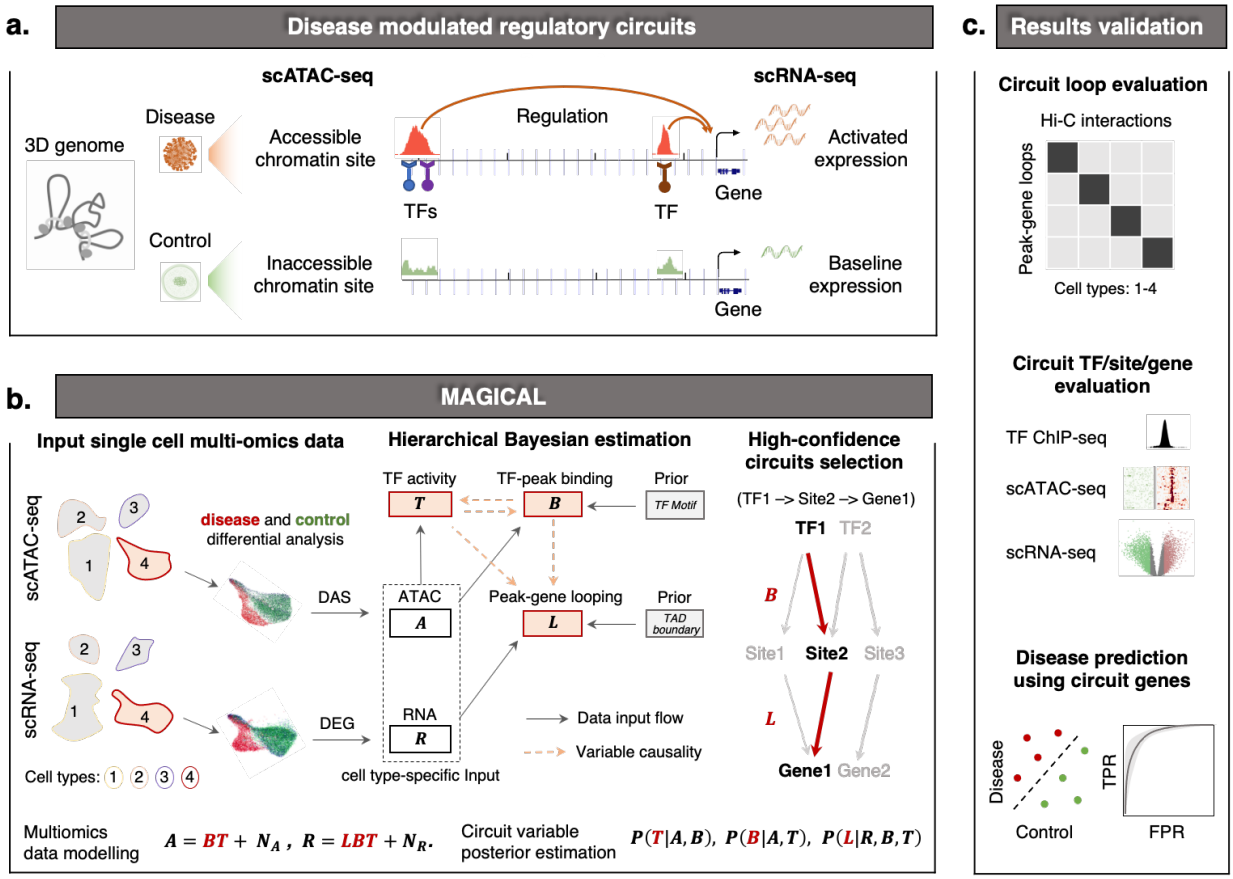

Figure 1: **Overview of MAGICAL for mapping disease-associated regulatory circuits from scRNA-seq and scATAC-seq data**. (**a**) Altered gene expression in cells between disease and control conditions can be attributed to chromatin accessibility changes of proximal and distal chromatin sites regulated by TFs. (**b**) To identify disease-associated regulatory circuits in a selected cell type (including ATAC assay cells and RNA assay cells from samples being compared), MAGICAL selects DAS as candidate regions and DEG as candidate genes. Then, the filtered ATAC data and RNA data of DAS and DEG are used as input to a hierarchical Bayesian framework pre-embedded with the prior TF motifs and TAD boundaries. The chromatin activity $\boldsymbol{A}$ is modelled as a linear combination of TF-peak binding confidence $\boldsymbol{B}$ and the hidden TF activity $\boldsymbol{T}$, with contamination of data noise $\boldsymbol{N_A}$. The gene expression $\boldsymbol{R}$ is modelled as a linear combination of $\boldsymbol{B}$, $\boldsymbol{T}$, and peak-gene looping confidence $\boldsymbol{L}$, with contamination of data noise $\boldsymbol{N_R}$. MAGICAL estimates the posterior probabilities $\boldsymbol{P}(\boldsymbol{B}|\boldsymbol{A},\boldsymbol{T})$, $\boldsymbol{P}(\boldsymbol{T}|\boldsymbol{A},\boldsymbol{B})$ and $\boldsymbol{P}(\boldsymbol{L}|\boldsymbol{R},\boldsymbol{B},\boldsymbol{T})$ by iteratively sampling variables $\boldsymbol{B}$, $\boldsymbol{T}$, and $\boldsymbol{L}$ to optimize against the data noise $\boldsymbol{N_A}$ and $\boldsymbol{N_R}$ in both modalities. Finally, regulatory circuits with high posterior probabilities of $\boldsymbol{B}$ and $\boldsymbol{L}$ (e.g., a high confidence circuit with inferred interactions between TF1, Site2 and Gene1) are selected. (**c**) We evaluate the accuracy and cell type specificity of the inferred peak-gene looping interactions by checking their enrichment with cell-type matched chromatin interactions in Hi-C experiments. For the identified TFs, peaks, and genes in circuits, we checked the accuracy of each using independent ChIP-seq, scATAC-seq, and scRNA-seq data. Finally, as a demonstration of the utility of MAGICAL, we used the circuit target genes as features to predict disease states.

Comparative analysis of performance

MAGICAL is a scalable framework. It can infer regulatory circuits of TFs, chromatin regions, and genes with differential activities between contrast conditions or infer regulatory circuits with active chromatin regions and genes in a single condition. Because existing integrative methods[11,12,18] can only be applied to single-condition data, to provide a comparative assessment of the performance of MAGICAL, we restricted MAGICAL to the single-condition data analysis possible with existing methods.

For peak-gene looping inference, we compared MAGICAL to the TRIPOD[11] and FigR[18] methods, using the same benchmark single-cell multiome datasets as used by the authors reporting these methods. In the comparison of MAGICAL with TRIPOD using a 10X multiome single-cell dataset [https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k], MAGICAL-inferred peak-gene loops showed significantly higher enrichment of experimentally observed chromatin interactions in blood cells in the 4DGenome database[19] (p-value<0.0001, **Fig. S2a**), the same validation data used by TRIPOD developers. MAGICAL also significantly outperformed FigR on the application to a GM12878 SHARE-seq dataset[10]. In that case, the peak-gene loops in MAGICAL-selected circuits had significantly higher enrichment of H3K27ac HiChIP chromatin interactions[20] than did FigR (p-value<0.0001, **Fig. S2b**).

Because the MAGICAL framework, unlike TRIPOD and FigR, used chromatin TAD boundaries as prior information, we evaluated whether the improvement in performance resulted solely from this additional information. To investigate this, we eliminated the use of TAD boundaries and modified MAGICAL for this test by assigning candidate linkages between peaks and genes within 500Kb (a naïve distance prior). As shown in **Fig.S2a** and **S2b,** even without the TAD prior information, MAGICAL still outperformed the competing methods (p-value <0.001). Overall, these results suggest that in addition to the benefit of priors, explicit modeling of signal and noise in both chromatin

accessibility and gene expression data increased the accuracy of peak-gene looping identification.

MAGICAL analysis of COVID-19 single-cell multiomics data

To demonstrate the accuracy of the primary application of MAGICAL on contrast condition data to infer disease-modulated circuits, we applied MAGICAL to sample-paired peripheral blood mononuclear cell (PBMC) scRNA-seq and scATAC-seq data from SARS-CoV-2 infected individuals and healthy controls[5]. Because immune responses in COVID-19 patients differ according to disease severity[21,22], MAGICAL inferred the regulatory circuits for mild and severe clinical groups separately. The chromatin sites and genes in the identified circuits were validated using newly generated and publicly available independent COVID-19 single-cell datasets (**Fig.2a**). We primarily focused on three cell types that have been found to show widespread gene expression and chromatin accessibility changes in response to SARS-CoV-2 infection[23,24], including CD8 effector memory T (TEM) cells, CD14 monocytes (Mono), and natural killer (NK) cells. In total, MAGICAL identified 1,489 high confidence circuits (1,404 sites and 391 genes) in these cell types for mild and severe clinical groups (**Table S1).**

To confirm the circuit chromatin sites selected by MAGICAL for mild COVID-19, we generated an independent PBMC scATAC-seq dataset from six SARS-CoV-2-infected subjects with mild symptoms and three uninfected (PCR-negative) controls (**Fig. 2b**; **Table S2**). About 25,000 quality cells were selected after QC analysis. These cells were integrated, clustered and annotated using ArchR[25] **(Fig. S3; Table S3).** Peaks were called from each cell type using MACS2[26]. In total, 284,909 peaks were identified (**Table S4**). For the three selected cell types, differential analysis between COVID-19 and control returned 3,061 sites for CD8 TEM, 1,301 sites for CD14 Mono, and 1,778 sites for NK (**Table S5;** see **Methods**). This produced three validation peak sets for mild COVID-19 infection. For severe COVID-19, an existing study focused on T cells identified specific chromatin activity changes with severe COVID-19 in CD8 T cells[27].

We used their reported chromatin sites for validating the circuit chromatin sites identified in CD8 T cells. In all four validation sets, the precision (proportion of sites that are differential in the validation data) of the MAGICAL-selected chromatin sites is significantly higher than the original DAS (*p*-value <0.001, **Fig. 2c** and **2d;** see **Methods).**

When multiple potential chromatin regulatory loci are identified in the vicinity of a specific gene, it is commonly assumed that the locus closest to the TSS is likely to be the most important regulatory site. Challenging this assumption, however, are the results of experimental studies showing that genes may not be regulated by the nearest region[28,29]. Supporting the importance of more distal regulatory loci, MAGICAL-selected chromatin sites significantly outperformed the nearest DAS to the TSS of DEG or all DAS within the same TAD with DEG, and the improvement is substantial (precision is ~50% better with MAGICAL, *p*-value<0.05, **Fig. 2c** and **2d**).

To validate the circuit genes modulated by mild or severe COVID-19, we used genes reported by external COVID-19 single-cell studies[21,30,31]. In total, we collected six validation gene sets (three cell types for mild COVID-19 and three cell types for severe COVID-19). The precision of MAGICAL-selected circuit genes is significantly higher than that of original DEG in all validations (precision is ~30% better with MAGICAL, *p*-value<0.05, **Fig. 2e** and **2f**). These results confirmed the increased accuracy of disease association for both chromatin sites and genes in MAGICAL-identified regulatory circuits.

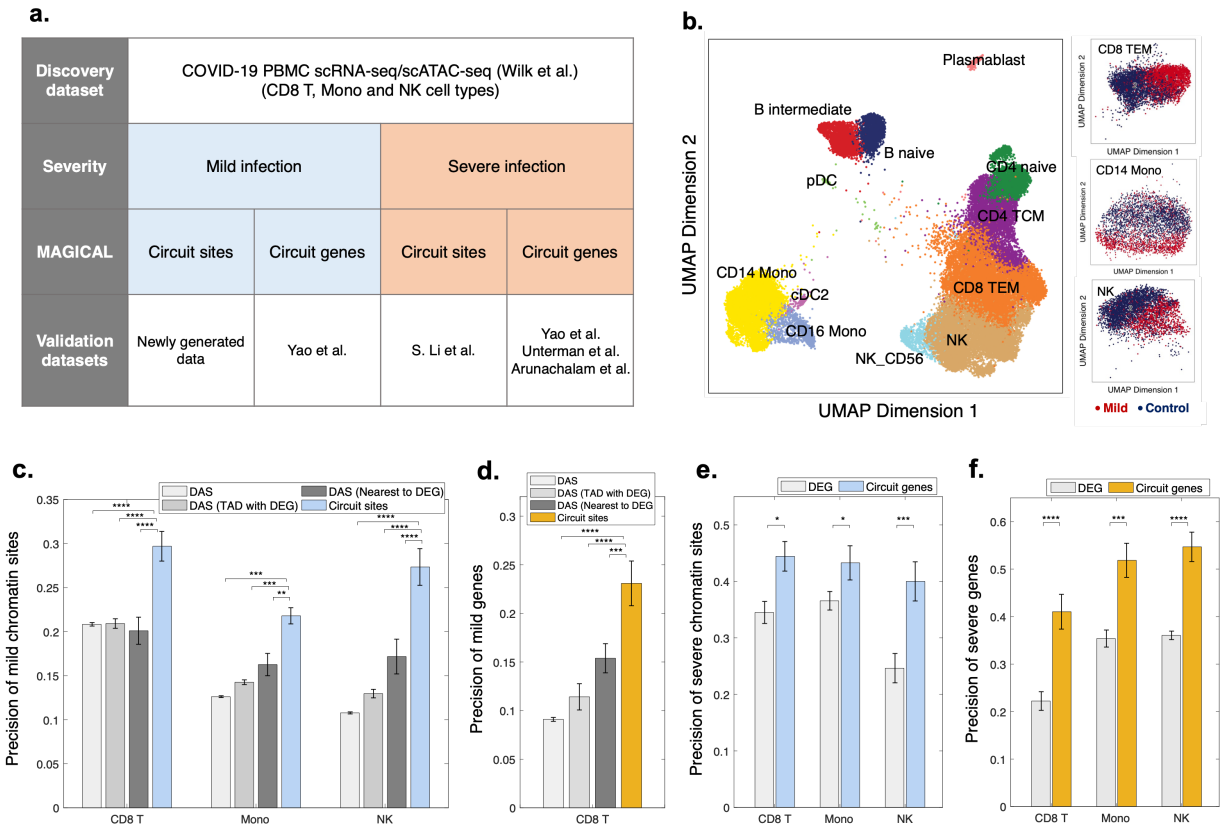

Figure 2. Validation of COVID-19-associated circuit chromatin sites and genes. (**a**) We applied MAGICAL to a COVID-19 PBMC single-cell multiomics dataset and identified circuits for the clinical mild and severe groups, respectively. We validated the circuit-associated chromatin sites and genes using newly generated and independent COVID-19 single-cell datasets. (**b**) UMAPs of a newly generated scATAC-seq dataset including 16K cells from six COVID-19 subjects and 9K cells from three showed chromatin accessibility changes in CD8 TEM, CD14 Mono, and NK cell types. (**c-d**) The precision of MAGICAL-selected circuit sites is significantly higher than the that of the original DAS, the nearest DAS to DEG or all DAS in the same TAD with DEG. (**e-f**) The precision of circuit sites are significantly higher than the that of DEG. (**c**)(**e**) for mild COVID-19 and (**d**)(**f**) for severe COVID-19. The error bars represent the standard error. Precision is defined as the proportion of sites/genes to be differentially accessible and differentially expressed in the same cell type between infection and control conditions. *p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, **** p-value < 0.0001. The error bars represent the standard error.

MAGICAL analysis of *S. aureus* sepsis single-cell multiomics data

We applied MAGICAL to the clinically important challenge of distinguishing methicillin-resistant (MRSA) and methicillin-susceptible *S. aureus* (MSSA) infections[32-34]. We

profiled paired scRNA-seq and scATAC-seq data using human PBMCs from adults who were blood culture positive for *S. aureus*, including 10 MRSA and 11 MSSA, and from 23 uninfected control subjects (**Fig. 3a; Table S6**). To integrate scRNA-seq data from all samples, we implemented a Seurat[35]-based batch correction and cell type annotation pipeline (see **Methods**). In total, 276,200 quality cells were selected and labeled (**Fig. 3b; Fig. S4; Table S7**). For scATAC-seq data, we integrated the fragment files from quality samples using ArchR[25] and selected and annotated 70,174 quality cells (**Fig**. **3c; Fig. S5** and **S6; Table S8**). In total, 388,860 peaks were identified (**Fig.S6b**; **Table S9; see Methods**). 13 major cell types that surpassed the 200-cell threshold in both scRNA-seq and scATAC-seq data were selected for subsequent analysis. Differential analysis for three contrasts (MRSA vs Control, MSSA vs Control, and MRSA vs MSSA) in each cell type returned a total of 1,477 DEG and 23,434 DAS (**Fig. S7** and **S8; Table S10** and **S11;** see **Methods**).

MAGICAL identified 1,513 high-confidence regulatory circuits (1,179 sites and 371 genes) within cell types for three contrasts (MRSA vs Control, MSSA vs Control, and MRSA vs MSSA) (**Table S4;** see **Methods**). It has been reported that activation of CD14 monocytes plays a principal role in response to *S. aureus* infection[36,37]. In MAGICAL analysis, CD14 monocytes showed the highest number of regulatory circuits (**Fig.3d**). Comparing circuits between cell types we found that these disease-associated circuits are cell type-specific (**Fig.3e**). For example, circuits rarely overlapped between very distinct cell types like monocytes and T cells. Between CD14 mono and CD16 mono, or between subtypes of T cells, most circuits are still specific for one cell type. These circuits were further validated using cell type-specific chromatin interactions reported in a reference promoter capture (pc) Hi-C dataset[13]. In all the cell types for which the cell type-specific pcHi-C data was available (B cells, CD4 T cells, CD8 T cells, CD14 monocytes), the circuit peak-gene interactions showed significant enrichment of pcHi-C interactions in matched cell types (**Fig.3f**; *p*-value < 0.01; see **Methods**). For comparison, we also performed the peak-gene interaction enrichment analysis between different cell types, finding significantly lower enrichment levels. These results indicate cell-type specificity of MAGICAL-identified circuits.

In CD14 monocytes, MAGICAL identified AP-1 complex proteins as the most important regulators, especially at chromatin sites showing increased activity in infection cells (**Fig.3g**). This finding is consistent with the importance of these complexes in gene regulation in response to a variety of infections[5,38,39]. Supporting the accuracy of the identified TFs, we compared circuit chromatin sites with ChIP-seq peaks from the Cistrome database[40]. The most similar TF ChIP-seq profiles were from AP-1 complex JUN/FOS proteins in blood or bone marrow samples (**Fig.S9**). Moreover, functional enrichment analysis[41] of the circuit genes showed that cytokine signaling, a known pathway mediated by AP-1 factors and associated with the inflammatory responses in macrophages[42,43], was the most enriched (adjusted *p*-value 2.4e-11).

MAGICAL modeled regulatory effects of both proximal and distal regions on genes. We examined the chromatin site location relative to the target gene TSS, for circuits chromatin sites and genes identified for CD14 monocytes. Compared to all ATAC peaks called around the circuit genes, an increased proportion of circuit chromatin sites were located 15Kb to 25Kb away from the TSS (**Fig.3h**). This pattern is consistent with the 24Kb median enhancer distance found by CRISPR-based perturbation in a blood cell line[44]. In addition, nearly 50% of circuit chromatin sites were overlapping with enhancer-like regions in the ENCODE database[45], further emphasizing that MAGICAL circuits are enriched in distal regulatory loci. We also found that these circuit chromatin sites were significantly enriched in inflammatory-associated GWAS loci reported in the GWAS catalog database[46], suggesting active host epigenetic responses to infectious diseases (**Fig.S10**; *p*-value < 0.005; see **Methods**). Notably, one distal chromatin site (hg38 chr6: 32,484,007-32,484,507) looping to HLA-DRB1 is within the most significant GWAS region (hg38 chr6: 32,431,410-32,576,834) associated with *S. aureus* infection[47].

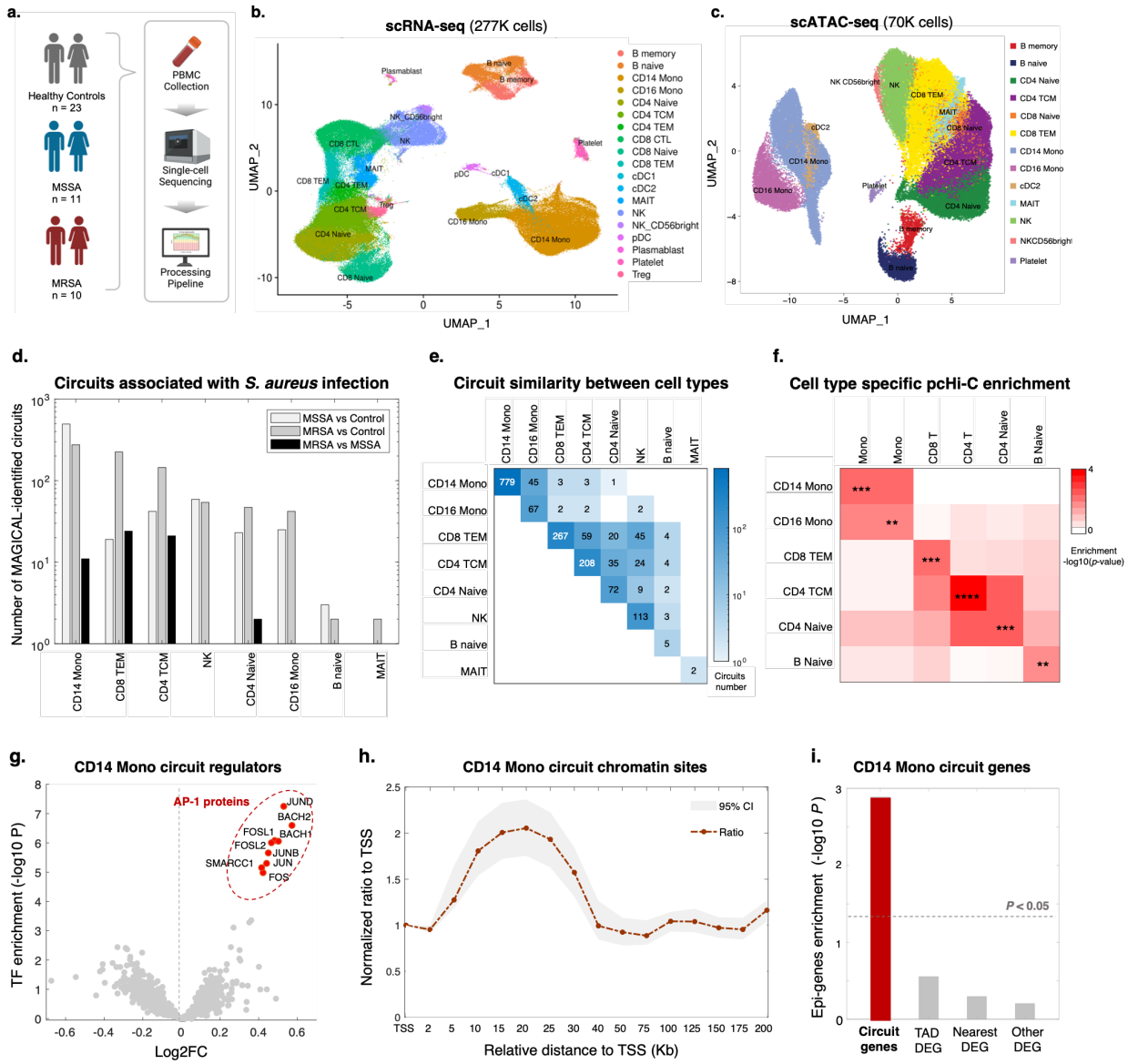

Figure 3. MAGICAL accurately identified distal regulatory chromatin sites and epi-driven genes associated with *S. aureus* infection. (**a**) We collected PBMC samples from 10 MRSA-infected, 11 MSSA-infected, and 23 healthy control subjects and generated same-sample scRNA-seq and scATAC-seq data using separate assays. (**b**) UMAP of integrated scRNA-seq data with 18 PBMC cell subtypes. (**c**) UMAP of integrated scATAC-seq data with 13 PBMC cell subtypes. Under-represented subtypes including cDC1, CD4 TEM, CD8 CTL, pDC, and Plasmablast, altogether representing less than 5% of cells in the scRNA-seq data, were not recovered from the scATAC-seq data. (**d**) The number of MAGICAL-identified regulatory circuits for each cell type and in contrast analysis. (**e**) The number of shared and specific circuits between cell types. (**f**) Enrichment of circuit peak-gene interactions in each cell type with cell type-specific pcHi-C interactions (***p*-value< 0.01, ****p-value* < 0.001, *****p-value*< 0.0001). (**g-i**) We specifically analyzed MAGICAL-identified regulatory circuits for CD14 monocytes. (**f**) TF motif enrichment analysis in circuit sites showed that AP-1 proteins are mostly significantly

enriched at chromatin regions with increased accessibility in the infection condition. The log2FC is calculated for each TF by dividing the number of binding sites with increased chromatin activity in the infection condition by the number of sites with decreased activity. (**g**) In comparison to all accessible chromatin sites, an increased proportion of MAGICAL circuit chromatin sites were in the range of 15Kb to 25Kb relative to gene TSS (the gray area represents the 95% confidence interval); and (**i**) the circuit genes were significantly enriched with experimentally confirmed epi-genes in monocytes.

We finally compared circuit genes to existing epi-genes whose transcriptions were significantly driven by epigenetic perturbations in CD14 monocytes[48]. MAGICAL-identified circuit genes were significantly enriched with epi-genes (p-value < 0.005, **Fig.3i**) while the remaining DEG not selected by MAGICAL, or those mappable with DAS either within the same topological domains or closest to each other showed no evidence of being epigenetically driven. These results suggest that MAGICAL accurately identified regulatory circuits activated in response to *S. aureus* infection.

S. aureus infection prediction

Early diagnosis of *S. aureus* infection and the strain antibiotic sensitivity is critical to appropriate treatment for this life-threatening condition. We first evaluated whether the MAGIC-identified circuit genes that are in common to MRSA and MSSA could provide a robust signature for predicting the diagnosis of *S. aureus* infection in general. Within each cell type, we selected circuit genes common to both the MRSA and MSSA analyses, resulting in 152 genes (**Fig.4a; Table S12**). To evaluate this *S. aureus* infection, we collected external, public expression data of *S. aureus* infection from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). In total, we found one adult whole-blood[49] and two pediatric PBMC bulk microarray datasets[50,51] that comprised a total of 126 *S. aureus* infected subjects and 68 uninfected controls. The use of pediatric validation data has the advantage of providing a much more rigorous test of the robustness of MAGICAL-identified circuit genes for classifying disease samples in this very different cohort.

To allow validation using public bulk transcriptome datasets, we refined the 152 circuit genes set by selecting those with robust performance in our dataset at pseudobulk level. We calculated an AUROC for each circuit gene by classifying *S. aureus* infection and control subjects using pseudobulk gene expression (aggregated from the discovery scRNA-seq data). 117 circuit genes with AUROCs greater than 0.7 were selected (**Table S13; Fig. S11a**). Functional gene enrichment analysis showed that IL-17 signaling was significantly enriched (*p*-value 2.4e-4), including genes from AP-1, Hsp90, and S100 families. IL-17 had been found to be essential for the host defense against cutaneous *S. aureus* infection in mouse models[52]. We trained an SVM model using the selected circuit genes as features and the discovery pseudobulk gene expression data as input. We then applied the trained SVM model to each of the three validation datasets. The model achieved high prediction performance on all datasets, showing AUROCs from 0.93 to 0.98 (**Fig.4a**).

This generalizability of circuit genes for predicting infection in different cohorts suggested that MAGICAL identifies regulatory processes that are fundamental to the host response to *S. aureus* sepsis. We further evaluated this by comparing the 117 circuit genes to 366 filtered DEG (with per gene AUROC>0.7 in the discovery pseudobulk gene expression data). We examined the differential expression π-value[53] (a statistic score that combines both fold change and p-values) of genes in the validation datasets and found significantly higher π-values for the circuit genes (**Fig. S11b**, *p*-value 9.0e-3).

S. aureus antibiotic sensitivity prediction

We then addressed the challenging problem of predicting strain antibiotic sensitivity in *S. aureus* infection. When we tested the predictive models trained with DEG for the contrast of MRSA and MSSA on three pediatric PBMC microarray datasets[50,51,54] (comprising a total of 66 MRSA and 45 MSSA samples), we did not find predictive value (median of prediction AUCs close to 0.5) (**Fig.S12b-d;** see **Methods**). And in all tests, the statistical difference between DEG-based prediction scores of the MRSA and MSSA

samples in the validation datasets was never significant. These results suggest that using host scRNA-seq data alone fails to identify robust features for predicting the antibiotic sensitivity of the infected strain. These echo previous studies showing that in challenging cases, differential expression analysis using RNA-seq data had limited power to identify robust features for disease-control sample classification[55].

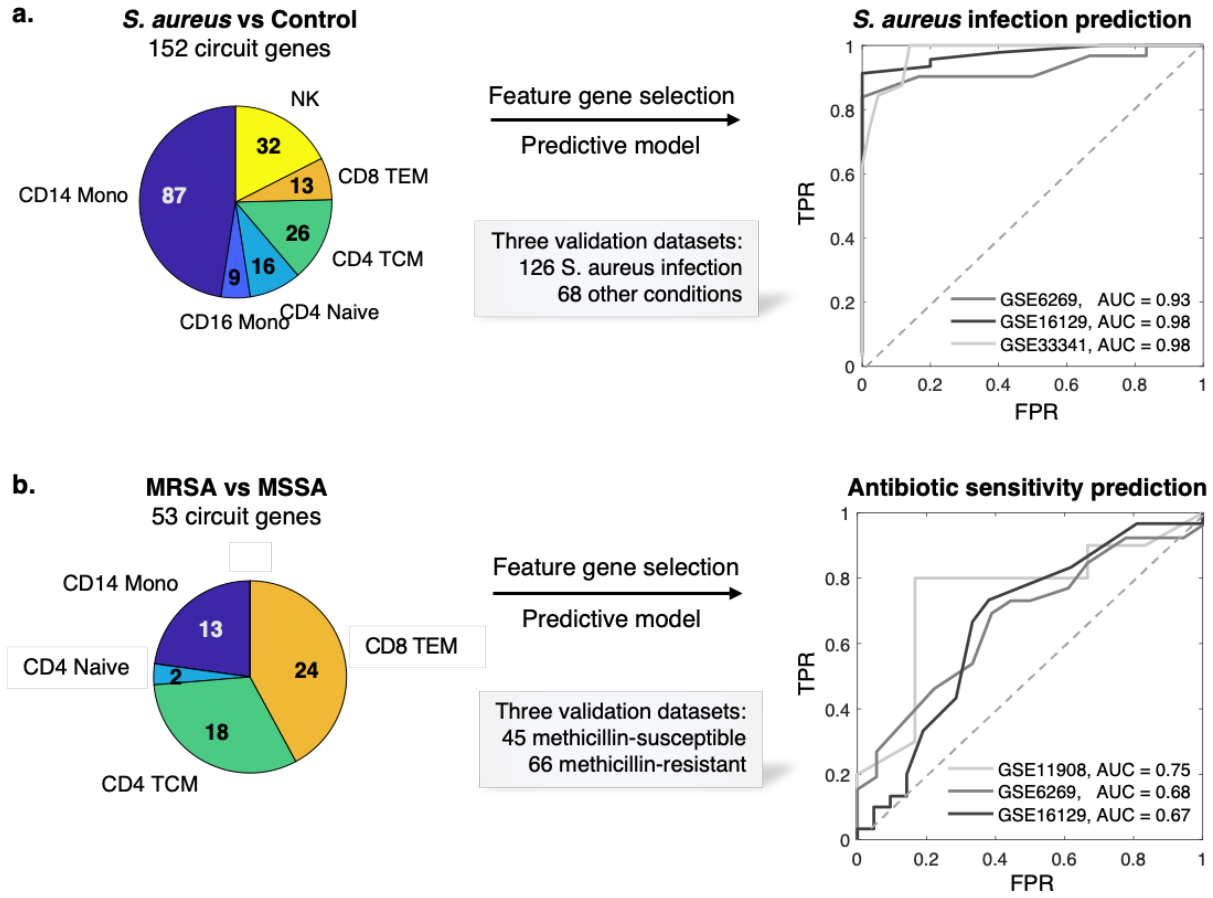

Figure 4: MAGICAL-identified circuit genes robustly predict *S. aureus* infection and bacteria antibody sensitivity. (**a**) Circuit genes in common to MRSA and MSSA infections achieved a near-perfect classification of *S. aureus* infected and uninfected samples in multiple independent datasets (one adult dataset and two pediatric datasets). (**b**) Circuit genes that differed between MRSA and MSSA showed predictive value of antibiotic sensitivity in independent patient samples (three pediatric datasets).

With MAGICAL we identified 53 circuit genes from the comparative multiomics data analysis between MRSA and MSSA (**Table S14**). A new model trained using 32 circuit genes that were robustly differential in the discovery pseudobulk data (per gene discovery AUROC > 0.7, **Fig. S12a**) best distinguished antibiotic-resistant and

antibiotic-sensitive samples in all three validation datasets, with AUROCs from 0.67 to 0.75 (**Fig.4b**). And the statistical difference between prediction scores of MRSA and MSSA samples was significant (p-value = 9.2e-3). The success of the circuit gene-based model demonstrated that MAGICAL captured generalizable regulatory differences in the host immune response to these closely related bacterial infections.

Discussion

MAGICAL addressed the previously unmet need of identifying differential regulatory circuits based on single cell multiomics data from different conditions. Critically, it identifies regulatory circuits involving distal chromatin sites. The previously difficult-to-predict distal regulatory regions are increasingly recognized as key for understanding gene regulatory mechanisms. MAGICAL provides a robust framework to infer regulatory circuits contributing to disease states.

Using newly generated PBMC single cell multiomics data, MAGICAL identified the regulatory circuits characteristic of the human immune response to *S. aureus* infection. Our thorough analysis of MAGICAL circuits and independent data validation demonstrate the ability of MAGICAL to use epigenetic context to identify robust immune cell responses specific to these infections. Despite profound differences between the discovery and the validation cohorts (adult vs. pediatric) and technologies and data resolution (single-cell sequencing vs. bulk microarray), MAGICAL-identified circuit genes achieved high prediction performance on the latter.

MAGICAL is a general tool for single-cell multiomics data analysis and provides a better lens to identify disease-associated regulation within cell types. We have applied MAGICAL to additional datasets (data not shown) for diagnosis gene signature development. For example, we generated new PBMC sample-matched single-cell multiomics data for HIV and influenza (flu) infection and used MAGICAL was to identify circuits genes for each by contrasting against the control condition. These circuit genes achieved high prediction accuracy on independent validation datasets (AUC 0.95 for

HIV and AUC 0.85 for Influenza). We also applied MAGICAL to true multiome data collected from skeletal muscle before and after acute endurance exercise. In this application, we resolved new regulators controlling responses to exercise that differed among muscle fiber types. These diverse applications demonstrate the broad utility of MAGICAL.

MAGICAL uses DAS and DEG called from a pre-selected cell type. For less distinct cell types or conditions, it is harder for MAGICAL to infer circuits at cell type resolution as there will be few candidate peaks and genes. Also, MAGICAL analyzes each cell type separately, and cell type specificity is not directly modeled for disease circuit identification. Incorporating an approach to directly identify cell type-specific circuits regulated in disease conditions would be valuable. In future work, we hope to extend the MAGICAL framework to improve circuit identification when cell types are poorly defined and to model cell type specificity.

Methods

S. aureus patient and control samples selection

Patients with culture-confirmed *S. aureus* bloodstream infection transferred to DUMC are eligible if pathogen speciation and antibiotic susceptibilities are confirmed by the Duke Clinical Microbiology Laboratory. Patient or patient's legally authorized representative provides signed informed consent allowing participation unless patient expires prior to notification of blood culture results. DNA and RNA samples, PBMCs, clinical data, and the bacterial isolate from the subject are cataloged using an IRB-approved Notification of Decedent Research. We excluded samples if prior enrollment of the patient in this investigation (to ensure statistical independence of observations) or they are polymicrobial (i.e., more than one organism in blood or urine culture). In total, 21 adult patients were selected with 10 MRSAs and 11 MSSAs. None of them received any antibiotics in the 24 h before the bloodstream infection. Control samples were obtained from uninfected healthy adults matching the sample number and age range of the patient group. In total, 23 samples were collected from two cohorts: 14 controls provided by from the Weill Cornell Medicine, New York, NY, and 9 controls (provided by the Battelle Memorial Institute, Columbus, OH. Meta information of the selected subjects were provided in **Table S6**.

PBMC thawing

Frozen PBMC vials were thawed in a 37°C-waterbath for 1 to 2 minutes and placed on ice. 500µl of RPMI/20% FBS was added dropwise to the thawed vial, the content was aspirated and added dropwise to 9ml of RPMI/20% FBS. The tube was gently inverted to mix, before being centrifuged at 300xg for 5 min. After removal of the supernatant, the pellet was resuspended in 1-5ml of RPMI/10% FBS depending on the size of the pellet. Cell count and viability were assessed with Trypan Blue on a Countess II cell counter (Invitrogen).

S. aureus scRNA-seq data generation

ScRNA-seq was performed as described (10x Genomics, Pleasanton, CA), following the Single Cell 3' Reagents Kits V3.1 User Guidelines. Cells were filtered, counted on a Countess instrument, and resuspended at a concentration of 1,000 cells/µl. The number of cells loaded on the chip was determined based on the 10X Genomics protocol. The 10X chip (Chromium Single Cell 3' Chip kit G PN-200177) was loaded to target 5,000-10,000 cells final. Reverse transcription was performed in the emulsion and cDNA was amplified following the Chromium protocol. Quality control and quantification of the amplified cDNA were assessed on a Bioanalyzer (High-Sensitivity DNA Bioanalyzer kit) and the library was constructed. Each library was tagged with a different index for multiplexing (Chromium i7 Multiplex Single Index Plate T Set A, PN-2000240) and quality controlled by Bioanalyzer prior to sequencing.

S. aureus scRNA-seq data integration and cell type annotation

Reads of scRNA-seq experiments were aligned to human reference genome (hg38) using Cell Ranger v3.1.0 and. The filtered feature-by-barcode count matrices were then processed using Seurat[35]. Quality cells were selected as those with more than 400 features (transcripts), fewer than 5,000 features, and less than 10% of mitochondrial content (**Fig. S4**; **Table S7**). Cell cycle phase scores were calculated using the canonical markers for G2M and S phases embedded in the Seurat package. Finally, the effects of mitochondrial reads and cell cycle heterogeneity were regressed out using SCTransform. To integrate cells from heterogeneous disease samples, we first built a reference by integrating and annotating cells from the uninfected control samples using a Seurat-based pipeline. For batch correction, we identified the intrinsic batch variants and used Seurat to integrate cells together with the inferred batch labels. All control samples were integrated into one harmonized query matrix. Each cell was assigned a cell type label by referring to the same reference PBMC single cell dataset used above. The cell type label of each cell cluster was determined by most cell labels in each. Canonical markers were used to refine the cell type label assignment. This integrated control object was used as reference to map the infected samples. To avoid artificially removing the biological variance between each infected sample during batch correction,

we computationally predicted and manually refined cell types for each sample. All infection samples were projected onto the UMAP of the control object for visualization purpose. In total, 276,200 high-quality cells and 19 cell types with at least 200 cells in each were selected for the subsequent analysis.

Nuclei isolation for scATACseq

Thawed PBMCs were washed with PBS/0.04% BSA. Cells were counted and 100,000-1,000,000 cells were added to a 2mL-microcentrifuge tube. Cells were centrifuged at 300xg for 5min at 4°C. The supernatant carefully completely removed, and 0.1X lysis buffer (1x: 10mM Tris-HCl pH 7.5, 10mM NaCl, 3mM MgCl2, nuclease-free H2O, 0.1% v/v NP-40, 0.1% v/v Tween-20, 0.01% v/v digitonin) was added. After 3min incubation on ice, 1ml of chilled wash buffer was added. The nuclei were pelted at 500xg for 5min at 4°C and resuspended in a chilled diluted nuclei buffer (10X Genomics) for scATAC-seq. Nuclei were counted and the concentration was adjusted to run the assay.

S. aureus scATAC-seq data generation

ScATAC-seq was performed immediately after nuclei isolation and following the Chromium Single Cell ATAC Reagent Kits V1.1 User Guide (10x Genomics, Pleasanton, CA). Transposition was performed in 10μl at 37°C for 60min on at least 1,000 nuclei, before loading of the Chromium Chip H (PN-2000180). Barcoding was performed in the emulsion (12 cycles) following the Chromium protocol. After post GEM cleanup, libraries were prepared following the protocol and were indexed for multiplexing (Chromium i7 Sample Index N, Set A kit PN-3000427). Each library was assessed on a Bioanalyzer (High-Sensitivity DNA Bioanalyzer kit).

S. aureus scATAC-seq data integration and cell type annotation

Reads of scATAC-seq experiments were aligned to human reference genome (hg38) using 10x Genomics Cell Ranger software (version 1.2). The resulting fragment files were processed using ArchR[25]. Quality cells were selected as those with TSS enrichment > 12, the number of fragments >3000 and <30000, and nucleosome ratio <2 (**Figures S5** and **S6**; **Table S8**). The likelihood of doublet cells was computationally

assessed using ArchR's addDoubletScores function and cells were filtered using the ArchR's filterDoublets function with default settings. Cells passing quality and doublet filters from each sample were combined into a linear dimensionality reduction using ArchR's addIterativeLSI function with the input of the tile matrix (read counts in binned 500bps across the whole genome) with iterations = 2 and varFeatures = 20000. This dimensionality reduction was then corrected for batch effect using the Harmony method[56], via ArchR's addHarmony function. The cells were then clustered based on the batch-corrected dimensions using ArchR's addClusters function. We annotated scATAC-seq cells using ArchR's addGeneIntegrationMatrix function, referring to a labeled multimodal PBMC single cell dataset (https://atlas.fredhutch.org/nygc/multimodal-pbmc/). Doublet clusters containing a mixture of many cell types were manually identified and removed. In total, 70,174 high-quality cells and 13 cell types with at least 200 cells in each were selected. Peaks were called for each cell type using ArchR's addReproduciblePeakSet function with the MACS2 peak caller[26] (**Fig. S6b**). In total, 388,859 peaks were identified (**Table S9**).

S. aureus differential peak and gene calling

For each cell type, differentially accessible chromatin sites (DAS) between contrast conditions (MRSA vs Control, MSSA vs Control or MRSA vs MSSA) were called from the single cell chromatin accessibility count data using the "getMarkerFeatures" function of ArchR v1.0.2[25], with parameter settings as testMethod = "wilcoxon", bias = "log10(nFrags)", normBy = "ReadsInPeaks", and maxCells = 15000. Peaks with single cell differential statistics FDR < 0.05, |log2FC|>0.1, and actively accessible in at least 10% cells (pct>0.1) from either condition were selected as DAS. Due to the high false positive rate in single cell-based differential analysis[57], we further refined the DAS by fitting a linear model to the aggregated and normalized pseudobulk chromatin accessibility data and tested DAS individually about their covariance with sample conditions[58]. Refined DAS passing pseudobulk differential statistics *p*-value <0.05 and |log2FC|>0.3 between the contrast conditions were selected as the final DAS (**Table S11**).

For the same cell type, differentially expressed genes (DEG) between contrast conditions were first selected using single cell gene expression count data. We called differential genes using the "Findmarkers" function of the Seurat V4 package[35] with default parameters and selected DEG with Wilcoxon test FDR < 0.05, |log2FC|>0.1 and actively expressed in at least 10% cells (pct>0.1) from either condition were selected. To correct potential bias caused by the different sequencing depth between samples, we run DEseq2[58] on the aggregated pseudobulk gene expression data. Refined DEG passing pseudobulk differential statistics *p*-value <0.05 and |log2FC|>0.3 were selected as the final DEG (**Table S10**).

MAGICAL

To build candidate regulatory circuits, TFs were mapped to the selected DAS by searching for 870 human motifs from the chromVARmotifs library (https://github.com/GreenleafLab/chromVARmotifs) using ArchR's addMotifAnnotations function. The binding DAS were then linked with DEG by requiring them in the same TAD within boundaries. Then, a candidate circuit is constructed with a chromatin region and a gene in the same domain, with at least one TF motif match in the region.

For each cell type (i.e. i-th cell type), MAGICAL inferred the confidence of TF-peak binding and peak-gene looping in each candidate circuit using a hierarchical Bayesian framework with two models: a model of TF-peak binding confidence ($\boldsymbol{B}$) and hidden TF activity ($\boldsymbol{T}$) to fit chromatin accessibility ($\boldsymbol{A}$); a second model of peak-gene interaction ($\boldsymbol{L}$) and the refined (noise removed) regulatory region activity ($\boldsymbol{BT}$) to fit gene expression ($\boldsymbol{R}$).

$$\boldsymbol{A}_{P\times K_{A,S},i} = \boldsymbol{B}_{P\times M,i}\boldsymbol{T}_{M\times K_{A,S},i} + \boldsymbol{N}_{P\times K_{A,S},i}, \quad (1)$$

$$\boldsymbol{R}_{G\times K_{R,S},i} = \boldsymbol{L}_{G\times P,i}\boldsymbol{B}_{P\times M,i}\boldsymbol{T}_{M\times K_{R,S},i} + \boldsymbol{N}_{G\times K_{R,S},i}, \quad (2)$$

$\boldsymbol{A}_{P\times K_{A,S},i}$: a P by $K_{A,S}$ matrix with each element $a_{p,k_{A,s},i}$ representing the ATAC read count of p-th chromatin site (ATAC peak) in $k_{A,s}$-th cell in s-th sample.

$\boldsymbol{R}_{G\times K_{R,S},i}$: a G by $K_{R,S}$ matrix with each element $r_{g,k_{R,s},i}$ representing the RNA read count of g-th gene in $k_{R,s}$-th cell of s-th sample.

$\boldsymbol{N}_{P\times K_{A,S},i}$ and $\boldsymbol{N}_{G\times K_{R,S},i}$ represented data noise in corresponding to $\boldsymbol{A}_{P\times K_{A,S},i}$ and $\boldsymbol{R}_{G\times K_{R,S},i}$.

$\boldsymbol{B}_{P\times M,i}$: a P by M matrix with each element $b_{p,m,i}$ representing the binding confidence of m-th TF on p-th candidate chromatin site.

$\boldsymbol{L}_{G\times P,i}$: a G by P matrix with each element $l_{p,g,i}$ representing the interaction between p-th chromatin site and g-th gene.

$\boldsymbol{T}_{M\times K_{A,S},i}$: a M by $K_{A,S}$ matrix with each element $t_{m,k_{A,s},i}$ representing the hidden TF activity of m-th TF in $k_{A,s}$-th ATAC cell of s-th sample.

$\boldsymbol{T}_{M\times K_{R,S},i}$: a M by $K_{R,S}$ matrix with each element $t_{m,k_{R,s},i}$ representing the hidden TF activity of m-th TF in $k_{R,s}$-th RNA cell of s-th sample.

$\boldsymbol{T}_{M\times K_{A,S},i}$ and $\boldsymbol{T}_{M\times K_{R,S},i}$ were both extended from the same $\boldsymbol{T}_{M\times S,i}$ (with elements $t_{m,s,i}$) by assuming that in i-th cell type and s-th sample, m-th TF's regulatory activities in all ATAC cells and all RNA cells followed an identical distribution of a single variable $t_{m,s,i}$. Therefore, $K_{A,S,i}$ and $K_{R,S,i}$ can be different numbers and MAGICAL will only estimate the matrix $\boldsymbol{T}_{M\times S,i}$.

To select high-confidence regulatory circuits, MAGICAL estimated the confidence (probability) of TF-peak binding $\boldsymbol{B}_{P\times M,i}$ and peak-gene interaction $\boldsymbol{L}_{G\times P,i}$ together with the hidden variable $\boldsymbol{T}_{M\times S,i}$ in a Bayesian framework.

$$P(\boldsymbol{B},\boldsymbol{T},\boldsymbol{L}|\boldsymbol{A},\boldsymbol{R}) \propto P(\boldsymbol{R}|\boldsymbol{L},\boldsymbol{B},\boldsymbol{T})P(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{T})P(\boldsymbol{L})P(\boldsymbol{B})P(T). \quad (3)$$

Based on the regulatory relationship among chromatin sites, upstream TFs, and downstream genes (as illustrated in **Fig. 1**), the posterior probability of each variable can be approximated as:

$$P(\boldsymbol{T}|\boldsymbol{A},\boldsymbol{B}) \propto P(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{T})P(\boldsymbol{T}), \tag{4}$$

$$P(\boldsymbol{B}|\boldsymbol{A},\boldsymbol{T}) \propto P(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{T})P(\boldsymbol{B}), \tag{5}$$

$$P(\boldsymbol{L}|\boldsymbol{R},\boldsymbol{B},\boldsymbol{T}) \propto P(\boldsymbol{R}|\boldsymbol{L},\boldsymbol{B},\boldsymbol{T})P(\boldsymbol{L}). \tag{6}$$

Although the prior states of $b_{p,m,i}$ and $l_{p,g,i}$ were obtained from the prior information of TF motif-peak mapping and topological domain-based peak-gene pairing, their values were unknown. We assumed zero-mean Gaussian priors for $\boldsymbol{B}$, $\boldsymbol{L}$ and the hidden variable $\boldsymbol{T}$ by assuming that positive regulation and negative regulation would have the same priors, which is likely to be true given the fact that there were usually similar numbers of up-regulated and down-regulated peaks and genes after the differential analysis. We set a high variance (non-informative) in each prior distribution to allow the algorithm to learn the distributions from the input data.

$$b_{p,m,i} \sim normal(\mu_B, \sigma_B^2), \tag{7}$$

$$t_{m,s,i} \sim normal(\mu_T, \sigma_T^2), \tag{8}$$

$$l_{p,g,i} \sim normal(\mu_L, \sigma_L^2). \tag{9}$$

The likelihood functions $P(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{T})$ and $P(\boldsymbol{R}|\boldsymbol{L},\boldsymbol{B},\boldsymbol{T})$ represent the fitting performance of the estimated variables to the input data. These two conditional probabilities are equal to the probabilities of the fitting residues $\boldsymbol{N}_{P\times K_{A,S},i}$ and $\boldsymbol{N}_{G\times K_{R,S},i}$, for which we assumed zero-mean Gaussian distributions and used an Inverse Gamma distribution with hyperparameters α and β to control the variance of fitting residues (very low probabilities on large variances).

$$\boldsymbol{A}|\boldsymbol{B},\boldsymbol{T} \sim normal(\mu_{N_A}, \sigma_{N_A}^2),\ \sigma_{N_A}^2 \sim inversegamma\ (\alpha_{N_A}, \beta_{N_A}), \tag{10}$$

$$\boldsymbol{R}|\boldsymbol{L},\boldsymbol{B},\boldsymbol{T} \sim normal(\mu_{N_R}, \sigma_{N_R}^2),\ \sigma_{N_R}^2 \sim inversegamma\ (\alpha_{N_R}, \beta_{N_R}), \tag{11}$$

Then, the posterior probability of each variable defined in Eq. (4-6) was still a Gaussian distribution with poster mean $\hat{\mu}$ and variance $\hat{\sigma}$ as shown below:

$$\hat{b}_{p,m,i} \sim normal(\hat{\mu}_{B,m,i}, \hat{\sigma}^2_{B,m,i}), \tag{12}$$

$$\hat{t}_{m,s,i} \sim normal(\hat{\mu}_{T,m,s,i}, \hat{\sigma}^2_{T,m,s,i}), \tag{13}$$

$$\hat{l}_{p,g,i} \sim normal(\hat{\mu}_{L,i}, \hat{\sigma}^2_{L,i}). \tag{14}$$

Gibbs sampling was used to iteratively learn the posterior distribution mean and variance of each set of variables and draw samples of their values accordingly.

For the TF-peak binding events, the posterior mean $\hat{\mu}_{B,m,i}$ and variance $\hat{\sigma}^2_{B,m,i}$ were estimated specifically for m-th TF since the number of binding sites and the positive or negative regulatory effects between TFs could be very different.

$$\hat{\mu}_{B,m,i} = \frac{\sum_s \sum_k t_{m,s,i}(a_{p,ks,i} - \sum_{m'} b_{p,m',i} t_{m',s,i})\sigma_B^2 + \mu_{B,t}\sigma_{N_A}^2}{\sum_s K_{A,s} t^2_{m,s,i}\sigma_B^2 + \sigma_{N_A}^2} \text{ and } \hat{\sigma}^2_{B,m,i} = \frac{\sigma_{N_A}^2 \sigma_B^2}{\sum_s K_{A,s} t^2_{m,s,i}\sigma_B^2 + \sigma_{N_A}^2}. \tag{15}$$

For TF activities, the posterior mean $\hat{\mu}_{T,m,s,i}$ and variance $\hat{\sigma}^2_{T,m,s,i}$ were estimated specifically for m-th TF and s-th sample using chromatin accessibility data as follows:

$$\hat{\mu}_{T,m,s,i} = \frac{\sum_p \sum_k b_{p,m}(a_{p,k,s,i} - \sum_{m'} b_{p,m'} t_{m',s})\sigma_T^2 + \mu_T\sigma_{N_A}^2}{\sum_p K_{A,s} b^2_{p,m,i}\sigma_T^2 + \sigma_{N_A}^2} \text{ and } \hat{\sigma}^2_{T,m,s,i} = \frac{\sigma_{N_A}^2 \sigma_T^2}{\sum_p K_{A,s} b^2_{p,m,i}\sigma_T^2 + \sigma_{N_A}^2}. \tag{16}$$

Then, based on the estimated distribution parameters of $\hat{\mu}_{T,m,s,i}$ and $\hat{\sigma}^2_{T,m,s,i}$ of $\hat{t}_{m,s,i}$, for $k_{R,s}$-th RNA cell in the same *s*-th sample we draw a TF regulatory activity sample as $\hat{t}_{m,k_R,s,i}$. For p-th peak, we were able to reconstruct its chromatin activity in the RNA cell as $\hat{a}_{p,k_R,s,i} = \sum_m \hat{b}_{p,m,i}\hat{t}_{m,k_R,s,i}$, and for g-th gene, we further estimated the interaction confidence $\hat{l}_{p,g,i}$ between p-th peak and g-th gene. The peak-gene interaction distribution parameters $\hat{\mu}_{L,i}$ and $\hat{\sigma}^2_{L,i}$ were estimated as follows:

$$\hat{\mu}_{L,i} = \frac{\sum_s \sum_k \hat{a}_{p,k_R,s,i}\left(r_{g,k,s,i} - \sum_{p'} l_{g,p'} \hat{a}_{p',k_R,s,i}\right)\sigma_L^2 + \mu_L \sigma_{N_R}^2}{\sum_s \sum_{k_{R,s}} (\hat{a}_{p,k_R,s,i})^2 \sigma_L^2 + \sigma_{N_R}^2} \text{ and } \hat{\sigma}_L^2 = \frac{\sigma_{N_R}^2 \sigma_L^2}{\sum_s \sum_{k_{R,s}} (\hat{a}_{p,k_R,s,i})^2 \sigma_L^2 + \sigma_{N_R}^2}. \quad (17)$$

In *n*-th round of Gibbs estimation, after learning all distributions, we estimated the confidence of each linkage by linearly mapping the sampled values of $\hat{b}_{p,m,i}$ and $\hat{l}_{p,g,i}$ in the range of (-∞, ∞) to probabilities in (0,1) as follows:

$$P\left(state\left(b_{p,m,i}|n\right) = 1\right) = \frac{\exp\left\{\left(\hat{b}_{p,m,i} - \hat{\mu}_{B,m,i}\right)/2\hat{\sigma}_{B,m,i}^2\right\}}{\exp\left\{\left(\hat{b}_{p,m,i} - \hat{\mu}_{B,m,i}\right)/2\hat{\sigma}_{B,m,i}^2\right\} + \exp\left\{\left(0 - \hat{\mu}_{B,m,i}\right)/2\hat{\sigma}_{B,m,i}^2\right\}}. \quad (18)$$

$$P\left(state\left(l_{p,g,i}|n\right) = 1\right) = \frac{\exp\left\{\left(\hat{l}_{p,g,i} - \hat{\mu}_{L,i}\right)/2\hat{\sigma}_{L,i}^2\right\}}{\exp\left\{\left(\hat{l}_{p,g,i} - \hat{\mu}_{L,i}\right)/2\hat{\sigma}_{L,i}^2\right\} + \exp\left\{\left(0 - \hat{\mu}_{L,i}\right)/2\hat{\sigma}_{L,i}^2\right\}}. \quad (19)$$

Binary state samples were then drawn based on the confidence of each linkage and were then used to initiate the next round of estimations. After running a long sampling process (in total *N* rounds) and accumulating enough samples on the binary states of TF-peak bindings and peak-gene interactions, we calculated the sampling frequency of each linkage as a posterior probability.

$$\begin{cases} P\left(state\left(b_{p,m,i}\right) = 1\right) = \frac{\sum_n state\left(b_{p,m,i}|n\right)}{N} \\ P\left(state\left(l_{p,g,i}\right) = 1\right) = \frac{\sum_n state\left(l_{p,g,i}|n\right)}{N} \end{cases} \quad (20)$$

MAGICAL analysis of *S.aureus* PBMC single-cell multiomics data

For each cell type, given DAS and DEG of contrast conditions (MRSA vs Control, MSSA vs Control or MRSA vs MSSA), MAGICAL was first initialized by mapping prior TF motifs from the 'chromVARmotifs' library to DAS using ArchR's addMotifAnnotations. Because there is no PBMC cell type Hi-C data publicly available, we are using TAD boundaries from a lymphoblastoid cell line, GM12878, which was originally generated by EBV transformation of PBMCs[59]. The TAD boundary structure is closely conserved between the lymphoblastoid cell lines and primary PBMC[60] and between cell types[61,62].

We called TAD boundaries from a GM12878 cell line Hi-C profile[62] using TopDom[63]. About 6000 topological domains were identified. For each contrast, we built candidate circuits by pairing DAS with TF binding sites with DEG in the same domain. MAGICAL was run 10000 times to ensure that the sampling process converged to stable states. This process was repeated for all cell types and the top 10% high confidence circuit predictions were selected from each cell type for validation analysis.

MAGICAL analysis of COVID-19 PBMC single-cell multiomics data

As a proof of concept for contrast condition single cell multiomics data analysis, MAGICAL was applied to a public PBMC COVID-19 single-cell multiomics dataset[5] with samples collected from patients with different severity and heathy controls. Processed scRNA-seq data was downloaded from the COVID-19 Cell Atlas (https://www.covid19cellatlas.org/). For each of the three selected cell subtypes (CD8 TEM, CD14 Mono, and NK), from the original publication we downloaded DEG for two contrasts: mild vs control and severe vs control. As no processed scATAC-seq files were provided, the raw fragment files were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/, under accession number GSE174072) and quality cells are selected and further processed by following all steps as introduced in the paper. For each of the selected cell types, DAS were called respectively for mild vs control and severe vs control using ArchR's functions and thresholds as introduced in the paper. MAGICAL was initialized by mapping prior TF motifs from the 'chromVARmotifs' library to DAS using ArchR's addMotifAnnotations. As explained above, we used TAD boundary information of ~6000 domains identified in GM12878 cell line[62] as prior. Then, DAS with TF binding sites were paired with DEG in the same TAD and the initial candidate regulatory circuits were constructed. Respectively for mild and severe COVID-19, MAGICAL was run 10000 times to ensure that the sampling process converged to stable states. This process was repeated for all selected cell types. The chromatin sites and genes in the top 10% predicted high confidence circuits in each cell type were selected as disease associated (**Table S1**) and further validated on newly generated and external high quality single cell datasets.

New COVID-19 PBMC samples

To validate chromatin sites associated with mild COVID-19, PBMC samples were obtained from the COVID-19 Health Action Response for Marines (CHARM) cohort study, which has been previously described[64]. The cohort is composed of Marine recruits that arrived at Marine Corps Recruit Depot—Parris Island (MCRDPI) for basic training between May and November 2020, after undergoing two quarantine periods (first a home-quarantine, and next a supervised quarantine starting at enrolment in the CHARM study) to reduce the possibility of SARS-CoV-2 infection at arrival. Participants were regularly screened for SARS-CoV-2 infection during basic training by PCR, serum samples were obtained using serum separator tubes (SST) at all visits, and a follow-up symptom questionnaire was administered. At selected visits, blood was collected in BD Vacutainer CPT Tube with Sodium Heparin and PBMC were isolated following the manufacturer's recommendations. We used PBMC samples from six participants (five males and one female) who had a COVID-19 PCR positive test and had mild symptoms (sampled 3-11 days after the first PCR positive test), and from three control participants (three males) that had a PCR negative test at the time of sample collection and were seronegative for SARS-CoV-2 IgG. New scATAC-seq data were generated following the same protocol as described above (**Table S2**).

COVID-19 PBMC scATACseq data integration

Reads of scATAC-seq experiments were aligned to human reference genome (hg38) using 10x Genomics Cell Ranger software (version 1.2). The resulting fragment files were processed using ArchR[25]. Quality cells were selected as those with TSS enrichment > 12, the number of fragments >3000 and <30000, and nucleosome ratio <2. The likelihood of doublet cells was computationally assessed using ArchR's addDoubletScores function and cells were filtered using the ArchR's filterDoublets function with default settings. A total of 15,836 high quality cells in the infection group and 9,125 cells in the control group were selected after QC analysis (**Fig. S3**; **Table S3**). These cells were combined into a linear dimensionality reduction using ArchR's addIterativeLSI function with the input of the tile matrix (read counts in binned 500bps across the whole genome) with iterations = 2 and varFeatures = 20000. The cells were

then clustered using ArchR's addClusters function. We annotated scATAC-seq cells using ArchR's addGeneIntegrationMatrix function, referring to a labeled multimodal PBMC single cell dataset (https://atlas.fredhutch.org/nygc/multimodal-pbmc/). Doublet clusters containing a mixture of many cell types were manually identified and removed. Peaks were called for each cell type using ArchR's addReproduciblePeakSet function with peak caller MACS2[26] (**Fig. S3**). In total, 284,525 peaks were identified (**Table S4**).

COVID-19 PBMC scATAC-seq data differential peak calling

For each of the three selected cell types (CD8 TEM, CD14 Mono and NK), peaks were first called using ArchR's addReproduciblePeakSet function with MACS2 peak caller[26]. Then, differentially accessible chromatin sites between COVID-19 and control conditions were called from the single cell chromatin accessibility count data using ArchR's "getMarkerFeatures" function, with parameter settings as testMethod = "wilcoxon", bias = "log10(nFrags)", normBy = "ReadsInPeaks", and maxCells = 15000. Peaks with single cell differential statistics FDR < 0.05, |log2FC|>0.1, and actively accessible in at least 10% cells (pct>0.1) from either condition were selected. We further refined these differential peaks by fitting a linear model to the aggregated and normalized pseudobulk chromatin accessibility data[58]. Refined peaks passing pseudobulk differential statistics *p*-value <0.05 and |log2FC|>0.3 between the contrast conditions were finally selected as the validation peak set (**Table S5**).

COVID-19-associated circuit peaks and genes accuracy evaluation

The number of peaks/genes reported by each COVID-19 study would be different due to the difference in the number of recruited patients and collected cells. To overcome the issue caused by the imbalanced number between discovery and validation dataset or between differential peaks/genes and circuit sites/genes in comparison, in each comparison, the larger peak/gene set was randomly down sampled to match the smaller number of peaks/genes in the other set. The precision (site reproduction rate) is calculated to assess the accuracy of each peak/gene set. This calculation was repeated 10 times for each comparison and the median value was used for precision evaluation.

MAGICAL analysis of 10X PBMC single-cell true multiome data

For benchmarking, MAGICAL was applied to a 10X PBMC single cell multiome dataset (https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k), including 108377 ATAC peaks, 36601 genes, and 11909 cells from 14 cell types. MAGICAL used the same candidate peaks and genes as selected by TRIPOD[11] for fair performance comparison. Two different priors were used to pair candidate peaks and genes: (1) the peaks and genes were within the same TAD from the GM12878 cell line; (2) the centers of peaks and the TSS of genes were within 500K bps. MAGICAL inferred regulatory circuits with each prior and used the top 10% predictions for accuracy assessment. High confidence peak-gene interactions predicted by TRIPOD on the same data were directly downloaded from the supplementary tables of their publication[11]. Two baseline approaches of peak-gene pairing were included: pairing all peaks with each gene if they are in the same TAD or pairing only the nearest peak to gene based on their genomic distance. To fairly assess the accuracy of MAGICAL weighted peak-gene interactions and the results (paired or non-paired) from TRIPOD or baseline approaches, we selected the top 10% predictions by MAGICAL as the final peak-gene pairing. We overlapped these pairs with the curated 3D genome interactions in blood context from the 4DGenome database[19] and calculated the precision for each approach.

MAGICAL analysis of GM12878 cell line SHARE-seq data

For benchmarking, MAGICAL was also applied to a GM12878 cell line SHARE-seq dataset[10]. Processed data files were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/, under accession number GSE140203). For fair comparison, MAGICAL used the same candidate peaks and genes as selected by FigR[18]. MAGICAL was initialized with two different priors to pair candidate peaks and genes: (1) the peaks and genes were within the same prior TAD from the GM12878 cell line; (2) the centers of peaks and the TSS of genes were within 500k bps. MAGICAL inferred regulatory circuits under each setting and used the top 10% predictions for accuracy assessment. High confidence peak-gene interactions predicted by FigR were directly downloaded from the supplementary tables of the original publication[10].

Similarly, the top 10% predictions by MAGICAL and interactions paired by the two baseline approaches mentioned above were selected. We overlapped peak-gene interactions predicted by each approach with GM12878 H3K27ac HiChIP chromatin interactions[20] for precision evaluation.

Precision evaluation for predicted peak-gene interactions

To assess the precision of the predicted circuit peak-gene interactions, we assumed an corrected inferred peak-gene pair should be also connected by a chromatin interaction reported by Hi-C or similar experiments. To check this, each peak was extended to 2kb long and then checked for overlapping with one end of a physical chromatin interaction. For genes, we checked if the gene promoter (-2kb to 500b of TSS) overlapped the other end of the interaction. Precision was calculated as the proportion of overlapped interactions among the predicted peak-gene interactions by each method. To control the peak-gene interaction number difference between methods, we randomly sampled the same number of interactions for each method (bigger set were down sampled) when we calculated the precision. This process was repeated 10 times and the median value was used as the final precision estimation.

Enrichment analysis for cell type specific interactions

To assess the cell type specificity of predicted circuits interactions, we checked the enrichment of cell type-specific chromatin interactions in MAGICAL-predicted circuits. Cell type specific promoter-capture Hi-C interactions[13] detected in CD14 monocytes, CD4 T cells, CD8 T cells and B cells were downloaded from https://osf.io/u8tzp/files/osfstorage. The overlap between MAGICAL-predicted peak-gene interactions and chromatin interactions identified in pcHi-C experiments is assessed the same way as above. In this analysis, we used enrichment p-value instead of precision or AUC. Because the multiome data used for MAGICAL analysis (PBMC cell types from *S. aureus* infection and control samples) and these pcHi-C experiments (primary hematopoietic cell types) were collected from close but different contexts, we only expected a significant overlap. For circuits predicted in each cell type, the enrichment *p*-value of cell type-matched pcHi-C interactions was calculated, using all

candidate peak-gene pairs as background. The enrichment was also assessed in non-macthed cell types for cell type specificity evaluation.

GWAS enrichment analysis

To assess the enrichment of GWAS loci of inflammatory diseases in circuit chromatin sites in each cell type, significant GWAS loci were downloaded from GWAS catalog (https://www.ebi.ac.uk/gwas/) for inflammatory diseases including celiac disease, Crohn's disease, inflammatory bowel disease, type 1 diabetes, multiple sclerosis, primary biliary cirrhosis, rheumatoid arthritis, systemic lupus erythematosus, ulcerative colitis and psoriasis and control diseases including Alzheimer's, ADHD, bipolar depression, Schizophrenia, Parkinson's, type 2 diabetes. GREGOR[65] was used to assess the enrichment of GWAS loci at which either the index SNP or at least one of its LD proxies overlaps with a circuit chromatin site, using pre-calculated LD data from 1000G EUR samples. The enrichment *p*-value of each disease GWAS was converted to a z-score for group comparison between inflammatory diseases and control diseases. With each cell type, enrichment scores for traits with fewer than 5 overlapped GWAS SNPs with circuit sites were hold out. Also, as all reference data used by GREGOR is hg19 based, genome coordinates of testing regions were mapped from hg38 to hg19.

Feature gene selection for predicting *S. aureus* infection

To refine circuit genes lately used for predicting infection diagnosis in microarray gene expression data, the capability of each circuit gene on distinguishing infection and control samples was assessed using sample level pseudobulk gene expression data, aggregated from the discovery scRNA-seq datasets. The total number of reads of each sample was normalized to 1e7. For each circuit gene in common to MRSA and MSSA, AUROC (area under the ROC curve) was calculated by sorting the S. aureus samples (labeled to '1') and control samples (labeled to '0') based on the normalized pseudobulk gene expression and comparing sample ranks against sample labels in the discovery scRNA-seq data. We prioritized circuit genes based on their discovery AUROC and tested top ranked genes with a varying threshold from 0.5 to 0.85. An SVM model was trained using circuit genes as features and their normalized pseudobulk expression data

of 21 infection samples and 23 control samples as input. Genes over AUROC > 0.7 provided the most robust prediction performance across all validation datasets (**Fig. S11a**; **Table S13**).

Feature gene selection for predicting *S. aureus* antibiotic sensitivity

For each gene identified in the comparative analysis of MRSA and MSSA, AUROC was calculated by sorting the MRSA samples (labeled to '1') and MSSA samples (labeled to '0') based on the normalized pseudobulk gene expression and comparing sample ranks against sample labels. To distinguish MRSA and MSSA strain infections (predicting antibiotic sensitivity based on host responses), an SVM model was trained using circuit genes as features and their normalized pseudobulk expression data of 10 MRSA samples and 11 MSSA samples as input. After a comprehensive evaluation using genes above discovery AUROC cutoffs, 32 circuit genes with discovery AUROC > 0.7 were selected due to their high prediction performance on validation datasets (**Fig. S12**; **Table S14**). For DEG, we also tested the top DEG prioritized by discovery AUROC or by other approaches like the Minimum Redundancy Maximum Relevance (MRMR) algorithm or LASSO regression.

Data availability

Data from scRNA-seq and scATAC-seq of *S.aureus* infection and control PBMC samples and scATAC-seq of validation COVID-19 positive and PCR negative control PBMC samples have been deposited with the Gene Expression Omnibus under accession no. GSE220190. All the other data used in this study are publicly available and the URLs are provided in the corresponding sections in Methods.

Code availability

A Github repository for source code of MAGICAL and processed single cell datasets is available at https://github.com/xichensf/magical.

Acknowledgments

We thank the Single-cell and Spatial Technologies team at the Center for Advanced Genomics Technology, Department of Genetics and Genomic Sciences, the Icahn School of Medicine at Mount Sinai for providing the experimental, computational, data resources, and staff expertise. The Navy COVID-19 study protocol was approved by the Naval Medical Research Center institutional review board (protocol number NMRC.2020.0006) in compliance with all applicable Federal regulations governing the protection of human subjects. The staphylococcus sepsis protocol was reviewed and approved by the Duke Medical School institutional review board (protocol number Pro00102421). This study was funded by the Defense Advanced Research Projects Agency contract number N6600119C4022 (S.C.S.), Defense Health Agency grant 9700130 through the Naval Medical Research Center (A.G.L.), National Institutes of Health grant R01GM071966 (O.G.T.), and Simons Foundation grant 395506 (O.G.T.).

Author Contributions

S.C.S, O.G.T, and E.Z conceived the study and supervised the research. X.C. designed and implemented the computational framework, conducted benchmarks and case studies with Y.W., wrote the code, and set up the web access with the help of A.T. A.C. was involved in the *S. aureus* study. W.S.C. managed and processed single-cell sequencing data with help from A.R., G.N., and A.V. S.K. and D.C. conducted the public microarray data search. F.R.Z., V.N., M.C.G., and R.S. generated the PBMC single-cell multiomics data for the *S. aureus* infected and control subjects. The *S. aureus* patient blood samples were provided by C.W.W, V.G.F., F.R, and M.D. The control samples were provided by R.R.S. and L.C.N. I.R. and C.M.M. provided immunological interpretations of the results. I.R., A.G.L., and A.S.S. provided the validation PBMC scATAC-seq data of COVID-19 patients and uninfected controls. S.C.S, O.G.T, X.C., E.Z., A.C., and C.T. wrote the first draft of the manuscript. All authors proofread the submitted version.

Disclosures

A.G.L. is a military service member. This work was prepared as part of his official duties. Title 17, US Code §105 provides that copyright protection under this title is not available for any work of the US Government. Title 17, US code §101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties. The views expressed in the article are those of the authors and do not necessarily express the official policy and position of the US Navy, the Department of Defense, the US Government, or the institutions affiliated with the authors.

V.G.F. reports personal fees from Novartis, Debiopharm, Genentech, Achaogen, Affinium, Medicines Co., MedImmune, Bayer, Basilea, Affinergy, Janssen, Contrafect, Regeneron, Destiny, Amphliphi Biosciences, Integrated Biotherapeutics; C3J, Armata, Valanbio; Akagera, Aridis, Roche, grants from NIH, MedImmune, Allergan, Pfizer, Advanced Liquid Logics, Theravance, Novartis, Merck; Medical Biosurfaces; Locus; Affinergy; Contrafect; Karius; Genentech, Regeneron, Deep Blue, Basilea, Janssen; Royalties from UpToDate, stock options from Valanbio and ArcBio, Honoraria from Infectious Diseases of America for his service as Associate Editor of Clinical Infectious Diseases, and a patent sepsis diagnostics pending.

S.C.S. is a founder of GNOMX Corp.

Reference

1. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* **20**, 437-455 (2019).
2. Kim, H.D., Shay, T., O'Shea, E.K. & Regev, A. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* **325**, 429-32 (2009).
3. mod, E.C. *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-97 (2010).
4. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods* **13**, 366-70 (2016).
5. Wilk, A.J. *et al.* Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19. *J Exp Med* **218**(2021).
6. Krijger, P.H. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol* **17**, 771-782 (2016).
7. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).
8. Kreitmaier, P., Katsoula, G. & Zeggini, E. Insights from multi-omics integration in complex disease primary tissues. *Trends Genet* (2022).
9. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e21 (2019).
10. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116 e20 (2020).
11. Jiang, Y. *et al.* Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions. *Cell Syst* **13**, 737-751 e4 (2022).
12. Cao, Z.J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* **40**, 1458-1466 (2022).
13. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e19 (2016).
14. Arnold, S.R. *et al.* Changing patterns of acute hematogenous osteomyelitis and septic arthritis: emergence of community-associated methicillin-resistant Staphylococcus aureus. *J Pediatr Orthop* **26**, 703-8 (2006).
15. Saavedra-Lozano, J. *et al.* Changing trends in acute osteomyelitis in children: impact of methicillin-resistant Staphylococcus aureus infections. *J Pediatr Orthop* **28**, 569-75 (2008).
16. Liao, J.C. *et al.* Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A* **100**, 15522-7 (2003).
17. Tran, L.M., Brynildsen, M.P., Kao, K.C., Suen, J.K. & Liao, J.C. gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab Eng* **7**, 128-41 (2005).
18. Kartha, V.K. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genom* **2**(2022).
19. Teng, L., He, B., Wang, J. & Tan, K. 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics* **31**, 2560-4 (2015).

20. Mumbach, M.R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49**, 1602-1612 (2017).
21. Arunachalam, P.S. *et al.* Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* **369**, 1210-1220 (2020).
22. Lucas, C. *et al.* Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463-469 (2020).
23. Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* **369**(2020).
24. Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **182**, 1419-1440 e23 (2020).
25. Granja, J.M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**, 403-411 (2021).
26. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-40 (2012).
27. Li, S. *et al.* Epigenetic Landscapes of Single-Cell Chromatin Accessibility and Transcriptomic Immune Profiles of T Cells in COVID-19 Patients. *Front Immunol* **12**, 625881 (2021).
28. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet* **51**, 1442-1449 (2019).
29. Chen, X. *et al.* Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. *Cell Syst* **12**, 353-362 e6 (2021).
30. Yao, C. *et al.* Cell-Type-Specific Immune Dysregulation in Severely Ill COVID-19 Patients. *Cell Rep* **34**, 108590 (2021).
31. Unterman, A. *et al.* Single-cell multi-omics reveals dyssynchrony of the innate and adaptive immune system in progressive COVID-19. *Nat Commun* **13**, 440 (2022).
32. Magill, S.S. *et al.* Changes in Prevalence of Health Care-Associated Infections in U.S. Hospitals. *N Engl J Med* **379**, 1732-1744 (2018).
33. Tong, S.Y., Davis, J.S., Eichenberger, E., Holland, T.L. & Fowler, V.G., Jr. Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev* **28**, 603-61 (2015).
34. Marquez-Ortiz, R.A. *et al.* USA300-related methicillin-resistant Staphylococcus aureus clone is the predominant cause of community and hospital MRSA infections in Colombian children. *Int J Infect Dis* **25**, 88-93 (2014).
35. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e29 (2021).
36. Skjeflo, E.W., Christiansen, D., Espevik, T., Nielsen, E.W. & Mollnes, T.E. Combined inhibition of complement and CD14 efficiently attenuated the inflammatory response induced by Staphylococcus aureus in a human whole blood model. *J Immunol* **192**, 2857-64 (2014).
37. Kusunoki, T., Hailman, E., Juan, T.S., Lichenstein, H.S. & Wright, S.D. Molecules from Staphylococcus aureus that bind CD14 and stimulate innate immune responses. *J Exp Med* **182**, 1673-82 (1995).

38. Ludwig, S. *et al.* Influenza virus-induced AP-1-dependent gene expression requires activation of the JNK signaling pathway. *J Biol Chem* **276**, 10990-8 (2001).
39. Gjertsson, I., Hultgren, O.H., Collins, L.V., Pettersson, S. & Tarkowski, A. Impact of transcription factors AP-1 and NF-kappaB on the outcome of experimental Staphylococcus aureus arthritis and sepsis. *Microbes Infect* **3**, 527-34 (2001).
40. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **12**, R83 (2011).
41. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res* **50**, D687-D692 (2022).
42. Kyriakis, J.M. Activation of the AP-1 transcription factor by inflammatory cytokines of the TNF family. *Gene Expr* **7**, 217-31 (1999).
43. Hannemann, N. *et al.* The AP-1 Transcription Factor c-Jun Promotes Arthritis by Regulating Cyclooxygenase-2 and Arginase-1 Expression in Macrophages. *J Immunol* **198**, 3605-3614 (2017).
44. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377-390 e19 (2019).
45. Consortium, E.P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
46. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
47. DeLorenze, G.N. *et al.* Polymorphisms in HLA Class II Genes Are Associated With Susceptibility to Staphylococcus aureus Infection in a White Population. *J Infect Dis* **213**, 816-23 (2016).
48. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).
49. Ahn, S.H. *et al.* Gene expression-based classifiers identify Staphylococcus aureus infection in mice and humans. *PLoS One* **8**, e48979 (2013).
50. Ramilo, O. *et al.* Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* **109**, 2066-77 (2007).
51. Ardura, M.I. *et al.* Enhanced monocyte response and decreased central memory T cells in children with invasive Staphylococcus aureus infections. *PLoS One* **4**, e5446 (2009).
52. Cho, J.S. *et al.* IL-17 is essential for host defense against cutaneous Staphylococcus aureus infection in mice. *J Clin Invest* **120**, 1762-73 (2010).
53. Xiao, Y. *et al.* A novel significance score for gene selection and ranking. *Bioinformatics* **30**, 801-7 (2014).
54. Chaussabel, D. *et al.* A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**, 150-64 (2008).
55. Wenric, S. & Shemirani, R. Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies. *Front Genet* **9**, 297 (2018).
56. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289-1296 (2019).
57. Squair, J.W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).

58. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
59. Anderson, M.A. & Gusella, J.F. Use of cyclosporin A in establishing Epstein-Barr virus-transformed human lymphoblastoid cell lines. *In Vitro* **20**, 856-8 (1984).
60. Tan, L., Xing, D., Chang, C.H., Li, H. & Xie, X.S. Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924-928 (2018).
61. McArthur, E. & Capra, J.A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am J Hum Genet* **108**, 269-283 (2021).
62. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
63. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* **44**, e70 (2016).
64. Letizia, A.G. *et al.* SARS-CoV-2 seropositivity and subsequent infection risk in healthy young adults: a prospective cohort study. *Lancet Respir Med* **9**, 712-720 (2021).
65. Schmidt, E.M. *et al.* GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601-6 (2015).