Airqtl dissects cell state-specific causal gene regulatory networks with efficient single-cell eQTL mapping

Lingfei Wang*

Department of Genomics and Computational Biology
University of Massachusetts Chan Medical School
Worcester, MA, USA

January 15, 2025

Abstract

Single-cell expression quantitative trait loci (sceQTL) mapping offers a powerful approach for understanding gene regulation and its heterogeneity across cell types and states. It has profound applications in genetics and genomics, particularly causal gene regulatory network (cGRN) inference to unravel the molecular circuits governing cell identity and function. However, computational scalability remains a critical bottleneck for sceQTL mapping, prohibiting thorough evaluation and optimization of statistical accuracy. We present airqtl, a novel method to overcome these challenges through algorithmic advances and efficient implementations of linear mixed models. Airqtl achieves superior time complexity and over eight orders of magnitude of acceleration, enabling objective method evaluation and optimization. Airqtl offers *de novo* inference of robust, experimentally validated cell state-specific cGRNs that reflect perturbation outcomes. Our results dissect the drivers of cGRN heterogeneity and underscore the value of natural genetic variations in primary human cell types for biologically relevant single-cell cGRN inference.

*Lingfei.Wang@umassmed.edu

Introduction

Single-cell technology offers a unique opportunity to infer gene regulatory networks (GRNs) at unprecedented scale. Recent computational advances [1–4] have enabled the inference of cell type-specific and dynamic GRNs from a single sample. hese approaches significantly reduce experimental costs and effort while enhancing the specificity of GRNs for distinct cell types, developmental stages, and extracellular conditions (collectively referred to as "states"). However, many existing methods focus on optimizing predictive performance between genes, often confusing causation with reverse causation or confounding on its own. These noncausal GRNs often fail to predict perturbation outcomes, limiting their reliability in applications such as disease understanding and drug discovery. While some approaches introduced causal evidence with transcription factors (TFs) and their DNA-binding events, these partial GRNs overlook other important regulatory mechanisms, such as signaling pathways, chromatin modifications, and non-coding RNAs.

Causal inference, particularly with perturbations, provides a robust framework for distinguishing causation and accounting for diverse regulatory mechanisms. This paradigm has been successful across various natural and social sciences, including single-cell experiments such as Perturb-seq [5–8]. However, these studies primarily involve engineered cell lines subjected to strong artificial perturbations. Since GRNs are highly dynamic and cell state-specific [1–4], the relevance of these findings to primary cell types in natural human populations with both strong and mild perturbations remains suboptimal. In contrast, Mendelian randomization (MR) leverages natural genetic perturbations as instrumental variables to infer causality between organismal and/or molecular traits [9–13]. MR has been effectively employed to infer tissue-specific causal GRNs (cGRNs) from population-scale bulk RNA sequencing (RNA-seq), using expression quantitative trait loci (eQTLs) that directly regulate cis-gene expression and indirectly influence trans-genes [14].

Applying MR to population-scale single-cell RNA-seq (scRNA-seq) promises greater cell state specificity and cost-effectiveness than bulk RNA-seq for cGRN inference. While dozens of population-scale scRNA-seq studies have been conducted (e.g. [15–19]), the challenges in single-cell eQTL (sceQTL) mapping [15, 20] have limited its utility for downstream cGRN inference. On the statistical front, single-cell data sparsity necessitates specialized normalization techniques [5, 21, 22], while the shared genome of cells from the same donor requires (generalized) linear mixed models [16, 23]. However, the primary bottleneck lies in computational scalability. Without scalable methods, it is impractical to benchmark and improve statistical performance at reasonable cost and time. Consequently, most population-scale scRNA-seq studies rely on *ad hoc* sceQTL mapping computer programs, with few capable of mapping all cis- and trans-sceQTLs across all single nucleotide polymorphism (SNP)–gene pairs. Some studies employ pseudo-bulk strategies for eQTL mapping, but they face challenges with continuous cell states and technical confounders [15, 20]. Accurate and efficient

eQTL mapping is also pivotal in human genetics [24] and genomic foundation models [25, 26]. The absence of a scalable sceQTL mapping method has hindered progress in these domains.

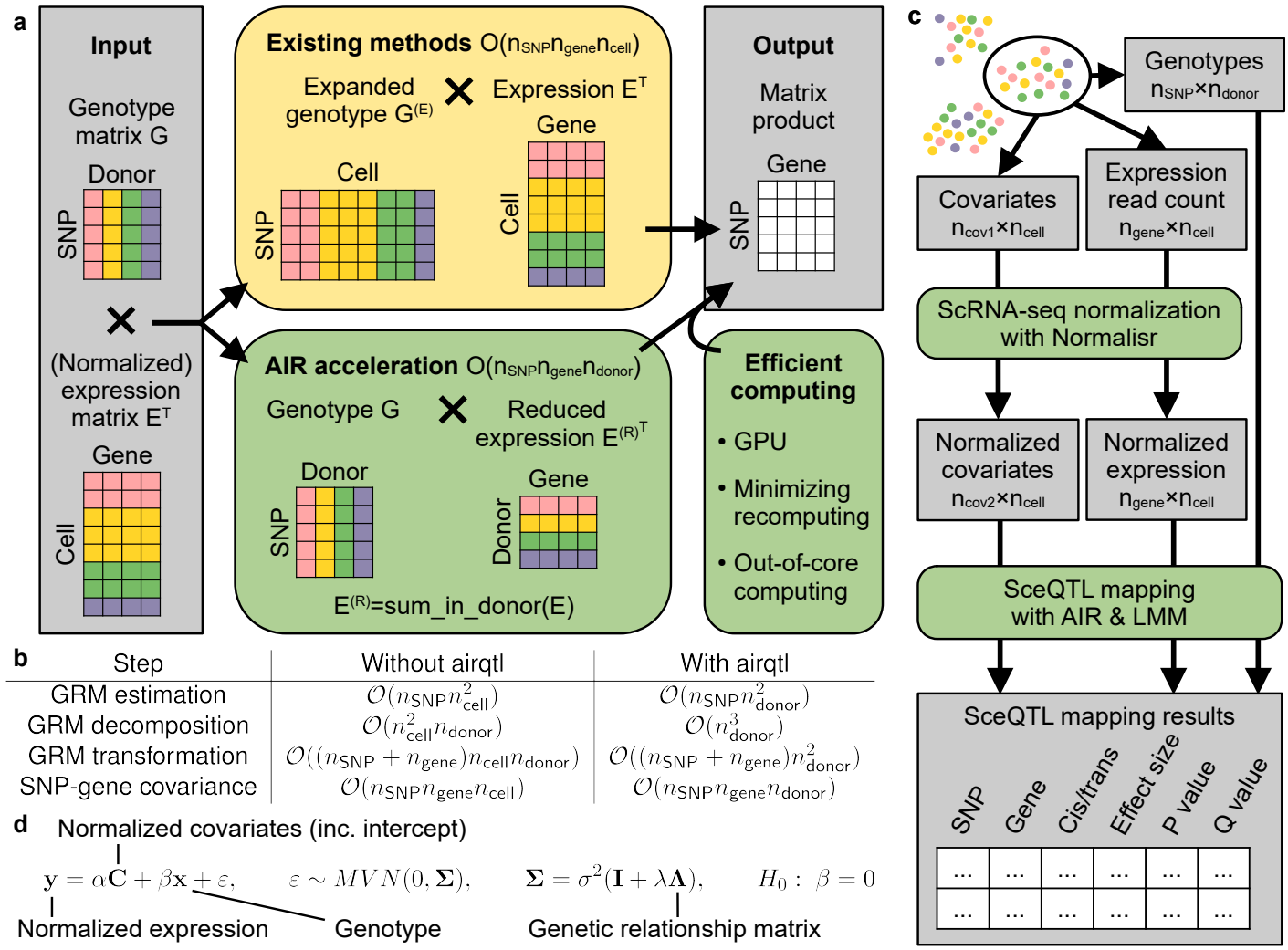
In this paper, we present airqtl, a novel method to overcome these challenges in sceQTL mapping and infer cell state-specific cGRNs. Airqtl introduces a novel exact algorithm that accelerates basic matrix operations by leveraging the donor-cell hierarchy in population-scale single-cell datasets. This advancement improves the time complexity of linear mixed models beyond state-of-the-art methods such as FaST-LMM [27] and GEMMA [28]. Together with GPU utilization and other implementation optimizations, airqtl achieves over eight orders of magnitude in computational acceleration at a comparable or higher statistical accuracy. This scalability facilitates objective benchmarking and refinement of cell state-specific sceQTL mapping and admixed population modeling. Finally, airqtl's efficient and accurate sceQTL mapping allows the *de novo* inference of robust, experimentally validated cell state-specific cGRNs. Comparative analyses reveal intrinsic and extrinsic drivers of cGRN heterogeneity, underlining the importance of cell state-specific experimental data and computational methods.

Results

Airqtl overview

The core of airqtl is the Array of Interleaved Repeats (AIR), a compressed data structure designed to efficiently handle operations between donor-level (e.g., genome) and cell-level (e.g., transcriptome) variables. By leveraging the donor-cell hierarchy, AIR enables novel exact accelerations in basic matrix operations. For instance, a critical step of sceQTL mapping involves computing the covariance between all SNPs and normalized gene expression levels using matrix multiplication (**Fig. 1a**). Conventional methods adapted from bulk eQTL mapping would expand the genotype matrix to the cell level, resulting in $\mathcal{O}(n_{\text{SNP}} n_{\text{gene}} n_{\text{cell}})$ time complexity. In contrast, AIR reduces the gene expression matrix to the donor level, achieving a significantly lower complexity of $\mathcal{O}(n_{\text{SNP}} n_{\text{gene}} n_{\text{donor}})$ (**Methods**). This principle underpins similar accelerations across all steps of sceQTL mapping with linear mixed models (LMM), resulting in a total time complexity of $\mathcal{O}(n_{\text{gene}} n_{\text{cell}} + (n_{\text{SNP}} + n_{\text{donor}})(n_{\text{gene}} + n_{\text{donor}}) n_{\text{donor}})$ (**Fig. 1b**, **Methods**, **Supplementary Note**). Compared to state-of-the-art LMM methods such as FaST-LMM [27] and GEMMA [28], which operate at $\mathcal{O}((n_{\text{SNP}} + n_{\text{donor}})(n_{\text{gene}} + n_{\text{cell}}) n_{\text{cell}})$ time complexity, AIR delivers a $\sim$1000-fold algorithmic speedup for sceQTL mapping with 100,000 cells and 1,000 donors per cell type.

Airqtl performs sceQTL mapping for each cell state separately using a streamlined two-step process (**Fig. 1c**, **Methods**). First, quality-controlled scRNA-seq read counts and covariates are normalized for each



Step	Without airqtl	With airqtl
GRM estimation	$\mathcal{O}(n_{\text{SNP}} n_{\text{cell}}^2)$	$\mathcal{O}(n_{\text{SNP}} n_{\text{donor}}^2)$
GRM decomposition	$\mathcal{O}(n_{\text{cell}}^2 n_{\text{donor}})$	$\mathcal{O}(n_{\text{donor}}^3)$
GRM transformation	$\mathcal{O}((n_{\text{SNP}} + n_{\text{gene}}) n_{\text{cell}} n_{\text{donor}})$	$\mathcal{O}((n_{\text{SNP}} + n_{\text{gene}}) n_{\text{donor}}^2)$
SNP-gene covariance	$\mathcal{O}(n_{\text{SNP}} n_{\text{gene}} n_{\text{cell}})$	$\mathcal{O}(n_{\text{SNP}} n_{\text{gene}} n_{\text{donor}})$

$$\mathbf{y} = \alpha \mathbf{C} + \beta \mathbf{x} + \varepsilon, \qquad \varepsilon \sim MVN(0, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} = \sigma^2(\mathbf{I} + \lambda \mathbf{\Lambda}), \qquad H_0: \ \beta = 0$$

Figure 1: **Overview of acceleration from AIR and sceQTL mapping with airqtl. a** AIR acceleration demonstrated in the computation of SNP-gene covariance via matrix multiplication. AIR achieves a time complexity of $\mathcal{O}(n_{\text{donor}})$ whereas existing methods require $\mathcal{O}(n_{\text{cell}})$. Additional optimizations in airqtl further enhance time and memory efficiency (bottom right). **b** Exact acceleration offered by AIR/airqtl as shown by time complexity comparisons with state-of-the-art algorithms for each step of sceQTL mapping using LMM. **c** Workflow of sceQTL mapping with airqtl for each selected cell type. **d** Linear mixed model and null hypothesis tested by airqtl for sceQTL mapping between each genotype and gene expression level. Box color: method (yellow: existing; green: AIR/airqtl; gray: data). Matrix color (**a**) and schematic low-dimensional embedding color (**c**): donor.

cell state using Normalisr [5]. Normalisr is optimized for scRNA-seq normalization and differential expression analysis with linear models, delivering both high accuracy and exceptional speed compared to generalized linear model-based methods. Normalisr's accuracy and efficiency has been uniquely demonstrated in mapping the cis- and trans-effects of CRISPRi gRNAs and inferring cGRNs from large-scale studies with high multiplicity of infection, a scenario that closely resembles sceQTL mapping.

Next, airqtl applies LMM to identify sceQTLs from the normalized data (**Fig. 1d**, **Methods**). LMM is particularly suited for scRNA-seq data as it accounts for the genetic similarity among cells from the same donor using the genetic relationship matrix (GRM). This approach is analogous to its use in genome-wide association studies (GWAS) for handling population structures, such as twins, to improve statistical power and reduce false positives [29]. Besides the algorithmic accelerations provided by AIR (**Fig. 1b**), airqtl implements additional optimizations such as GPU utilization (**Fig. 1a**). Airqtl also efficiently maps cell type/state-specific sceQTLs with an extended LMM including all linear and quadratic interaction terms that receives similar accelerations (**Methods**).

Efficient and accurate sceQTL mapping with airqtl

We conducted comprehensive benchmarking to evaluate the performance improvements offered by airqtl. In GWAS and eQTL mapping, simulated datasets have been pivotal for method benchmarking due to the lack of ground-truth information [27, 30–32]. Following this precedent, we simulated population-scale scRNA-seq datasets of various sizes, incorporating ground-truth characteristics closely resembling real-world studies (**Methods**). These simulations were based on Randolph et al [19], a published scRNA-seq dataset comprising influenza A virus (flu) or mock (NI)-exposed human peripheral blood mononuclear cells (PBMCs) from male donors of African and European descent. This dataset was subsequently used for sceQTL mapping and cGRN inference in this study. We compared airqtl with CellRegMap [23], the only existing method explicitly designed for sceQTL mapping with frequentist inference. Bayesian [33] and heuristic [34] methods were excluded due to their reduced compatibility with downstream analyses or population-scale scRNA-seq datasets (**Methods**).

Due to CellRegMap's inability to scale to full cis- and trans-sceQTL mapping in large population-scale scRNA-seq datasets, we first tested both methods on small-scale datasets (**Fig. 2a**), where airqtl already demonstrated substantial advantage. Based on these results, we extrapolated CellRegMap's running time for large-scale datasets and compared it to the actual runtime of airqtl (**Methods**). The results revealed that airqtl achieved over eight orders of magnitude of acceleration and six orders of magnitude of cost reduction (**Fig. 2b**).

Accurate effect size estimation is crucial for eQTL mapping and downstream analyses, such as cGRN

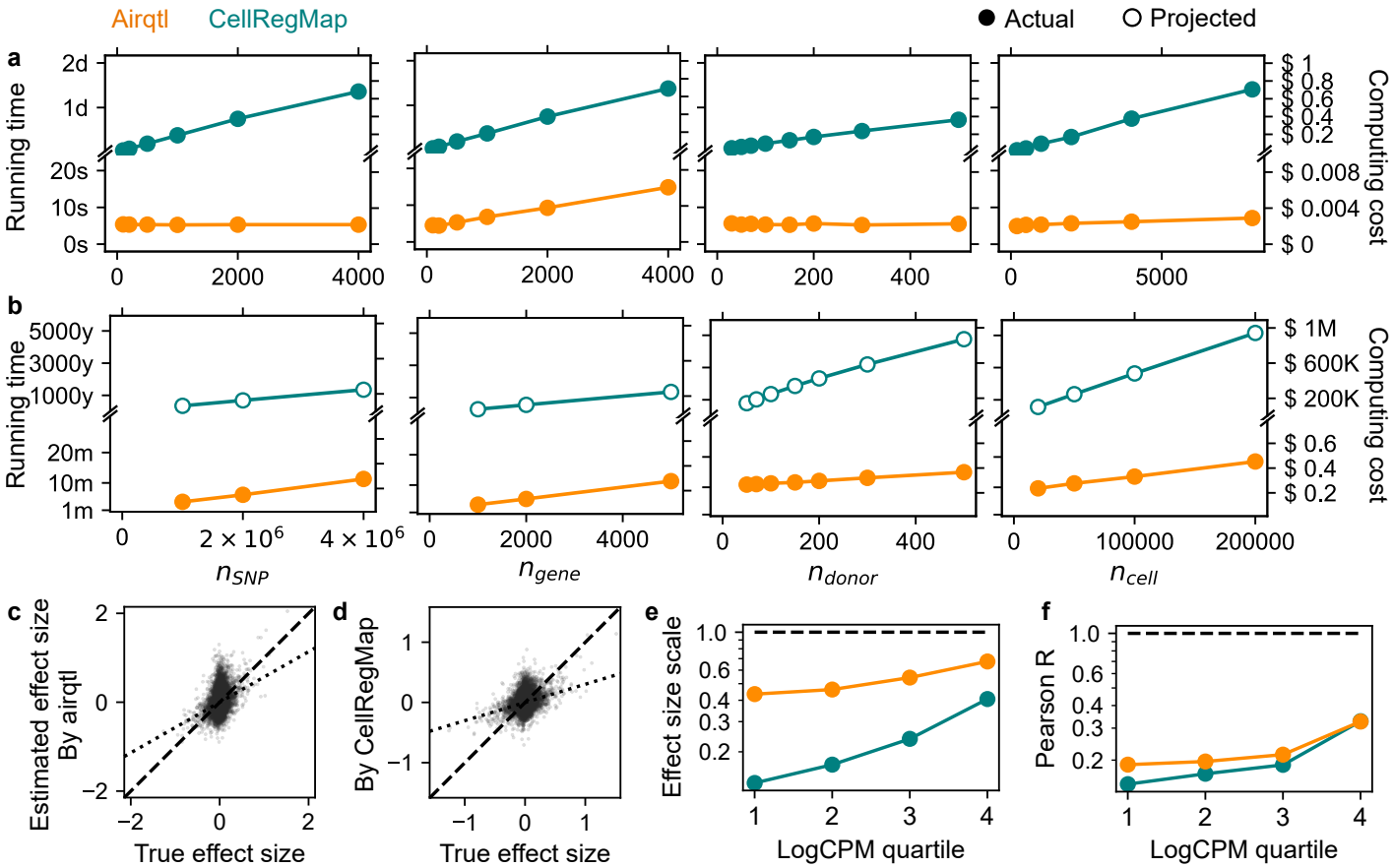


Figure 2: **Airqtl offers eight orders of magnitude of acceleration and improved effect size estimation in sceQTL mapping.** **ab** Running time (left Y) and computational cost (right Y) for sceQTL mapping on small-scale (**a**) and realistic large-scale (**b**) datasets with varying dimension sizes (X). **cd** Ground-truth (X) and estimated (Y) sceQTL effect sizes (dots) by airqtl (**c**) and CellRegMap (**d**) without stratification. Deviation of the best-fit linear model (dotted line) from the diagonal (dashed line) indicates the extent of overall effect size underestimation. **ef** Effect size estimation bias (**e**) and variance (**f**) for genes across expression quartiles (X). Perfect performance: dashed line. Color: method.

inference. However, it is often overlooked in single-cell differential expression studies and is particularly challenging due to expression-dependent biases [5]. To assess this issue in sceQTL mapping, we followed a bias-variance decomposition approach [5], stratifying errors by expression level (**Methods**). Airqtl greatly mitigated the expression-dependent underestimation bias of sceQTL effect sizes observed in CellRegMap (**Fig. 2c-f**). This improvement is attributed to the use of Normalisr for scRNA-seq normalization, which has been shown to reduce such biases in differential expression analyses [5]. This advantage is expected to be especially critical for cell type/state-specific sceQTL mapping, where expression-dependent biases can erroneously attribute changes in expression levels as cell type/state-specific sceQTLs.

Besides its advancements in efficiency and effect size estimation, airqtl demonstrated comparable or superior sensitivity and specificity (**Supplementary Fig. 1**, **Supplementary Data 1**, **Methods**). Taken together, airqtl provides a robust and comprehensive solution for efficient and accurate sceQTL mapping.

Efficient cell type-specific sceQTL mapping enabled comprehensive benchmarking and objective method optimization

The interaction between genotype and other factors is well established in genetics [35,36]. Numerous bulk and single-cell eQTL studies have revealed tissue or cell type-specific eQTL effects [15–19, 37]. While sceQTLs have been mapped separately for each cell type, statistical tests are essential to differentiate true cell type specificity from randomness in data but remain rare and underexplored. Airqtl overcomes these limitations by enabling efficient cell type-specific sceQTL mapping with an extended linear mixed model that incorporates linear and quadratic interactions between genotype, cell type, and other covariates (**Methods**). To systematically evaluate performance, we simulated population-scale scRNA-seq datasets with cell type-specific eQTLs based on CD4+ and CD8+ T cell subpopulations from the Randolph et al dataset [19] (**Methods**).

In terms of computational efficiency, airqtl was over five orders of magnitude faster and three orders of magnitude cheaper than CellRegMap (**Fig. 3ab**, **Methods**). This scalability enables genome-wide mapping of cell type-specific sceQTLs. Although CellRegMap's limited scalability prevented statistical benchmarking on large datasets, airqtl showed robust performance in estimating cell type-specific effect sizes (**Fig. 3cd**, **Supplementary Fig. 2**, **Methods**). Interestingly, cell type-specific effects were easier to estimate than sceQTL effects (**Fig. 2ef**), displaying lower bias and variance.

False positives in cell type specificity, in the form of biased null P-value distributions, can lead to unnecessary and costly follow-up experiments. Initial results from airqtl showed an overabundance of small null P-values for both non-eQTLs and non-specific eQTLs (Kolmogorov-Smirnov/KS P$< 10^{-99}$ and P$< 10^{-50}$ respectively, **Fig. 3e**, **Supplementary Data 2**, **Methods**). Given the weak but nonzero expression-dependent

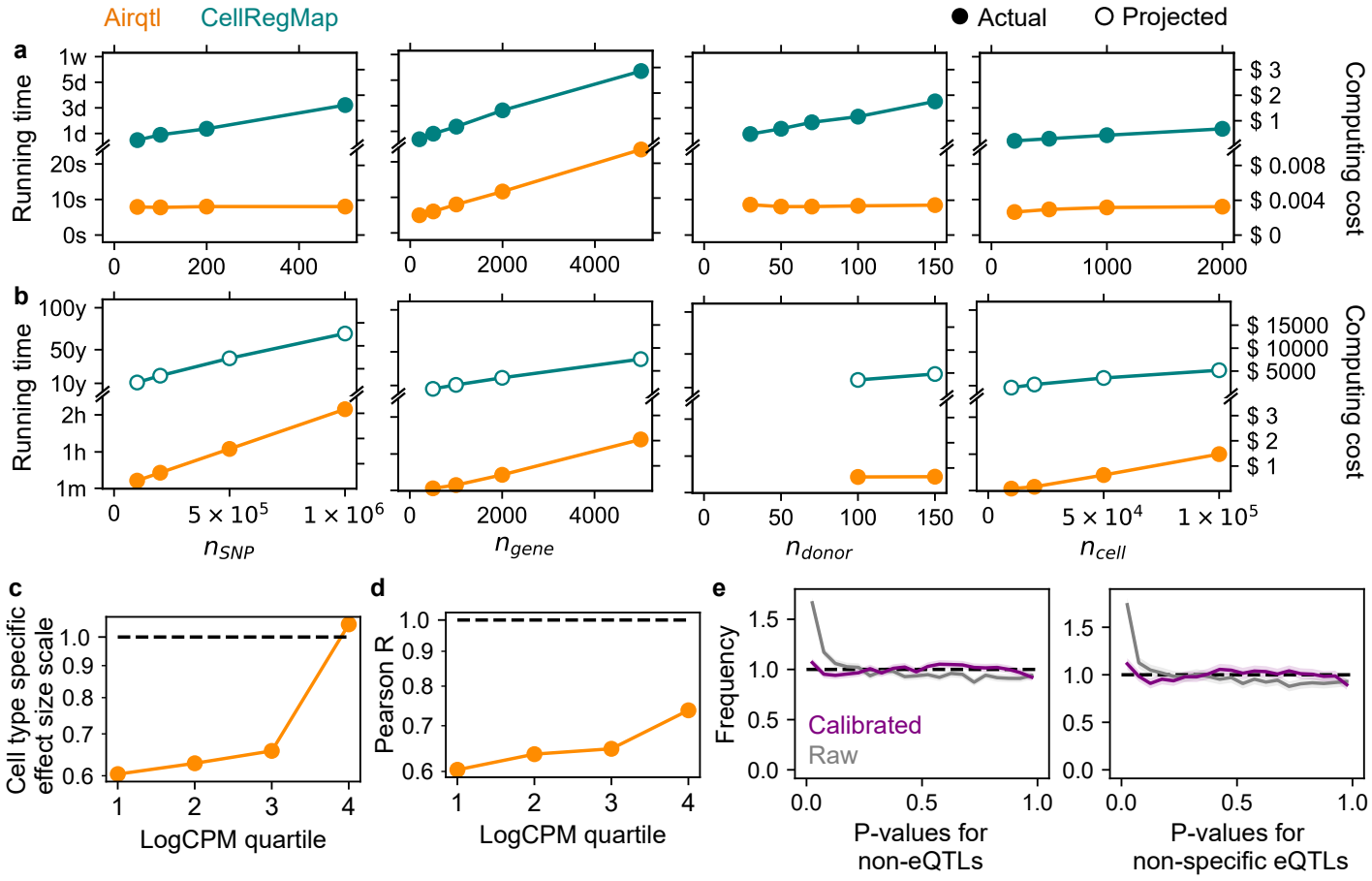


Figure 3: **Airqtl enables efficient cell type-specific sceQTL mapping, comprehensive benchmarking, and objective optimization. ab** Running time (left Y) and computational cost (right Y) for cell type-specific sceQTL mapping on small-scale (**a**) and large-scale (**b**) datasets with varying dimensions. Color: method. **cd** Low bias (**c**) and variance (**d**) in cell type-specific sceQTL effect size estimation by airqtl, stratified by expression quartiles (X). Perfect performance: dashed line. **e** Airqtl's scalability enabled benchmarking and optimization of P-value calibration for cell type specificity. Null P-value histograms are shown for non-eQTLs (left) and non-specific eQTLs (right) before (raw) and after (calibrated) calibration. Shades in histograms represent error bars ($2\sqrt{N}$, where N is the number of entries in each bin). Dashed lines indicate optimal performance as standard uniform distributions.

biases in sceQTL effect size estimation (**Fig. 2e**), this result was anticipated and motivated an optimization of the model. To address this, we calibrated P-values using Beta distributions following existing bulk eQTL studies [37, 38]. To overcome expression-dependent biases in scRNA-seq, we further incorporated stratification with additional parameters. We evaluated 10 candidate parameters, including known single-cell differential expression confounders, their derived parameters, and key population genetics parameters (**Methods**). Each parameter was used individually to parameterize the Beta distribution, which then models the null P-values obtained from separately simulated datasets (**Supplementary Fig. 3**). The learned parameterization was then used to calibrate P-values in the original dataset. Calibration using the parameter "n0min" (proportion of zero-read cells for one gene, minimum across cell types compared) yielded the best null P-value performance, with KS P=0.009 and P=0.12 for non-eQTLs and non-specific eQTLs, respectively (**Supplementary Data 2**, **Supplementary Fig. 4**). This optimization substantially improved the uniformity of null P-value distributions (**Fig. 3e**) and also the separate metric of Storey's π_1 (raw: 0.08, calibrated: 0.006) [39].

In summary, airqtl provides an efficient and scalable solution for cell type-specific sceQTL mapping, uniquely enabling comprehensive benchmarking and objective optimization of statistical performance.

Efficient sceQTL mapping enables causal inference of gene regulatory networks

Genetic variations are effective instrumental variables for inferring cGRNs via Mendelian randomization (MR), given their predominantly cis-regulatory effects. By identifying instrumental SNPs and their significant cis- and trans-genes, causal regulations from cis- to trans-genes can be inferred (**Fig. 4a**). Population-scale scRNA-seq studies offer unprecedented power for cell state-specific cGRN inference but also pose substantial computational challenges. Airqtl's unique efficiency provides a solution.

Population admixture is a well-known source of false positives in GWAS [40], yet its impact on eQTL mapping and cGRN inference remains less understood (**Fig. 4b**). The Randolph et al. dataset [19] includes admixed populations with strong inter-chromosomal genotype correlations (**Supplementary Fig. 5**). While principal component (PC) covariates of the genotype matrix can capture much of this correlation (**Supplementary Fig. 5**), they may also introduce spurious associations via collider bias [41]. An alternative approach involves estimating a genetic relationship matrix (GRM) from genotype data [42]. To assess these methods, we compared two donor-level GRMs (identity matrix versus genotype-based GRM) with and without PC covariates. Using SNPs with highly insignificant cis-associations (all raw P$>$0.5) as negative controls for trans-sceQTL mapping on different chromosomes (**Methods**), the vanilla approach (identity GRM without PC covariates) outperformed other methods in 7/12 cell states (cell type-flu exposure combination, **Fig. 4c**, **Supplementary Fig. 6**). Notably, adding PC covariates had little impact when a genotype-based GRM was used, indicating that

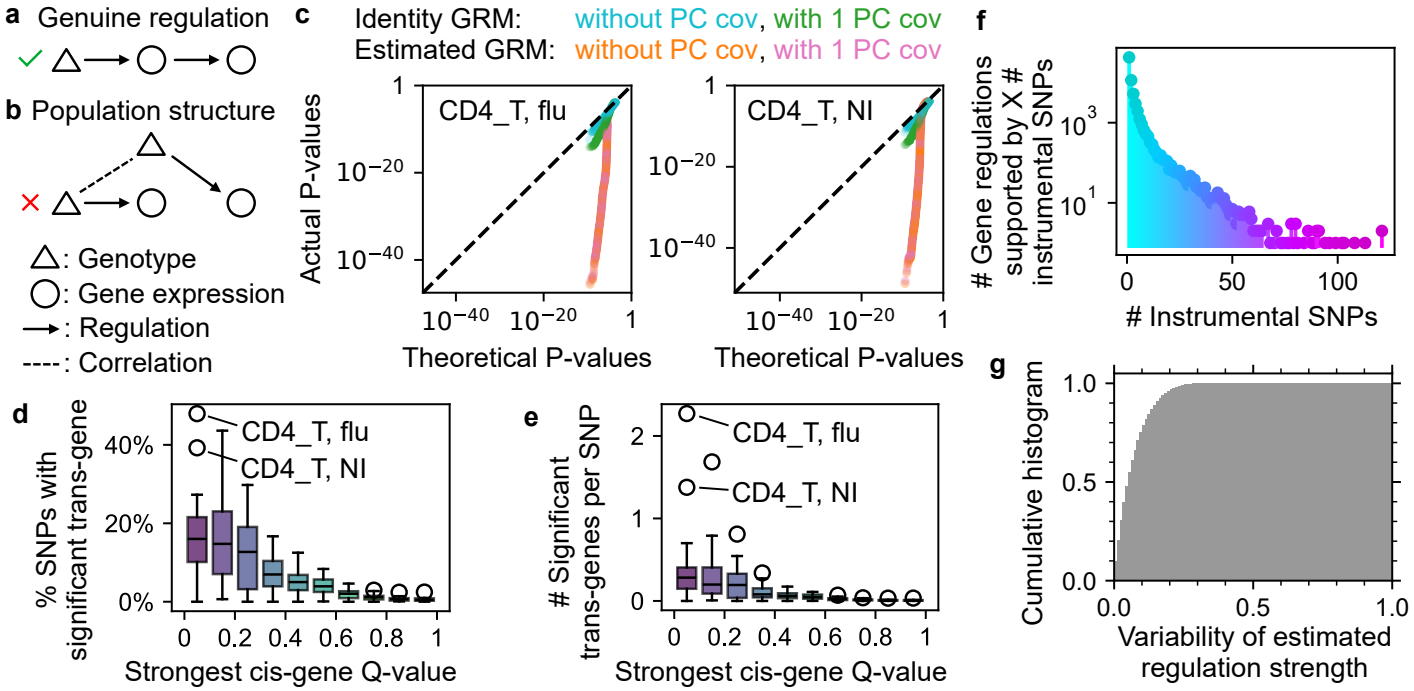


Figure 4: **Airqtl enables efficient inference of cGRNs through Mendelian randomization.** **ab** Scenarios for true positive (**a**) and false positive (**b**) cGRN inference using genotypic variation as an instrumental variable. Population structure (**b**) can lead to false positives in trans-eQTL mapping and cGRN inference. **c** The vanilla method (identity GRM without PC covariate) produced the best quantile-quantile (QQ) plots of null trans-sceQTL P-values for different cell states (panel) among different strategies to handle admixed populations (color). Dashed line: perfect performance. **de** Airqtl identified abundant trans-genes (Y) for cis-sceQTL SNPs, binned by the strongest cis-sceQTL Q-value (X). Data point: cell state at each bin. **f** Numerous inferred gene regulations (Y) were supported by multiple instrumental SNPs (X) in each cell state. **g** Gene regulation strength estimates were consistent across different instrumental SNPs.

GRMs effectively account for admixed population structure. Airqtl's speed enabled mapping of the trans-effects of insignificant SNPs, warranting further investigations into population admixture modeling.

Therefore, we used the vanilla model to map sceQTLs for each cell state (**Supplementary Data 3**). Strong cis-sceQTL SNPs (BH Q$<$0.1) were more likely to exhibit trans-associations and had more significant trans-genes (**Fig. 4de**, BH Q$<$0.5), supporting their utility for cell state-specific cGRN inference. To minimize pleiotropy-induced false positives, possible pleiotropic SNPs that may affect multiple cis-genes were excluded (**Methods**). Each remaining SNP, namely "instrumental SNP", could then infer causal gene regulations from its cis-gene to associated trans-genes via two-stage least squares (2SLS, **Methods**).

Multiple instrumental SNPs allows to evaluate and improve the robustness of inferred cGRNs. Combining gene regulations identified for each cell state, we identified 28,171 gene regulations supported by $\geq$2 unique instrumental SNPs (**Fig. 4f**). Among these, only 1 (0.004%) displayed conflicting directionality predictions (activation vs. repression, **Methods**). Additionally, gene regulation strength estimates were highly consistent across different instrumental SNPs, with 75% exhibiting very low relative variability ($<$10%, **Fig. 4g**, **Methods**). These findings reaffirm the robustness of cGRNs inferred from population-scale scRNA-seq data. Our final cell state-specific cGRNs consisted of gene regulations supported by at least two unique instrumental SNPs, whose regulation strength estimates were aggregated with median (**Supplementary Data 4**).

Comparative analyses revealed intrinsic and extrinsic drivers of cGRN cell state specificity

Single-cell cGRNs offer considerable advantages over noncausal GRNs for understanding molecular circuit rewiring, as they directly quantify perturbation outcomes and indirectly link to organismal trait outcomes. To delineate distinct modes of cGRN rewiring—including detectability, directionality, and strength changes—we first performed pairwise comparisons across cell states (**Fig. 5a**, **Methods**). Detectability varied widely between cell states, reflected in differing overlap rates. These variations stem from both genuine (in)activation of specific gene regulations and technical factors such as cell count. Among overlapping regulations, directionality was highly consistent, with differences confined to a small fraction (up to 3%). We quantified changes in regulation strength with Pearson R and regulation scale, akin to our benchmarks for sceQTL effect size estimation (**Fig. 2ef** and **Fig. 3cd**). Pearson R consistently exceeded 0.8 between all cell states, though regulation scale exhibited up to a twofold variation. Based on these metrics, cGRNs from the same cell type but different conditions were most similar, followed by those from identical conditions (**Fig. 5abc**). These findings suggest limited cGRN rewiring in regulation strength and directionality, driven predominantly by cell type and followed by external condition.

To dissect genuine (in)activation of specific gene regulations from technical detectability variations, we next

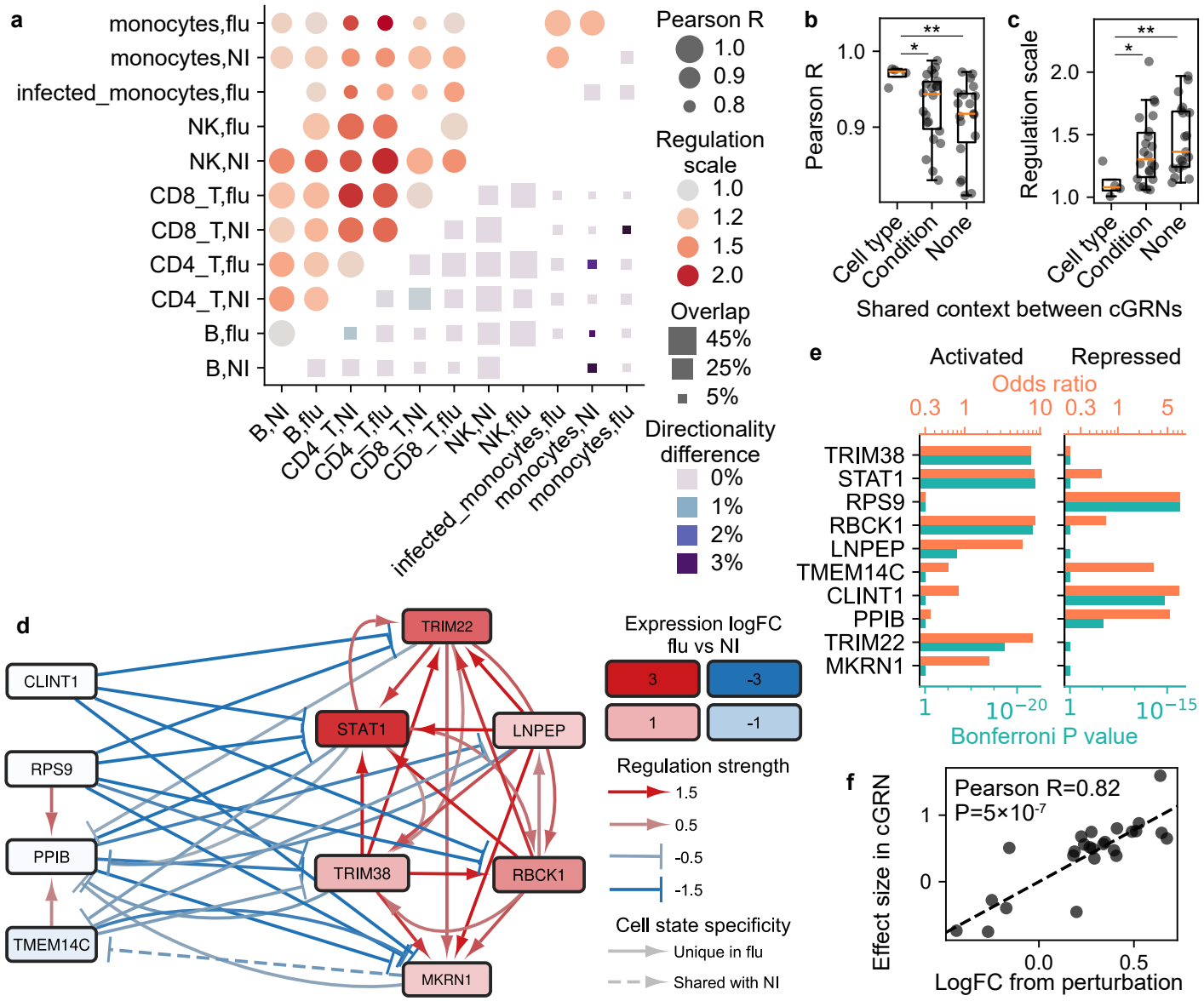


Figure 5: **Comparative analyses revealed intrinsic and extrinsic drivers of cGRN cell state specificity. a** State-specific cGRNs showed moderate overlap but high agreement among the overlapped gene regulations in pairwise comparisons. **bc** Pearson R (**b**) and regulation scale (**c**) comparisons across cGRNs with varying shared contexts (X) highlighted cell type as a stronger determinant of cGRN similarity than condition. Identical cGRNs result in Y=1 in both panels. Mann-Whitney U test P: <0.1 (*); <0.01 (**). **d** All gene regulations in the sub-cGRN of top master regulators were exclusively specific to CD4+ T cells under the flu condition. **e** Enrichment odds ratio (top X) and P value (bottom X) of "defense response to virus" among activated (left) and repressed (right) target genes for each master regulator (Y). **f** Validation of inferred cGRNs using published STAT1 perturbation experiment in CD4+ T cells. The estimated effect sizes of STAT1 regulating other genes (dot) from the cGRN (Y) showed high consistency with estimations from the perturbation experiment (X). Dashed line: best linear fit.

compared CD4+ T cells under flu vs. NI conditions, the most abundant cell states in the dataset with the highest numbers of trans-eQTLs (**Fig. 4de**). Under the flu condition, we identified 10,195 regulations involving 1,938 genes (**Supplementary Fig. 7**, **Supplementary Data 4**), of which 8,539 (84%) were cell state-specific and absent in NI. Notably, most regulators (641/681 or 94%) were not TFs with known motifs [4] (**Methods**), highlighting airqtl's unique capacity for *de novo* inference of cell state-specific cGRNs. A force-directed layout segregated genes in the cGRN into two distinct clusters that repressed each other (**Methods**), with upregulated genes clearly enriched in one cluster (**Supplementary Fig. 7**).

We further focused on the top 10 master regulators with the highest number of target genes. These master regulators segregated into two distinct groups, each connected exclusively by activation edges within the group and repression edges between groups (**Fig. 5d**). Notably, 50/51 (98%) regulations among these master regulators were specific to the flu condition. In contrast, the top 10 master regulators under the NI condition exhibited only 21/89 (24%) regulations as cell state-specific (**Supplementary Fig. 8**). Furthermore, 5/6 master regulators within the upregulated group (TRIM38, STAT1, RBCK1, LNPEP, and TRIM22) had no detectable target under the NI condition. A plausible explanation is that the NI condition predominantly maintains house-keeping regulations while lacking the key gene regulations activated under the flu condition. These findings underscore that active cGRNs are highly specific to the cell state, including both cell type and condition, thereby necessitating cell state-matched data for accurate characterization.

Most of the identified top master regulators [43–53] and their target genes are well-established components of cellular response to viral infections. The activation or repression targets of most (7/10) master regulators were most significantly enriched for the gene ontology (GO) category "defense response to virus" (Bonferroni hypergeometric P-values $<$0.05, **Fig. 5e**, **Supplementary Data 5**, **Methods**). Importantly, the activation/repression target gene set exhibiting this enrichment aligned with the regulatory groups of each master regulator shown in **Fig. 5d**. In contrast, target genes regulated in the opposite direction showed no statistically significant enrichment in any GO category. These enrichments were driven by diverse target genes, with $\sim$20% unique to a single master regulator (**Supplementary Fig. 9**). These findings confirm that the inferred cGRNs and their directional regulation information are biologically meaningful and highly specific to the cell state.

Finally, we validated the cell state-specific cGRNs using published perturbation experiments in CD4+ T cells. We queried the whole Gene Expression Omnibus (GEO) for datasets targeting any master regulator and identified one dataset [54] targeting STAT1, which also happened to match the viral exposure cell state. Among their overlapping targets, the cGRN's estimated effect sizes for STAT1 were highly consistent with the perturbation experiment, with Pearson R=0.82 and P=5×10^{-7} (**Fig. 5f**). This result highlights the capability of

our method to infer cell state-specific cGRNs that are experimentally validated and directly reflect perturbation outcomes.

Discussion

We developed airqtl, an efficient and accurate method for mapping sceQTLs and inferring cGRNs from population-scale scRNA-seq data. By integrating computational advances in algorithm design and implementation with established linear mixed models, airqtl achieves a substantial reduction in computational complexity compared to state-of-the-art methods. This yields orders-of-magnitude faster sceQTL and cell type-specific sceQTL mapping, all while maintaining comparable or superior accuracy. Its unprecedented efficiency enables comprehensive benchmarking and objective optimization of cell type-specific sceQTL mapping and population admixture handling. By efficiently mapping all cis- and trans-sceQTLs, airqtl uniquely infers *de novo* cell state-specific cGRNs that reflect perturbation outcomes in primary human cell types. The inferred cGRNs demonstrated consistency with established biological knowledge and were validated against published perturbation experiments.

Although we demonstrated airqtl's robust statistical accuracy in sceQTL mapping and introduced several optimizations, this study is not a systematic benchmarking or optimization of statistical performance. Instead, the orders of magnitude of acceleration offered by AIR overcame the scalability barrier of these questions and enabled their dedicated future studies. Our analyses also highlighted several key challenges to be further investigated in (cell type-specific) sceQTL mapping, including effect size estimation, biased null P-values, and the complexities of analyzing admixed populations. Beyond airqtl, AIR can similarly empower other existing or novel sceQTL mapping methods, as well as dynamic sceQTL or other molecular QTL mapping tasks that follow similar hierarchical structures. Consequently, the value of AIR acceleration remains significant, even if future methods achieve superior statistical performance.

We established the feasibility of inferring cell state-specific cGRNs from population-scale scRNA-seq datasets, paving the way for further benchmarking and improvement of accuracy in future studies. Unlike other GRNs reconstructed using DNA-binding data or expression predictivity, cGRNs directly and reliably capture perturbation outcomes. Our findings highlight the intrinsic and extrinsic drivers of cGRN specificity and underscore the necessity of cell state-matched data for deciphering cell state-specific gene regulations. Together with other evidence supporting cGRN cell type and state specificity [1–4], our results suggest that cGRNs inferred from population-scale scRNA-seq studies are more biologically relevant than bulk studies or single-cell CRISPR screens using artificial cell lines with strong loss-of-function perturbations.

Methods

Airqtl

Array of Interleaved Repeats (AIR)

AIR is a compressed data structure designed for efficient storage and computation of matrices with many repeated values, often resulting from hierarchical structures. It is implemented as a compressed, interoperable data format compatible with and based on torch.Tensor, a popular library for deep learning offering features such as GPU acceleration and automatic differentiation.

In sceQTL mapping, AIR is used to store genotype data for individual cells in population-scale scRNA-seq studies. For genotype data comprising n_{SNP} SNPs of n_{donor} donors in a population-scale scRNA-seq study involving n_{cell} cells ordered by donor, AIR encodes this information using: i) a donor-level genotype matrix with n_{SNP} rows and n_{donor} columns, and ii) the number of cells associated with each donor. AIR also enables efficient matrix operations between the genotype matrix and other cell-level variables, such as the transcriptome matrix comprising n_{gene} genes and n_{cell} cells. See **Supplementary Note** for more details.

SceQTL mapping — linear mixed model

For sceQTL mapping, we employed the LMM described in **Fig. 1d** using restricted maximum likelihood (reML) [27, 28]. Briefly, for each gene, λ was first estimated as $\hat{\lambda}$ via maximum likelihood under the null hypothesis ($\beta = 0$ for all SNPs). Genes with over 75% of unexplained variance attributable to donor-level variation (i.e., $\hat{\lambda}$ >3) were excluded from downstream analyses, as such high donor-level variation relative to scRNA-seq measurement noise likely indicates strong technical batch effects. Then, for each SNP, the reML was computed by fixing $\lambda = \hat{\lambda}$ and optimizing all other parameters (α, β, σ) to maximize the likelihood. The parameter estimates were obtained through reML. Hypothesis testing was conducted using the likelihood ratio test (LRT) to compare the full model against the null model under the same restriction $\lambda = \hat{\lambda}$. In practice, we equivalently tested conditional R^2 (coefficient of determination) to obtain P values after transforming the data with the GRM and removing covariates. Under the null hypothesis, the conditional R^2 follows a beta distribution, which can be well-approximated by a χ^2 distribution for computational efficiency (default in airqtl).

SceQTL mapping — linear mixed model for cell state specificity

We used the following model to map the cell state specificity of each sceQTL effect size:

$$\mathbf{y} = \alpha_0 + \alpha_1(\mathbf{C}, \mathbf{s}, \mathbf{x}) + \alpha_2(\mathbf{C}, \mathbf{s}) \otimes (\mathbf{C}, \mathbf{s}) + \alpha_3 \mathbf{C} \otimes \mathbf{x} + \alpha_4 \mathbf{x} \otimes \mathbf{x} + \beta \mathbf{s} \otimes \mathbf{x} + \varepsilon. \quad (1)$$

This model incorporates all linear and quadratic interaction terms between genotype, cell state, and other covariates. Here, $\mathbf{x}$ and $\mathbf{y}$ denote genotype and normalized expression level data for the SNP and gene considered, subjected to the same QC and normalization steps as in sceQTL mapping.

$\mathbf{C}$ and $\mathbf{s}$ include all covariates, with $\mathbf{s}$ specifically containing the cell state covariate (e.g., cell type) being tested for eQTL specificity. α_0 is the intercept, and parentheses indicate row-wise matrix concatenation. For $\mathbf{A} \in \mathbb{R}^{n_A \times k}$ and $\mathbf{B} \in \mathbb{R}^{n_B \times k}$, the Kronecker product $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{n_A n_B \times k}$ computes the element-wise product between rows of $\mathbf{A}$ and $\mathbf{B}$. In the case of self-product, $\mathbf{A} \otimes \mathbf{A} \in \mathbb{R}^{n_A(n_A+1)/2 \times k}$ includes only unique row pairs. Similarly, for $\mathbf{A} \in \mathbb{R}^{n_A \times k}$ and $\mathbf{x} \in \mathbb{R}^{1 \times k}$, $\mathbf{A} \otimes \mathbf{x} \in \mathbb{R}^{n_A \times k}$ broadcasts an element-wise product to each row. In practice, $\mathbf{x} \otimes \mathbf{x}$ was replaced with the presence of a dominant allele, $\mathbf{x}^{(\mathsf{dom})} = (x_1^{(\mathsf{dom})}, \dots)$, where $x_i^{(\mathsf{dom})} = 0$ or 1, depending on whether $x_i \neq 0$. Since humans are biallelic ($x_i = 0, 1, 2$), $\mathbf{x} \otimes \mathbf{x}$ and $\mathbf{x}^{(\mathsf{dom})}$ span the same linear space together with $\mathbf{x}$ and the intercept.

The term $\mathbf{s} \otimes \mathbf{x}$ models potential interactions between cell state and genotype, with the cell state-specific eQTL effect size represented by β. The residual variance ε is modeled as:

$$\varepsilon \sim MVN(0, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} = \sigma^2(\mathbf{I} + \lambda \mathbf{\Lambda}). \tag{2}$$

Hypothesis testing for the null hypothesis $\beta = 0$ was conducted using reML and the LRT, as in sceQTL mapping.

SceQTL mapping — algorithmic accelerations

We integrated all existing exact accelerations for LMM and included the unique optimizations enabled by AIR for basic matrix operations (**Supplementary Note**), facilitating efficient sceQTL mapping with LMM in the following steps:

- **GRM estimation**. For the donor-level genotype matrix $\mathbf{G} \in \mathbb{R}^{n_{\mathsf{SNP}} \times n_{\mathsf{donor}}}$, conventional methods expand this to the cell-level genotype matrix $\mathbf{G}^{(\mathsf{E})} \in \mathbb{R}^{n_{\mathsf{SNP}} \times n_{\mathsf{cell}}}$ to estimate the cell-level GRM via matrix multiplication $\mathbf{G}^{(\mathsf{E})} \left(\mathbf{G}^{(\mathsf{E})}\right)^T$ (after mean and variance normalization). This incurs a time complexity of $\mathcal{O}(n_{\mathsf{SNP}} n_{\mathsf{cell}}^2)$. AIR leverages repeated values to enable efficient matrix multiplications, reducing the time complexity to $\mathcal{O}(n_{\mathsf{SNP}} n_{\mathsf{donor}}^2)$.

- **GRM decomposition**. A major optimization introduced in FaST-LMM and GEMMA is the eigendecomposition (via singular value decomposition) of GRM and its subsequent transformation of other matrices. Conventional methods decompose the GRM $\mathbf{G}^{(\mathsf{E})} \left(\mathbf{G}^{(\mathsf{E})}\right)^T$ in $\mathcal{O}(n_{\mathsf{cell}}^2 n_{\mathsf{donor}})$ time, even after recognizing the GRM's low-rank nature. AIR further reduces the eigendecomposition time complexity to $\mathcal{O}(n_{\mathsf{donor}}^3)$.

- **GRM transformation**. Transforming the genotype and expression matrices with GRM eigenvectors, as required in the above optimization, involves matrix multiplication. Conventional methods compute this with a time complexity of $\mathcal{O}((n_{\text{SNP}}+n_{\text{gene}})n_{\text{cell}}n_{\text{donor}})$ after accounting for the GRM's low rank. Using AIR's efficient matrix multiplication, this complexity further is reduced to $\mathcal{O}((n_{\text{SNP}}+n_{\text{gene}})n_{\text{donor}}^2+n_{\text{gene}}n_{\text{cell}})$.

- **SNP-gene covariance**. The naive covariance matrix between genotype and gene expression (without considering GRM) takes $\mathcal{O}(n_{\text{SNP}}n_{\text{gene}}n_{\text{cell}})$ time to compute by conventional methods. AIR accelerates this computation to $\mathcal{O}(n_{\text{gene}}(n_{\text{SNP}}n_{\text{donor}}+n_{\text{cell}}))$ by leveraging its efficient matrix operations.

SceQTL mapping — normalization with Normalisr

Normalization of scRNA-seq read counts and covariates was performed following the GSE120861 tutorial for Normalisr. Briefly, the pipeline included the functions lcpm, normcov, scaling_factor, compute_var, and normvar from normalisr.normalisr. These steps transformed scRNA-seq read counts into the minimum mean square error (MMSE) estimator of log gene expression levels. Technical variations were extracted from the scRNA-seq read counts as additional covariates (in addition to the constant 1 intercept) for inclusion in the linear mixed model during sceQTL mapping and for variance normalization.

SceQTL mapping — P-value calibration for cell state specificity

We calibrated P-values for cell state-specific sceQTL mapping using null P-values derived from a separately simulated null dataset (**Supplementary Fig. 3**). The null P-value distribution was modeled as $Beta(e^{\alpha_0+\alpha_1 x}, e^{\beta_0+\beta_1 x})$. Unlike existing approaches [37, 38], we introduced x, one of the candidate parameters detailed below, computed separately for each gene or SNP to account for potential P-value distribution biases. Before calibration, x was normalized to have zero mean and unit variance, and this normalization was stored for subsequent steps. Parameters $\alpha_0, \alpha_1, \beta_0, \beta_1$ were estimated via maximum likelihood using stochastic gradient descent in PyTorch, with a learning rate of 10^{-6} over 20,000 epochs. To prevent divergence during estimation, P-values smaller than 10^{-300} were clipped to 10^{-300}. For calibration, x was computed for the original dataset, normalized as stored, and the P-value was adjusted by computing the cumulative distribution function (CDF) of the Beta distribution using the parameters learned above.

Candidate parameters for x:

- **Single-cell differential expression technical covariates**: lcpm (pseudo-bulk log(CPM+1) for each gene), n0 (proportion of cells with zero reads for each gene);

- **Derived parameters** (applied to each of the above covariates as x): $x$min (minimum of x's values computed separately for each cell state), $x$diff (difference in x's values between two cell states), $x$diffabs (absolute difference in x's values between two cell states);

- **Population genetics parameters**: maf (minor allele frequency for each SNP), l0 ($\hat{\lambda}$ for each gene).

CGRN inference — sceQTL mapping and filtering

For cGRN inference based on sceQTL summary statistics, we first performed full cis- (within 1Mbp window before or after the transcription start site) and trans-sceQTL mapping, limiting trans-sceQTL candidate outputs to those with $P < 10^{-4}$ for computational efficiency. Gene names containing '-', '.', or '_' were excluded. SNPs not identified as cis-sceQTLs for any gene (BH Q$<$0.1, computed separately for cis-sceQTL candidate pairs) were removed. SNPs considered potentially pleiotropic were excluded from cGRN inference unless associated with a single cis-gene (BH Q $<$ 0.1) while associations with all other cis-genes were highly insignificant (raw P $>$ 0.1). Among the remaining SNP-gene pairs, BH Q values were calculated separately for trans-sceQTL mapping. To accomplish this, we implemented a custom function for calculating BH Q values for P-values below a cutoff. Significant cis-sceQTLs (BH Q $<$0.1) and trans-sceQTLs (BH Q $<$0.5, indicating at least half of identified trans-sceQTLs were true positives) were retained as SNP-gene pairs for cGRN inference.

CGRN inference — effect size estimation

To estimate the effect size of each candidate causal gene regulation in each cell state, we looked for instrumental SNP$\rightarrow$cis-gene (regulator)$\rightarrow$trans-gene (target) relationships. Here the SNP serves as the instrumental variable, as described and identified as a cis-sceQTL for the regulator and a trans-sceQTL for the target above. Using each instrumental SNP, we estimated the gene regulation effect size as the ratio between trans- and cis-sceQTL effect sizes following 2SLS.

CGRN inference — inference robustness evaluation

To evaluate inference robustness, we tested how much the sign and value of estimated gene regulation effect sizes vary across instrumental SNPs. We defined the relative variability of gene regulation strength estimates as their difference between maximum and minimum values divided by the twice of maximum of absolute value. This metric gives 0 when all instrumental SNPs yield identical estimates, 0.5 when one SNP estimates no effect while the others preserve directionality, and 1 when the maximum and minimum estimates have equal absolute values but opposite signs.

CGRN inference — filters with instrumental SNPs

Only gene regulations with at least 2 unique instrumental SNPs were included in the cGRN. Here uniqueness is defined as having non-identical genotype data vectors between the two instrumental SNPs among the donors used to map cell state-specific sceQTLs. Differences merely due to opposite definitions of reference/alternative alleles were regarded as identical.

Data

Published population-scale scRNA-seq datasets

The Randolph et al. dataset [19] contains scRNA-seq data of 255,731 peripheral blood mononuclear cells (PBMCs) sampled from 89 donors of African and European ancestry. he PBMCs were subjected to two distinct conditions: mock treatment (NI) and infection with the H1N1 Cal/04/09 influenza A virus (flu). Donors were genotyped with whole genome sequencing. Cell type and donor annotations provided in the original dataset were retained for this study.

Quality control

We primarily followed the quality control established for high-MOI single-cell CRISPR screen in the Normalisr study [5], with additional criteria for donor-based statistics. Briefly, we removed cells with less than 100 expressed genes or 500 reads, genes expressed in less than 10% of donors, 100 cells, or 2% cells, (cells from) donors with 5 cells or fewer, and SNPs that could not be confidently identified for any donor in the original study.

The above removal was performed separately for each cell subset (e.g. by cell type according to existing annotations) in published (and possibly already QCed) population-scale scRNA-seq data and iteratively until no gene, cell, SNP, or donor was removed. Then, cell subsets with less than 500 cells, 40 donors, 10,000 SNPs, or 500 genes remaining were excluded from sceQTL mapping and downstream analyses. Here specifically, a gene was defined as “expressed” if it exhibited at least 1 read in a particular cell or any cell from a particular donor.

Simulation method for sceQTL mapping benchmarking datasets

Our simulation strategy employed permutation and resampling to maximize the similarity with real datasets for benchmarking purposes. To generate the a simulated dataset with n_{SNP} SNPs, n_{gene} genes, n_{donor} donors, and n_{cell} cells that mimics a real dataset, we followed the approach established in Normalisr [5] and introduced the

contributions of cis-eQTLs. First, to resample these dimensions, we removed non-autosomal genes and randomly downsampled to n_{gene} genes from the remaining ones. SNPs outside 1Mbp window of the transcriptional start site (TSS) of any remaining gene were removed and then downsampled to n_{SNP} randomly to maximize cis-eQTL simulation efficiency. Then, donors and cells were randomly downsampled or upsampled to n_{donor} and n_{cell}, respectively.

For cis-eQTL simulation, each SNP within a 1Mbp distance of any gene's TSS was randomly selected as a cis-eQTL for that gene with a 50% probability. Cis-eQTL effect sizes were then simulated by performing multivariate differential expression using Normalisr on the genotypes of all simulated cis-eQTLs in the resampled dataset. For the log expression of gene i normalized via Normalisr as y_i (see "SceQTL mapping — normalization with Normalisr"), the maximum likelihood estimator for the effect size of SNP j was computed as β_{ji} within the linear model:

$$y_i = \sum_k \alpha_{ki} c_k + \sum_j \beta_{ji} x_j + \varepsilon_i^{(0)}, \tag{3}$$

where y_i represents the normalized expression of gene i, c_k denotes the normalized covariate k (including the intercept), α_{ki} is the effect size of covariate k on gene i's expression, x_j is the genotype of SNP j, and $\varepsilon_i^{(0)} \sim i.i.d\ N\left(0, \left(\sigma^{(0)}\right)^2\right)$ is the cell-level error term. To account for effect size underestimation caused by genotype correlations from linkage disequilibrium (LD) and to extend the range of effect sizes covered in benchmarking, β_{ji} was scaled by a factor of 1.5. The α_{ki} term for the intercept was retained to preserve the mean expression level of gene i.

To simulate genotypes, the existing genotype values (0,1,2) in the resampled dataset for each SNP were randomly permuted across donors to remove LD. This allowed to benchmark false positives in cis-eQTL mapping while preserving the minor allele frequency. In this study, we regard LD a question to be addressed in fine-mapping, beyond the scope of the sceQTL mapping step. Additionally, since neither this study nor CellRegMap aimed to enhance the statistical accuracy of GRM estimation, the use of permuted genotypes provided an ideal test scenario for assessing their core functionalities without introducing unequal genetic relatedness among donors.

To simulate the (unnormalized) true log expression levels for gene i in each cell, y_i, we extended the log expression model from Normalisr:

$$y_i = \alpha_i + \sum_j \beta_{ji} x_j + \varepsilon_i^{(0)} + \varepsilon_i^{(1)}, \tag{4}$$

where α_i represents the intercept and mean expression level for gene i, x_j denotes the genotype of SNP j, β_{ji} is the effect size of SNP j on gene i, $\varepsilon_i^{(0)} \sim i.i.d\ N\left(0, \left(\sigma^{(0)}\right)^2\right)$ is the cell-level error term, and $\varepsilon_i^{(1)} \sim MVN\left(0, \left(\sigma^{(1)}\right)^2 \mathbf{\Lambda}\right)$ is the donor-level error term. Here, $\mathbf{\Lambda}$ is a block diagonal matrix of all-one square

matrices, each sized according to the number of cells from a given donor. The intercept α_i was estimated as the pseudo-bulk mean expression level, and β_{ji} was estimated during the cis-eQTL simulation step. We set $\sigma^{(0)} = 0.5$ and $\sigma^{(1)} = 1$ to ensure a substantial contribution from each source of variation.

To simulate scRNA-seq read counts, we first calculated the normalized true expression level as the softmax across all genes for each cell. This step provided gene-level normalization and incorporated the exponential transformation from the unnormalized log expression level y_i. Then, scRNA-seq read counts for each cell were sampled randomly across all genes using a binomial distribution, with probabilities proportional to the normalized true expression levels. The total number of reads per cell was randomly sampled from its distribution derived from the resampled dataset.

Details about our simulation method can be found in airqtl.sim at `https://github.com/grnlab/airqtl`.

Simulation method for cell type-specific sceQTL mapping benchmarking datasets

The simulation of cell type-specific sceQTLs was based on the sceQTL simulation method with the following modifications: i) Cell type (CD4 vs. CD8 T) was included as an additional one-hot encoded covariate throughout the simulation process. Its effect size on gene expression levels was estimated in α_{ki} in **Eq. 3** and retained for subsequent simulation. Technical covariates introduced by Normalisr were additionally normalized in advance to be orthogonal to cell type covariates, ensuring accurate effect size estimation. ii) Each cis-eQTL was assigned a 50% probability of having cell type-specific effects. For genes with cell type-specific cis-eQTLs, their expression level model (**Eq. 3**) incorporated additional linear terms derived from the interaction between cell type and genotype. The corresponding effect sizes were estimated and stored for use in the simulation. iii) The simulation of true log expression levels (**Eq. 4**) was updated to account for cell type-specific contributions, consistent with the modified effect size estimation model. Specifically, the two extra terms above were added along with contributions from technical covariates from Normalisr and the intercept. Their effect sizes were introduced as they were estimated above. Since these terms already characterized the mean expression levels, the α_i term was removed.

Simulation method for null cell type-specific sceQTL mapping for calibration

Simulation method for null calibration datasets followed the same steps as cell type-specific sceQTL mapping benchmarking datasets but with the modifications: i) Simulations were based on the dataset for which cell type-specific sceQTL mapping P-values were to be calibrated, with identical dimensions (n_{SNP}, n_{gene}, n_{donor}, and n_{cell}). Because simulated datasets were used for benchmarking, they were the input datasets for this simulation to mimic. ii) Genotypes were not permuted, as LD removal was only required for benchmarking. iii)

Cis-eQTLs were present but none of them was cell type-specific. iv) The β_{ji} values were scaled by a factor of 2 (instead of 1.5) to ensure adequate coverage of effect size ranges for calibration purposes.

Simulated datasets for benchmarking

To benchmark the efficiency of sceQTL mapping and cell type-specific sceQTL mapping, we generated a series of simulated datasets designed to mimic the CD4+ T cells (the largest cell subset for sceQTL mapping) or the CD4+ and CD8+ T cells (the largest and most biologically related cell subsets, for cell type-specific sceQTL mapping) from the quality-controlled real dataset by Randolph et al. [19]. Small-scale datasets, designed to accommodate the limited scalability of CellRegMap, had default dimensions of $n_{\text{SNP}} = 500$, $n_{\text{gene}} = 500$, $n_{\text{donor}} = 100$, and $n_{\text{cell}} = 1,000$ for sceQTL mapping, and $n_{\text{SNP}} = 200$, $n_{\text{gene}} = 1,000$, $n_{\text{donor}} = 50$, and $n_{\text{cell}} = 2,000$ for cell type-specific sceQTL mapping. Large-scale datasets had default dimensions of $n_{\text{SNP}} = 4,000,000$, $n_{\text{gene}} = 5,000$, $n_{\text{donor}} = 100$, and $n_{\text{cell}} = 50,000$ for sceQTL mapping, and $n_{\text{SNP}} = 200,000$, $n_{\text{gene}} = 2,000$, $n_{\text{donor}} = 100$, and $n_{\text{cell}} = 50,000$ for cell type-specific sceQTL mapping. Using these default configurations, we also generated related datasets by varying each of these dimensions to assess the scalability of each method under different scenarios.

For statistical performance benchmarking in sceQTL mapping or cell type-specific sceQTL mapping, we similarly generated one simulated dataset each with $n_{\text{SNP}} = 20,000, n_{\text{donor}} = 100, n_{\text{cell}} = 50,000$, and all post-QC genes.

In all cases, datasets were generated by regarding each donor-condition pair from the real dataset as a separate donor. This allows to expand the range of variability in the number of donors. Genotype permutations for LD removal also eliminate any residual population structure, ensuring that each simulated donor is distinct.

Benchmarking

Other methods

For CellRegMap, we followed its website and tutorials for single-cell preprocessing and association testing starting from raw read counts (https://github.com/annacuomo/CellRegMap_analyses/tree/4f26437d20e824e704385deeb2bec1d9f1f61f4e/endodiff/preprocessing, https://github.com/annacuomo/CellRegMap_analyses/blob/4f26437d20e824e704385deeb2bec1d9f1f61f4e/simulations/utils/run_tests.py, `https://limix.github.io/CellRegMap/usage.html`). To allow a consistent benchmarking, latent dimension covariates were not included (see below). For statistical performance benchmarking (below), only SNP-gene

relations in the cis region (1Mbp up or downstream) were computed for CellRegMap's efficiency.

We did not include GASPACHO [33] and other Bayesian methods that are less compatible with downstream analyses such as fine mapping, transcriptome-wide association studies, and Mendelian randomization. We did not include BOLT-LMM [31], SAIGE-QTL [34], or similar heuristic methods in the time complexity comparison or benchmarking. This is because their approximation is designed for GWAS and are prone to large errors without a large number of individuals, which is rarely met by current population-scale scRNA-seq studies. Nevertheless, airqtl is estimated to run approximately 10^3 times faster than SAIGE-QTL based on its projected running time in [34]. Our AIR data structure can also provide similar accelerations to these methods.

Latent dimensions as covariates

In our benchmarking, we did not account for latent dimensions (such as those derived from PCA) from genome and/or transcriptome data as covariates for several reasons. i) This study is focused on improving computational efficiency without sacrificing statistical performance. Identifying the optimal strategy for incorporating latent factors would be the subject of a separate study. ii) While latent dimension covariates may improve false discovery rate control in sceQTL mapping, they can also obscure true signals, reduce statistical power, and lead to spurious associations [41]. Although the impact of these drawbacks varies across dimension reduction methods and data types, we would ideally prefer to identify a method that improves false discovery rate control without introducing these trade-offs. iii) Our benchmark for sceQTL mapping employed independently permuted genotypes, which lacked population structure, and yielded satisfactory results without the need for latent dimension covariates. iv) Airqtl is based on a standard linear mixed model and offers the flexibility to incorporate any type of covariate, including latent dimensions, if necessary, as demonstrated using the real dataset from Randolph et al.

SceQTL mapping and cell type-specific sceQTL mapping efficiency benchmarking

We compared the efficiency of CellRegMap and airqtl on Google Cloud Compute Engine, using the machine type for which each method is designed, and recorded the running times with Snakemake. Specifically, airqtl was run on a2-highgpu-1g with 1 NVIDIA A100 40GB GPU. Since each CellRegMap process was single-threaded, we ran up to 32 CellRegMap processes in parallel on n2d-highmem-32 with 32 vCPUs. No resource competition for vCPUs or memory was observed among the 32 parallel processes. The computing cost was estimated by multiplying the running time by the spot price of the Google Cloud Compute Engine instance ($1.469/hour for a2-highgpu-1g and $0.668/hour for n2d-highmem-32). The cost for CellRegMap was further divided by 32, as each process utilized only 1 vCPU, and the price was linearly proportional to the number of

vCPUs. The datasets described in "Simulated datasets for benchmarking" were used for these evaluations.

Since CellRegMap could not scale to the large datasets, we projected its running time and associated cost based on its performance on smaller datasets, using the following linear model:

$$\log t = \alpha + \beta_{n_{\text{SNP}}} \log n_{\text{SNP}} + \beta_{n_{\text{gene}}} \log n_{\text{gene}} + \beta_{n_{\text{donor}}} \log n_{\text{donor}} + \beta_{n_{\text{cell}}} \log n_{\text{cell}} + \varepsilon, \tag{5}$$

where $\varepsilon \sim N(0, \sigma^2)$. For sceQTL mapping, all data points for CellRegMap in **Fig. 2a** were used to train this model and project the running times for CellRegMap in **Fig. 2b**. For cell type-specific sceQTL mapping, all data points for CellRegMap in **Fig. 3a** were used to train the model and project the running times for CellRegMap in **Fig. 3b**.

SceQTL mapping and cell type-specific sceQTL mapping statistical performance benchmarking

We used the datasets described in "Simulated datasets for benchmarking" for statistical performance benchmarking. We benchmarked the estimation accuracy of sceQTL effect sizes and cell type-specific sceQTL effect sizes with bias and variance, following [5]. Briefly, bias characterizes the overall over- or under-estimation of effect sizes and can be quantified in the linear model:

$$y = \beta x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \tag{6}$$

where y represents the estimated effect sizes, and x denotes the ground-truth effect sizes for cis-eQTLs in the simulated dataset. Therefore, β reflects the estimated effect size scale. The maximum likelihood estimator for β was computed as $\hat{\beta}$, and its deviation from 1 indicates systematic bias. Variance was estimated as the deviation of Pearson R from 1 between the estimated and ground-truth effect sizes. Given that expression-dependent underestimation of effect sizes is commonly observed in single-cell differential expression [5], a simpler version of sceQTL mapping, we performed this benchmarking both for all genes together and for genes stratified by expression level (pseudo-bulk logCPM) to assess the severity of this problem in sceQTL mapping and cell type-specific sceQTL mapping.

To benchmark the specificity of each method, we separately compared the sceQTL P-value distribution for non-eQTLs or non-specific sceQTLs against the uniform distribution. To assess the sensitivity of each sceQTL mapping method, we compared the P-value distribution of eQTLs against those of non-eQTLs. Similarly, to evaluate the sensitivity of each cell type-specific sceQTL mapping method, we compared the P-value distribution of cell type-specific eQTLs against those of non-eQTLs and non-specific eQTLs.

Type I error evaluations

We drew P-value histograms for sceQTL mapping or cell type-specific sceQTL mapping P-values for SNP-gene pairs categorized as non-eQTLs (SNP has no effect on gene expression), eQTLs (SNP has a nonzero effect on gene expression), non-specific eQTLs (SNP has a nonzero effect on gene expression independent of cell type), or cell type-specific eQTLs (SNP has a nonzero effect on gene expression dependent on cell type). The ground-truth eQTL status of these pairs was determined in the simulation data. Each histogram was divided into 20 equal bins from 0 to 1. Error bars were estimated as $2\sqrt{N}$, where N is the number of entries in each bin. Dashed lines represent the uniform distribution.

We used two-sided Kolmogorov-Smirnov test against standard uniform distribution to evaluate the specificity of each method. We also used Storey's π_1 estimated by fdrtool [55] to characterize the proportion of false positives identified by each method or configuration in fully null datasets.

State-specific cGRN inference and analysis

Population admixture detection and removal

We performed a Pearson correlation test between SNP pairs located on different chromosomes to avoid correlation due to linkage disequilibrium. To accelerate computation, only 100,000 post-QC SNPs were randomly sampled for this test. As a baseline, we also performed this test after permuting the donors separately for each SNP.

To test the effect of population admixture removal on genotype correlation, we introduced different numbers of top PCs of the genotype matrix for the tested SNPs as covariates for the (conditional) Pearson test.

To test the effect of population admixture removal on trans-sceQTL mapping, we used GRM estimated from the genotype matrix and/or introduced its the top principal component from the genotype matrix as a covariate without LD pruning.

Comparison of cell state-specific cGRNs

The overlap rate was first computed as the ratio between the number of overlapping gene regulations and the minimum number of gene regulations between the two cGRNs being compared. Overlapping gene regulations between two cell states were defined as those present in both cGRNs, irrespective of effect size or directionality estimation. Any deviation of the overlap rate from one may reflect contributions from differences in the cGRNs, differences in the detectability of individual gene regulations, and data randomness due to the finite number of cells.

Among the overlapping gene regulations, three metrics were computed to characterize the actual differences in cGRNs. The directionality difference was defined as the proportion of gene regulations that exhibited opposite signs in the estimated effect sizes. Pearson R was calculated for the estimated effect sizes between the two cell states to assess the overall similarity. Regulation scale was computed as the linear regression coefficient between the estimated effect sizes of the two cell states (inverted if < 1). Since only overlapping gene regulations were considered, these metrics are less influenced by factors other than differences in the cGRNs.

Visualization of cell state-specific cGRNs

We used Cytoscape (3.10.2) for cGRN visualization. Prefused force directed layout was applied on the largest connected component of the cGRN for CD4+ T cells under flu condition using regulation strength as edge weight.

Gene ontology enrichment analysis

For each master regulator, we used GOATOOLS (1.3.11) for GO enrichment analysis on its target genes using all other genes that have a regulator in the cGRN as background. This reduces the bias of background gene selection compared to using all genes or genes expressed in the corresponding cell state because they do not account for cell state specificity of cGRNs. Activated targets (with regulation strength$>$0) and repressed targets (with regulation strength$<$0) were analyzed for GO enrichment separately. To avoid evaluation biases for our expression-based analyses, for all GO analyses we excluded GO evidences that may also have an expression-based origin (IEP, HEP, RCA, TAS, NAS, IC, ND, and IEA). All three GO categories were used. Both over- and under-enrichments were calculated. Gene names were converted with mygene (3.2.2) when needed. Basic ontology (2024-09-08) and human GO association (2024-09-10) files were used.

Query of perturbation experiments

We queried GEO for perturbation experiments involving any of the 10 master regulators in CD4+ T cells under the flu condition. The experiments were required to involve primary CD4+ T cells or their derived cell lines. Perturbations should include overexpression, knockdown, or knockout of the master regulator, along with a control condition, followed by either bulk or single-cell RNA-seq or microarray measurements. Viral infection or exposure was not a requirement. Only target genes that were significantly identified in the perturbation experiment and also found in the inferred cGRN were included for comparison. Perturbation experiments with ≤ 5 overlapped targets were excluded from the comparison.

Conventions

All P-values are two-sided raw unless otherwise specified. For boxplots, boxes extend between 1st and 3rd quartiles. Lines indicate median. Whiskers extend from the box to the farthest data point within 1.5x of the inter-quartile range from the box. Fliers are shown if beyond the whiskers.

Data availability

The Randolph et al population-scale scRNA-seq dataset is publicly available on Zenodo under accession code 4273998. Transcription factor and motif list was downloaded from `https://resources.aertslab.org/cistarget/motif_collections/v10nr_clust_public/snapshots/motifs-v10-nr.hgnc-m0.00001-o0.0.tbl`. The differential expression analysis result of STAT1 perturbation was downloaded from Supplementary Data of [54].

Code availability

Airqtl (0.1.0) is publicly available at `https://github.com/grnlab/airqtl`.

Acknowledgements

LW wish to thank Heping Xu and Kirk Gosik for their early support and Xinghu Qin for helpful discussions.

References

[1] Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nature Methods* **20**, 1368–1378 (2023). URL `https://www.nature.com/articles/s41592-023-01971-3`. Publisher: Nature Publishing Group.

[2] Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023). URL `https://www.nature.com/articles/s41586-022-05688-9`. Publisher: Nature Publishing Group.

[3] Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* 1–8 (2022). URL `https://www.nature.com/articles/s41586-022-05279-8`. Publisher: Nature Publishing Group.

[4] Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods* **20**, 1355–1367 (2023). URL https://www.nature.com/articles/s41592-023-01938-4. Publisher: Nature Publishing Group.

[5] Wang, L. Single-cell normalization and association testing unifying CRISPR screen and gene co-expression analyses with Normalisr. *Nature Communications* **12**, 6395 (2021). URL https://www.nature.com/articles/s41467-021-26682-1.

[6] Brown, B. C., Morris, J. A., Lappalainen, T. & Knowles, D. A. Large-scale causal discovery using interventional data sheds light on the regulatory network architecture of blood traits (2023). URL https://www.biorxiv.org/content/10.1101/2023.10.13.562293v1. Pages: 2023.10.13.562293 Section: New Results.

[7] Rohbeck, M. *et al.* Bicycle: Intervention-Based Causal Discovery with Cycles. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, 209–242 (PMLR, 2024). URL https://proceedings.mlr.press/v236/rohbeck24a.html. ISSN: 2640-3498.

[8] Jiang, L. *et al.* Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens (2024). URL https://www.biorxiv.org/content/10.1101/2024.01.29.576933v2. Pages: 2024.01.29.576933 Section: New Results.

[9] Smith, G. D. & Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33**, 30–42 (2004). URL https://dx.doi.org/10.1093/ije/dyh132. Publisher: Oxford Academic.

[10] Chen, L. S., Emmert-Streib, F. & Storey, J. D. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology* **8**, R219 (2007). URL http://dx.doi.org/10.1186/gb-2007-8-10-r219.

[11] Wang, L. & Michoel, T. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PLOS Computational Biology* **13**, e1005703 (2017). URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005703.

[12] Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L. & Davey Smith, G. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nature Communications* **11**, 185 (2020). URL https://www.nature.com/articles/s41467-019-13921-9. Publisher: Nature Publishing Group.

[13] Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome Open Research* **4**, 186 (2023). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7384151/.

[14] Bankier, S. *et al.* Plasma cortisol-linked gene networks in hepatic and adipose tissues implicate corticosteroid-binding globulin in modulating tissue glucocorticoid action and cardiovascular risk. *Frontiers in Endocrinology* **14** (2023). URL https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1186252/full. Publisher: Frontiers.

[15] Kang, J. B., Raveane, A., Nathan, A., Soranzo, N. & Raychaudhuri, S. Methods and Insights from Single-Cell Expression Quantitative Trait Loci. *Annual Review of Genomics and Human Genetics* **24**, annurev–genom–101422–100437 (2023). URL https://www.annualreviews.org/doi/10.1146/annurev-genom-101422-100437.

[16] Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* 1–9 (2022). URL https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41586-022-04713-1. Publisher: Nature Publishing Group.

[17] Wijst, M. G. P. v. d. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* **50**, 493–497 (2018). URL https://www.nature.com/articles/s41588-018-0089-9.

[18] Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041. URL https://www-science-org.ezp-prod1.hul.harvard.edu/doi/full/10.1126/science.abf3041. Publisher: American Association for the Advancement of Science.

[19] Randolph, H. E. *et al.* Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* **374**, 1127–1133 (2021). URL https://www.science.org/doi/full/10.1126/science.abg0928. Publisher: American Association for the Advancement of Science.

[20] Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nature Reviews Genetics* 1–15 (2023). URL https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/s41576-023-00599-5. Publisher: Nature Publishing Group.

[21] Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019). URL https://doi.org/10.1186/s13059-019-1874-1.

[22] Cao, Y., Kitanovski, S., Küppers, R. & Hoffmann, D. UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nature Biotechnology* **39**, 158–159 (2021). URL https://www.nature.com/articles/s41587-020-00810-6. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 2 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Next-generation sequencing;RNA sequencing Subject_term_id: next-generation-sequencing;rna-sequencing.

[23] Cuomo, A. S. E. *et al.* CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Molecular Systems Biology* **18**, e10663 (2022). URL https://www.embopress.org/doi/full/10.15252/msb.202110663. Publisher: John Wiley & Sons, Ltd.

[24] Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature Genetics* **55**, 1866–1875 (2023). URL https://www.nature.com/articles/s41588-023-01529-1. Publisher: Nature Publishing Group.

[25] Huang, C. *et al.* Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics* **55**, 2056–2059 (2023). URL https://www.nature.com/articles/s41588-023-01574-w. Publisher: Nature Publishing Group.

[26] Sasse, A. *et al.* Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nature Genetics* **55**, 2060–2064 (2023). URL https://www.nature.com/articles/s41588-023-01524-6. Publisher: Nature Publishing Group.

[27] Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011). URL https://www.nature.com/articles/nmeth.1681.

[28] Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012). URL https://www.nature.com/articles/ng.2310.

[29] Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100–106 (2014). URL https://www.nature.com/articles/ng.2876.

[30] Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biology* **20**, 228 (2019). URL https://doi.org/10.1186/s13059-019-1836-7.

[31] Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015). URL https://www.nature.com/articles/ng.3190.

[32] Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012). URL https://academic.oup.com/bioinformatics/article/28/10/1353/213326.

[33] Kumasaka, N. *et al.* Mapping interindividual dynamics of innate immune response at single-cell resolution. *Nature Genetics* **55**, 1066–1075 (2023). URL https://www.nature.com/articles/s41588-023-01421-y. Number: 6 Publisher: Nature Publishing Group.

[34] Zhou, W. *et al.* Efficient and accurate mixed model association tool for single-cell eQTL analysis (2024). URL https://www.medrxiv.org/content/10.1101/2024.05.15.24307317v1. Pages: 2024.05.15.24307317.

[35] Plomin, R., DeFries, J. C. & Loehlin, J. C. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin* **84**, 309–322 (1977). Place: US Publisher: American Psychological Association.

[36] Via, S. & Lande, R. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution* **39**, 505–522 (1985). URL https://doi.org/10.1111/j.1558-5646.1985.tb00391.x.

[37] THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). URL https://www.science.org/doi/10.1126/science.aaz1776. Publisher: American Association for the Advancement of Science.

[38] Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016). URL https://dx.doi.org/10.1093/bioinformatics/btv722. Publisher: Oxford Academic.

[39] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003). URL http://www.pnas.org/content/100/16/9440.

[40] Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLOS Genetics* **14**, e1007309 (2018). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007309. Publisher: Public Library of Science.

[41] Grinde, K. E., Browning, B. L., Reiner, A. P., Thornton, T. A. & Browning, S. R. Adjusting for principal components can induce spurious associations in genome-wide association studies in admixed populations (2024). URL https://www.biorxiv.org/content/10.1101/2024.04.02.587682v1. Pages: 2024.04.02.587682 Section: New Results.

[42] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006). URL https://www.nature.com/articles/ng1847. Publisher: Nature Publishing Group.

[43] Tolomeo, M., Cavalli, A. & Cascio, A. STAT1 and Its Crucial Role in the Control of Viral Infections. *International Journal of Molecular Sciences* **23**, 4095 (2022). URL https://www.mdpi.com/1422-0067/23/8/4095. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[44] Tanabe, Y. *et al.* Cutting Edge: Role of STAT1, STAT3, and STAT5 in IFN-$\alpha\beta$ Responses in T Lymphocytes. *The Journal of Immunology* **174**, 609–613 (2005). URL https://doi.org/10.4049/jimmunol.174.2.609.

[45] Gent, M. v., Sparrer, K. M. J. & Gack, M. U. TRIM Proteins and Their Roles in Antiviral Host Defenses. *Annual Review of Virology* **5**, 385–405 (2018). URL https://www.annualreviews.org/content/journals/10.1146/annurev-virology-092917-043323. Publisher: Annual Reviews.

[46] Hu, M.-M. & Shu, H.-B. Multifaceted roles of TRIM38 in innate immune and inflammatory responses. *Cellular & Molecular Immunology* **14**, 331–338 (2017). URL https://www.nature.com/articles/cmi201666. Publisher: Nature Publishing Group.

[47] Pagani, I., Poli, G. & Vicenzi, E. TRIM22. A Multitasking Antiviral Factor. *Cells* **10**, 1864 (2021). URL https://www.mdpi.com/2073-4409/10/8/1864. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[48] Zhang, M. *et al.* Negative feedback regulation of cellular antiviral signaling by RBCK1-mediated degradation of IRF3. *Cell Research* **18**, 1096–1104 (2008). URL https://www.nature.com/articles/cr2008277. Publisher: Nature Publishing Group.

[49] Drajac, C. *et al.* Control of IFN-I responses by the aminopeptidase IRAP in neonatal C57BL/6 alveolar macrophages during RSV infection. *Mucosal Immunology* **14**, 949–962 (2021). URL https://www.sciencedirect.com/science/article/pii/S1933021922001921.

[50] Wang, X., Zhu, J., Zhang, D. & Liu, G. Ribosomal control in RNA virus-infected cells. *Frontiers in Microbiology* **13** (2022). URL https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.1026887/full. Publisher: Frontiers.

[51] Schor, S. *et al.* The cargo adapter protein CLINT1 is phosphorylated by the Numb-associated kinase BIKE and mediates dengue virus infection. *Journal of Biological Chemistry* **298** (2022). URL https://www.jbc.org/article/S0021-9258(22)00396-9/abstract. Publisher: Elsevier.

[52] Watashi, K. *et al.* Cyclophilin B Is a Functional Regulator of Hepatitis C Virus RNA Polymerase. *Molecular Cell* **19**, 111–122 (2005). URL https://linkinghub.elsevier.com/retrieve/pii/S1097276505013201.

[53] Inturi, R., Mun, K., Singethan, K., Schreiner, S. & Punga, T. Human Adenovirus Infection Causes Cellular E3 Ubiquitin Ligase MKRN1 Degradation Involving the Viral Core Protein pVII. *Journal of Virology* **92**, 10.1128/jvi.01154–17 (2018). URL https://journals.asm.org/doi/full/10.1128/jvi.01154-17. Publisher: American Society for Microbiology.

[54] Weinstock, J. S. *et al.* Gene regulatory network inference from CRISPR perturbations in primary CD4+ T cells elucidates the genomic basis of immune disease. *Cell Genomics* **4** (2024). URL https://www.cell.com/cell-genomics/abstract/S2666-979X(24)00290-8. Publisher: Elsevier.

[55] Strimmer, K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24**, 1461–1462 (2008). URL https://academic.oup.com/bioinformatics/article/24/12/1461/196272.

Supplementary Note for "Airqtl dissects cell state-specific causal gene regulatory networks with efficient single-cell eQTL mapping"

1 Array of Interleaved Repeats

1.1 Definition

Consider D dimensional tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_D}$. Define the indexing operator for tensor $\mathbf{X}$ as $\mathbf{X}[(i,j)] \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{i-1} \times n_{i+1} \times \cdots \times n_D}$ as selecting the j-th entry of the i-th dimension of $\mathbf{X}$, where $i \in \{1, \ldots, D\}$ and $j \in \{1, \ldots, n_i\}$.

Then, we say $\mathbf{X}$ is an Array of Interleaved Repeats (AIR) with $\mathbf{r} \in \mathbb{N}_+^s$ in dimension $d = 1, \ldots, D$ or $\mathbf{X} \in \mathbb{AIR}(d, \mathbf{r})$, subjecting to $\sum_{i=1}^{s} r_i = n_d$, iff $\forall i \in \{1, \ldots, s\}, j \in \{c_i, c_i + 1, \ldots, c_{i+1} - 1\}, X[(d,j)] = X[(d, c_i)]$. Here $\mathbf{c} \equiv (c_1, c_2, \ldots, c_{s+1}) \in \mathbb{N}_+^{s+1}$, where $c_1 = 1$, $(c_2, \ldots, c_{s+1}) = c_1 + \mathrm{cumsum}\, \mathbf{r}$, and cumsum is the cumulative sum.

More generally, we say $\mathbf{X} \in \mathbb{AIR}(d_1, \mathbf{r}_1; d_2, \mathbf{r}_2; \ldots)$ iff $\forall i, \mathbf{X} \in \mathbb{AIR}(d_i, \mathbf{r}_i)$. Here we introduce additional subscripts to each corresponding variable, such as $d, \mathbf{r}, s, \mathbf{c}$. For simplicity, we say $\mathbf{X} \in \mathbb{AIR}(d_1, d_2, \ldots, \mathbf{r})$ iff $\mathbf{X} \in \mathbb{AIR}(d_1, \mathbf{r}; d_2, \mathbf{r}; \ldots)$.

1.2 Implementation

AIR is implemented as a compressed data structure for torch.Tensor. For $\mathbf{X} \in \mathbb{AIR}(d_1, \mathbf{r}_1; d_2, \mathbf{r}_2; \ldots)$ where $d_1 > d_2 > \ldots$, its AIR includes the following elements:

- $\widetilde{\mathbf{X}}$: compressed D dimensional data as torch.Tensor, where $\forall i_1 \in \{1, \ldots, s_1\}, i_2 \in \{1, \ldots, s_2\}, \ldots$ for each s_j defined as s for the corresponding i_j as i in **Section. 1.1**, $\widetilde{\mathbf{X}}[(d_1, i_1)][(d_2, i_2)] \cdots = \mathbf{X}[(d_1, c_{1,i_1})][(d_2, c_{2,i_2})] \ldots$
- $\mathbf{d} = (d_1, d_2, \ldots)$: the dimensions of $\widetilde{\mathbf{X}}$ containing repeats. Each $d_i \in \mathbb{N}_+$.
- $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots)$: the numbers of repeats for each value in $\widetilde{\mathbf{X}}$. Each $\mathbf{r}_i \in \mathbb{N}_+^{s_i}$.
- $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2, \ldots)$: the cumulative sum precomputed from each $\mathbf{r}_i$. Each $\mathbf{c}_i \in \mathbb{N}_+^{s_i+1}$.

1.3 Example in sceQTL mapping

AIR can be used to efficiently store genotype data for individual cells in population-scale scRNA-seq and perform calculations. For a population-scale scRNA-seq study involving n_{cell} cells from n_{donor} donors, conventional

Supplementary Table 1: Time complexity comparison of computing addition $\mathbf{X}+\mathbf{Y}$ depending on whether AIR-based acceleration is applied. Here $\mathbf{r} \in \mathbb{N}_+^s$, $s < n_1$, and $s < n_2$.

$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{X}+\mathbf{Y}$	Without AIR	With AIR
$\mathbb{R}^{n_1 \times n_2}$	$\mathbb{R}^{n_1 \times n_2}$	$\mathbb{R}^{n_1 \times n_2}$	$\mathcal{O}(n_1 n_2)$	$\mathcal{O}(n_1 n_2)$
$\mathbb{R}^{n_1 \times n_2} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}^{n_1 \times n_2}$	$\mathbb{R}^{n_1 \times n_2}$	$\mathcal{O}(n_1 n_2)$	$\mathcal{O}(n_1 n_2)$
$\mathbb{R}^{n_1 \times n_2} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}^{n_1 \times n_2} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}^{n_1 \times n_2} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathcal{O}(n_1 n_2)$	$\mathcal{O}(s n_2)$
$\mathbb{R}^{n_1 \times n_2}$	$\mathbb{R}$	$\mathbb{R}^{n_1 \times n_2}$	$\mathcal{O}(n_1 n_2)$	$\mathcal{O}(n_1 n_2)$
$\mathbb{R}^{n_1 \times n_2} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}$	$\mathbb{R}^{n_1 \times n_2} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathcal{O}(n_1 n_2)$	$\mathcal{O}(s n_2)$

Supplementary Table 2: Time complexity comparison of computing multiplication $\mathbf{XY}$ depending on whether AIR-based acceleration is applied. Here $\mathbf{r} \in \mathbb{N}_+^s$, $s < n_X$, and $s < n_3$.

$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{XY}$	Without AIR	With AIR
$\mathbb{R}^{n_X \times n_3}$	$\mathbb{R}^{n_3 \times n_Y}$	$\mathbb{R}^{n_X \times n_Y}$	$\mathcal{O}(n_X n_Y n_3)$	$\mathcal{O}(n_X n_Y n_3)$
$\mathbb{R}^{n_X \times n_3} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}^{n_3 \times n_Y}$	$\mathbb{R}^{n_X \times n_Y} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathcal{O}(n_X n_Y n_3)$	$\mathcal{O}(s n_Y n_3)$
$\mathbb{R}^{n_X \times n_3} \cap \mathbb{AIR}(2, \mathbf{r})$	$\mathbb{R}^{n_3 \times n_Y}$	$\mathbb{R}^{n_X \times n_Y}$	$\mathcal{O}(n_X n_Y n_3)$	$\mathcal{O}(n_X n_Y s + n_Y n_3)$
$\mathbb{R}^{n_X \times n_3}$	$\mathbb{R}^{n_3 \times n_Y} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}^{n_X \times n_Y}$	$\mathcal{O}(n_X n_Y n_3)$	$\mathcal{O}(n_X n_Y s + n_X n_3)$
$\mathbb{R}^{n_X \times n_3} \cap \mathbb{AIR}(2, \mathbf{r})$	$\mathbb{R}^{n_3 \times n_Y} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}^{n_X \times n_Y}$	$\mathcal{O}(n_X n_Y n_3)$	$\mathcal{O}(n_X n_Y s)$
$\mathbb{R}^{n_X \times n_3} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathbb{R}$	$\mathbb{R}^{n_X \times n_3} \cap \mathbb{AIR}(1, \mathbf{r})$	$\mathcal{O}(n_X n_3)$	$\mathcal{O}(s n_3)$

methods need to expand the genotype matrix from donor dimension to cell dimension before computing its covariance with expression matrix. The expanded genotype matrix is $\mathbf{X} \in \mathbb{R}^{n_{\text{SNP}} \times n_{\text{cell}}}$ where n_{SNP} is the number of SNPs. Assuming the cells are ordered by donor, genotype matrix is an AIR because $\mathbf{X} \in \mathbb{AIR}(2, \mathbf{r})$ with $\mathbf{r}$ the number of cells from each donor whose $\widetilde{\mathbf{X}} \in \mathbb{R}^{n_{\text{SNP}} \times n_{\text{donor}}}$.

1.4 Acceleration in matrix operations

Here we provide several non-exhaustive examples of how AIR can accelerate basic matrix operations. Linear mixed models used for sceQTL mapping employ addition, multiplication, and singular value decomposition (SVD) on the genotype matrix or its self-product. Therefore, we focus on these operations in the following examples. For simplicity, we assume $\mathbf{X} \in \mathbb{AIR}(1, \mathbf{r})$ and $\mathbf{Y} \in \mathbb{AIR}(2, \mathbf{r})$.

1.4.1 Addition

For AIR, because addition can be operated directly on the compressed data, the time complexity can be improved if the other operand has the same AIR structure or is a scalar (**Tab. 1**).

1.4.2 Multiplication

In addition to the accelerations allowed in addition, multiplication also benefits from AIR when the other operand is not an AIR. For example, when $\mathbf{X} \in \mathbb{R}^{n_X \times n_3} \cap \mathbb{AIR}(2, \mathbf{r})$ and $\mathbf{Y} \in \mathbb{R}^{n_3 \times n_Y}$, it is obvious that $\mathbf{XY} = \widetilde{\mathbf{X}}\mathbf{Y}^{(\mathsf{R})}$, where each row of $\mathbf{Y}^{(\mathsf{R})} \in \mathbb{R}^{s \times n_Y}$ is $\mathbf{Y}_i^{(\mathsf{R})} = \sum_{j=c_i}^{c_{i+1}-1} \mathbf{Y}_j$. In this case, the time complexity would be improved to $\mathcal{O}(n_X n_Y s + n_Y n_3)$ from $\mathcal{O}(n_X n_Y n_3)$. Potential accelerations in other multiplication scenarios are shown in **Tab. 2**.

1.4.3 Singular value decomposition

Consider matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} \cap \mathbb{AIR}(1, \mathbf{r}_1; 2, \mathbf{r}_2)$. In other words,

$$\mathbf{X} = \begin{pmatrix} \tilde{x}_{1,1}\mathbf{1}_{r_{1,1}\times r_{2,1}} & \tilde{x}_{1,2}\mathbf{1}_{r_{1,1}\times r_{2,2}} & \dots & \tilde{x}_{1,s}\mathbf{1}_{r_{1,1}\times r_{2,s_2}} \\ \tilde{x}_{2,1}\mathbf{1}_{r_{1,2}\times r_{2,1}} & \tilde{x}_{2,2}\mathbf{1}_{r_{1,2}\times r_{2,2}} & \dots & \tilde{x}_{2,s}\mathbf{1}_{r_{1,2}\times r_{2,s_2}} \\ \dots & \dots & \dots & \dots \\ \tilde{x}_{s,1}\mathbf{1}_{r_{1,s}\times r_{2,1}} & \tilde{x}_{s,2}\mathbf{1}_{r_{1,s}\times r_{2,2}} & \dots & \tilde{x}_{s,s}\mathbf{1}_{r_{1,s}\times r_{2,s}} \end{pmatrix}, \tag{7}$$

where $\mathbf{1}_{i\times j}$ is a $i \times j$ matrix with all entries equal to 1. Its SVD can be performed as:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*, \tag{8}$$

where $\mathbf{\Lambda}$ is a rectangular diagonal matrix with diagonal elements $(\lambda_1, \lambda_2, \dots, \lambda_{\min(n_1,n_2)})$. This standard approach has time complexity $\mathcal{O}(n_1 n_2 \min(n_1, n_2))$.

Acceleration of SVD arises from the duplicate rows and columns in $\mathbf{X}$ due to its repeats. First, $\mathbf{X}$ only has (up to) $\min(s_1, s_2)$ nonzero eigenvalues. This allows a low-rank SVD but that would only reduce the time complexity to $\mathcal{O}(n_1 n_2 \min(s_1, s_2))$.

Additionally, we can perform an equivalent eigendecomposition on the rescaled compressed data that accounts for duplicate rows and columns:

$$\widetilde{\widetilde{\mathbf{X}}} = \left(\widetilde{\widetilde{x}}_{i,j}\right) \equiv \left(\sqrt{r_{1,i} r_{2,j}}\, \tilde{x}_{i,j}\right) = \widetilde{\widetilde{\mathbf{U}}}\widetilde{\widetilde{\mathbf{\Lambda}}}\widetilde{\widetilde{\mathbf{V}}}^*, \tag{9}$$

where similarly, $\widetilde{\widetilde{\mathbf{\Lambda}}} = \mathrm{diag}(\tilde{\tilde{\lambda}}_1, \tilde{\tilde{\lambda}}_2, \dots, \tilde{\tilde{\lambda}}_{\min(s_1,s_2)})$, $\widetilde{\widetilde{\mathbf{U}}} \equiv (\tilde{\tilde{u}}_{i,j})$, $\widetilde{\widetilde{\mathbf{V}}} \equiv (\tilde{\tilde{v}}_{i,j})$, and * means conjugate transpose. It is easy to see that $\mathbf{\Lambda} = \widetilde{\widetilde{\mathbf{\Lambda}}}$ (for nonzero entries). In addition, $\mathbf{U} \in \mathbb{R}^{n_1 \times \min(s_1,s_2)} \cap \mathbb{AIR}(1, \mathbf{r}_1)$ has its compressed data $\widetilde{\mathbf{U}} \equiv (\tilde{u}_{i,j}) = (\tilde{\tilde{u}}_{i,j}/\sqrt{r_{1,i}})$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times \min(s_1,s_2)} \cap \mathbb{AIR}(1, \mathbf{r}_2)$ has its compressed data $\widetilde{\mathbf{V}} \equiv (\tilde{v}_{i,j}) = (\tilde{\tilde{v}}_{i,j}/\sqrt{r_{2,i}})$. Therefore, AIR acceleration allows SVD to be directly performed on compressed data and further

reduces its time complexity to $\mathcal{O}(s_1 s_2 \min(s_1, s_2))$.

2 Supplementary Data

1. Supplementary Data 1. KS test statistics when comparing each method's sceQTL mapping P values between null and non-null cases or against standard uniform distribution.

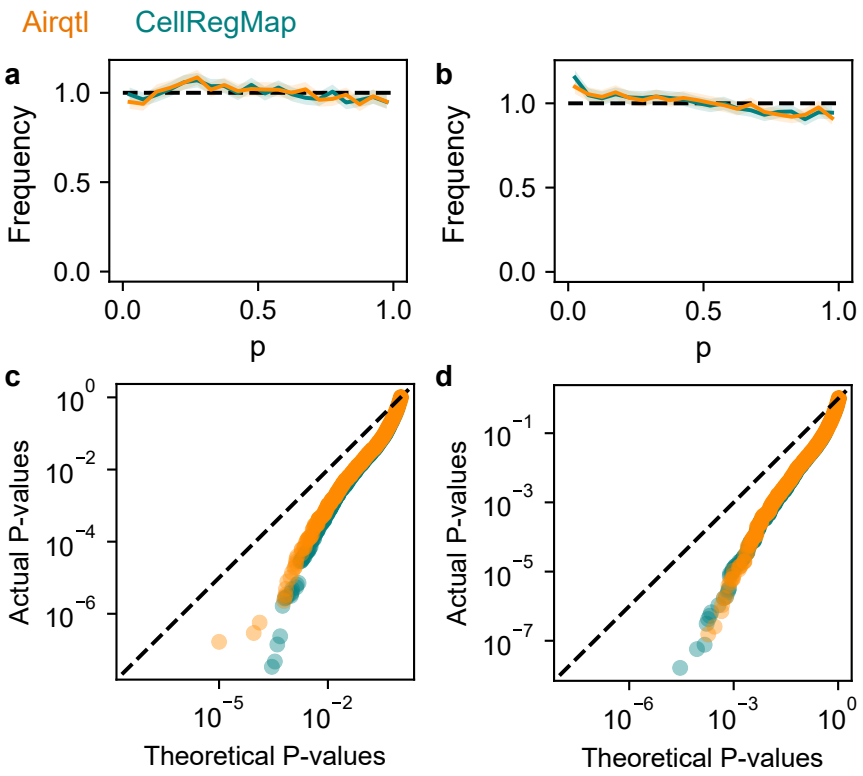
2. Supplementary Data 2. KS test statistics when comparing airqtl's cell type-specific sceQTL mapping P values between null (non-eQTLs and non-specific eQTLs separately) and non-null (specific eQTLs) cases or against standard uniform distribution.

3. Supplementary Data 3. SceQTL mapping summary statistics for each cell state in the Randolph et al dataset, with cutoffs first on raw P$< 10^{-4}$ for trans-sceQTLs and then BH Q$<$0.1 separately for cis- and trans-sceQTLs

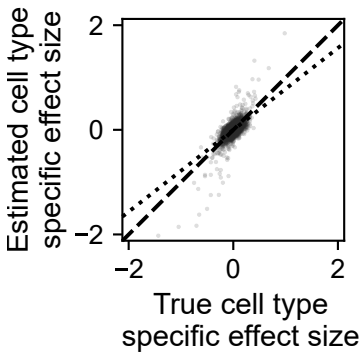
4. Supplementary Data 4. Node and edge properties of the inferred cell state-specific cGRNs

5. Supplementary Data 5. GO enrichment results for the target genes of each master regulator of CD4+ T cells under flu or NI condition, separately for activation and repression targets.

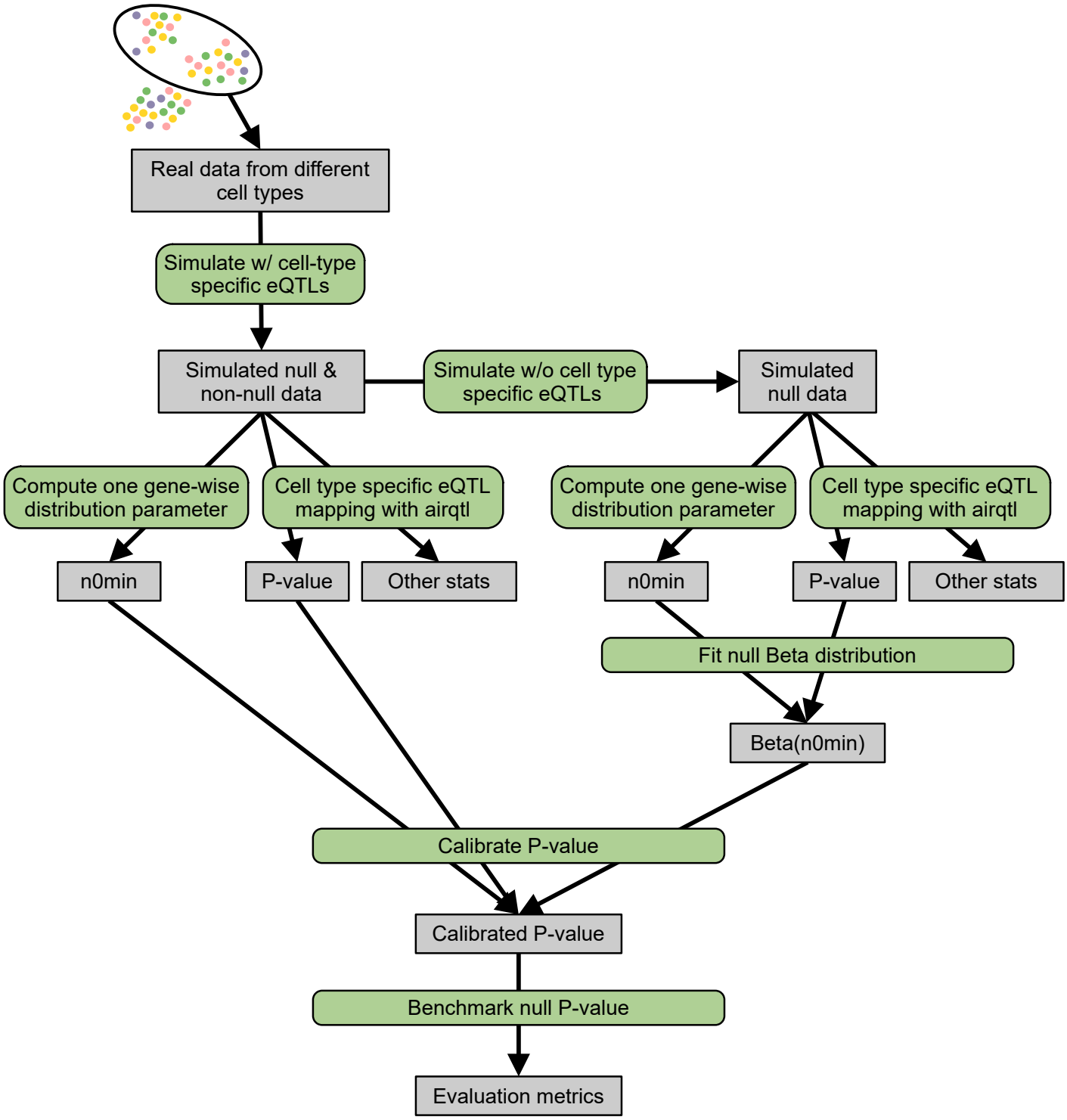
3 Supplementary Figures



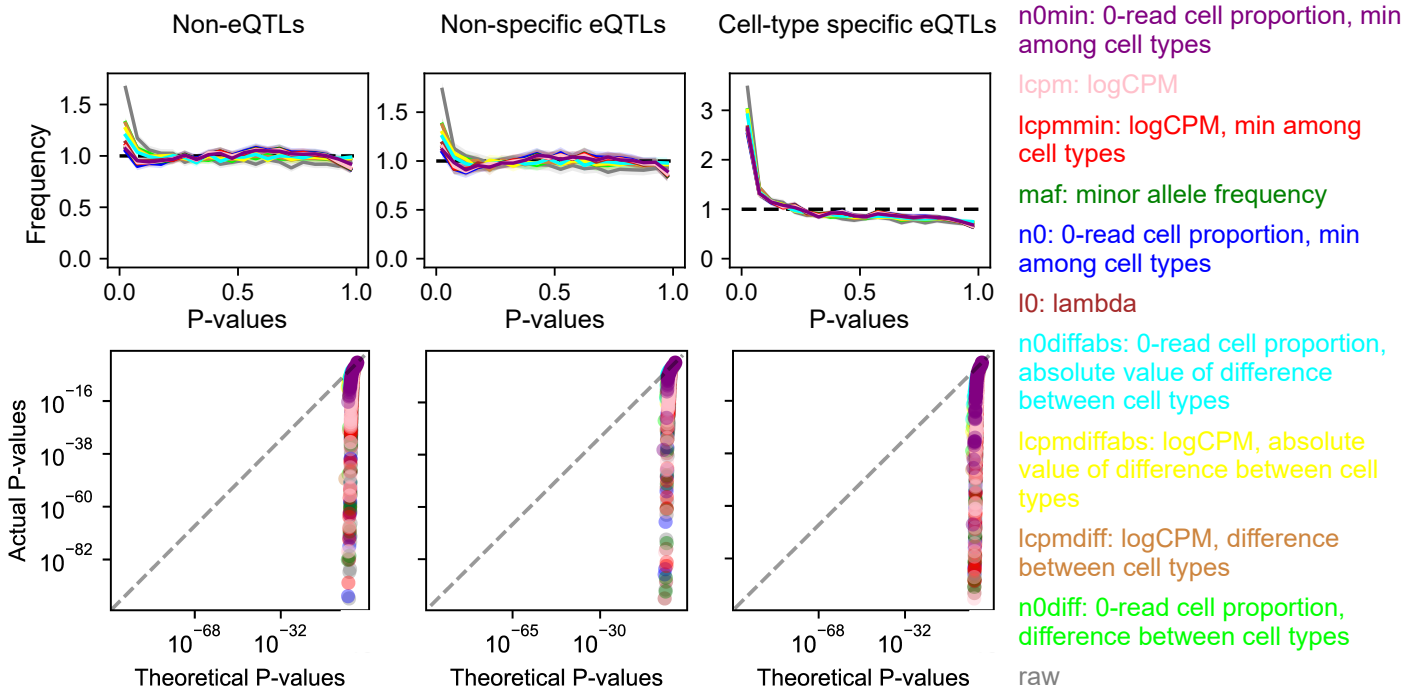
Supplementary Figure 1: **Sensitivity and specificity benchmarking for sceQTL mapping.** P-value distribution histograms (**ab**) and quantile-quantile plots (**cd**) for non-eQTLs (**ac**) and eQTLs (**bd**) as computed by different methods (color) from simulated data. Error bars in histograms were estimated as $2\sqrt{N}$ where N is the number of entries in each bin. Dashed lines indicate uniform distribution.



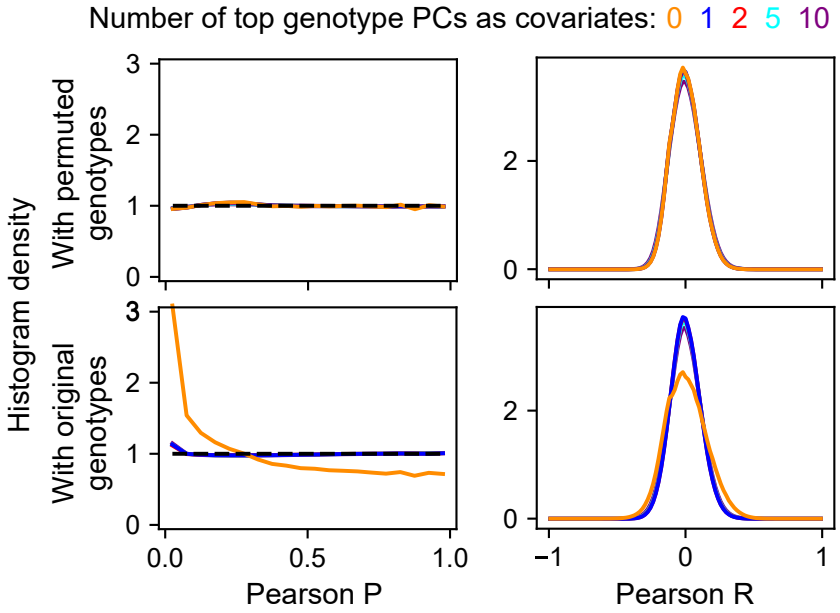
Supplementary Figure 2: **Airqtl provided accurate estimation of cell type-specific eQTL effect sizes.** Scatter plot of estimated (Y) and true (X) eQTL cell type-specific effect sizes as shown by individual SNP-gene pairs (dot) and best linear fit (dotted line). The dashed line indicates perfect performance with Y=X.



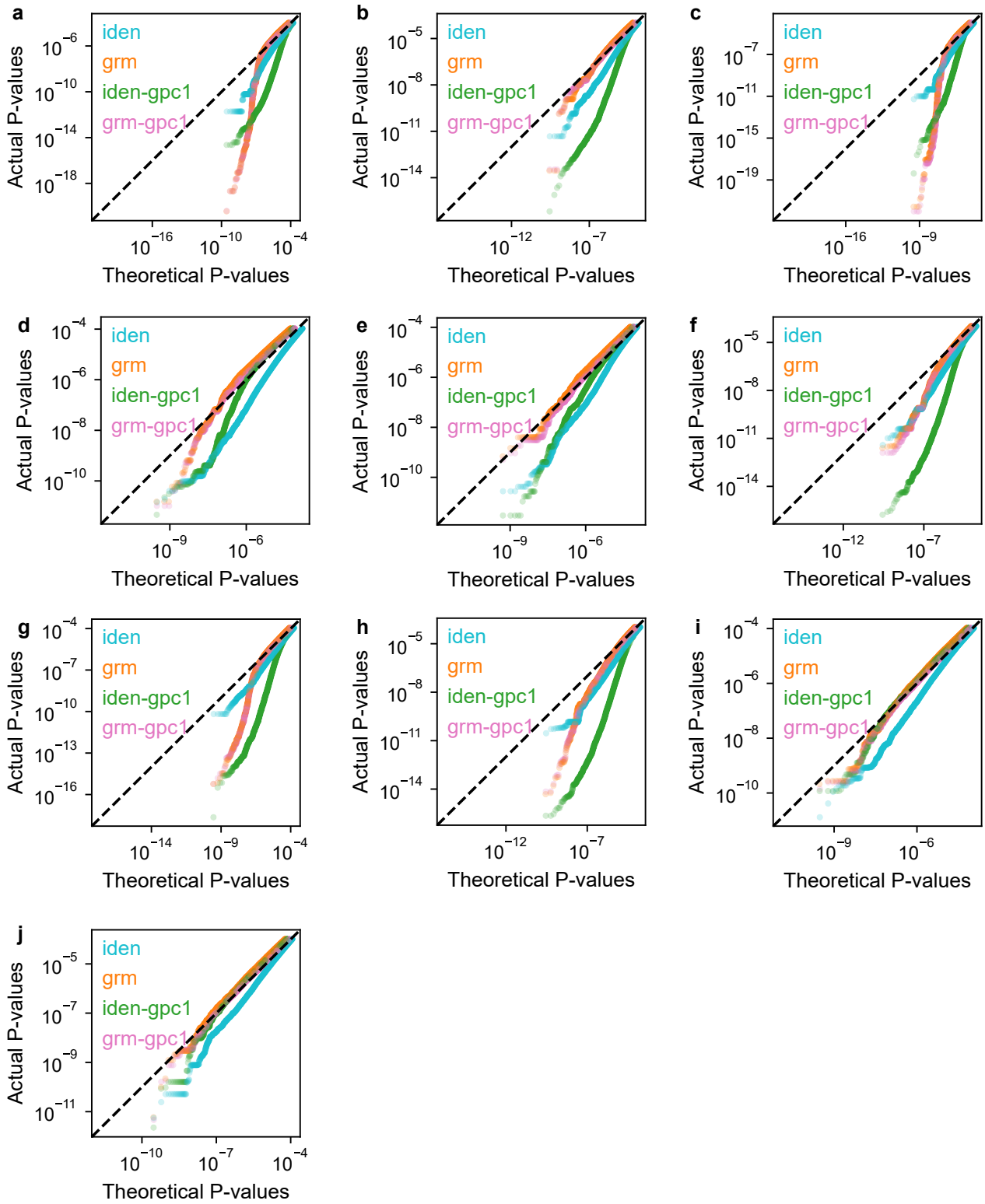
Supplementary Figure 3: **Workflow to objectively optimize parameter choice for P-value calibration.** Each candidate parameter was separately considered with this workflow to produce evaluation metrics. This workflow demonstration uses n0min as an example but other parameters (e.g. lcpm) followed the same workflow and use the same simulated data. Candidate parameter with the best evaluation results was selected for P-value calibration for this dataset (**Methods**).



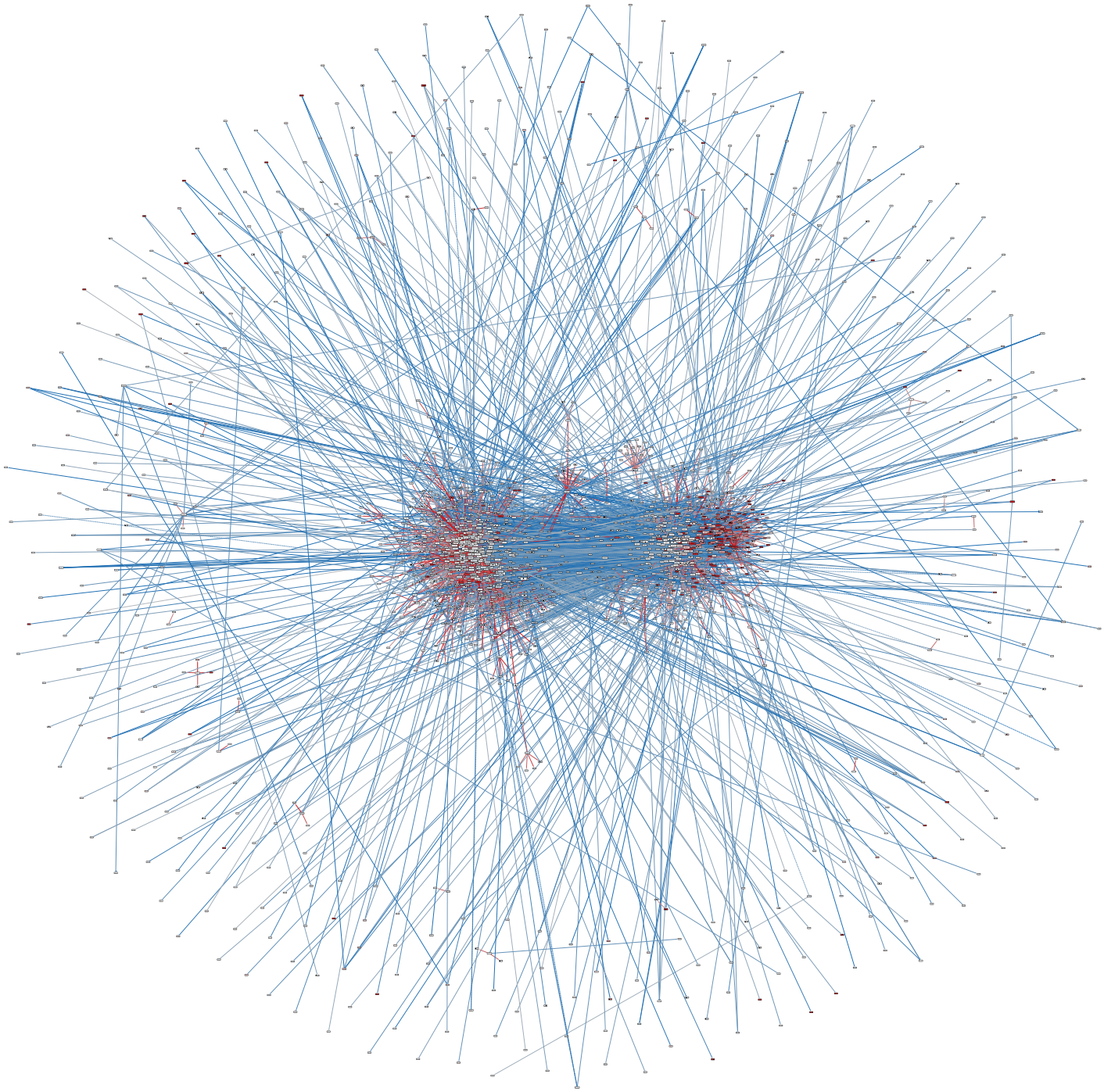
Supplementary Figure 4: **Objective optimization of cell type-specific sceQTL mapping through sensitivity and specificity benchmarking.** P-value distribution histograms (top) and quantile-quantile plots (bottom) for non-eQTLs (left), non-specific eQTLs (center), and cell type-specific eQTLs (right) as calibrated by different models, each with a different calibration parameter (color and legend, ordered by KS test P of non-specific eQTLs against uniform distribution). Raw P values were computed from the vanilla model with no calibration. Error bars in histograms were estimated as $2\sqrt{N}$ where N is the number of entries in each bin. Dashed lines indicate uniform distribution.



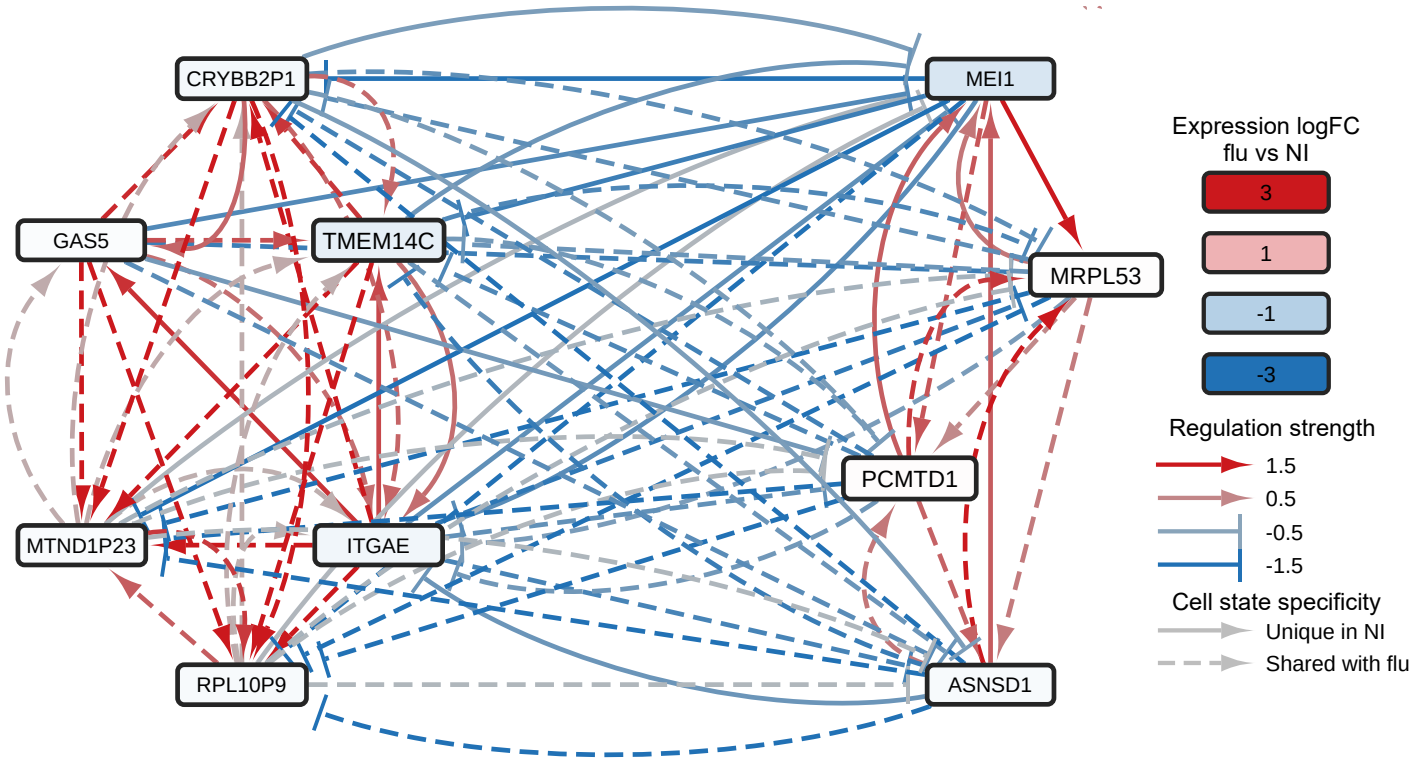
Supplementary Figure 5: **Pearson test for correlation between genotype pairs on different chromosomes reveals the best number of genotype PCs to model their covariance from population structure in Randolph et al dataset.** Pearson P-value (left) and R (right) distribution histograms for genotype pairs on different chromosomes to reflect correlation between them due to population structure and the best parameter to account for them. Tests were performed separately for original genotypes (bottom) and genotypes that were independently randomly permuted between donors (top). Different numbers of principal components computed from all genotypes were used as covariates (color). One genotype PC was the minimum number that provided the best removal of population structure effects.



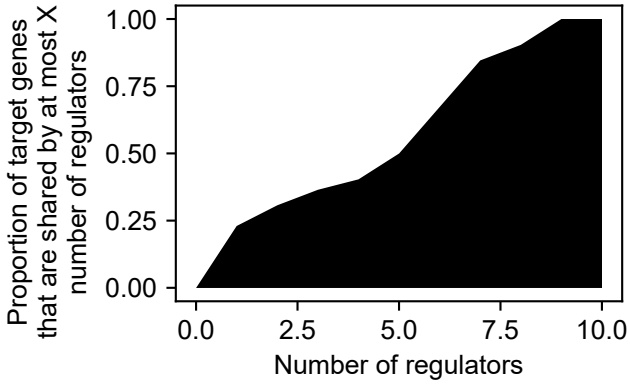
Supplementary Figure 6: **QQ plots of null trans-sceQTL P-values by different methods to accounting for admixed populations.** Panel corresponds to cell states, including B cells, flu (**a**); B cells, NI (**b**); CD8+ T cells, flu (**c**); CD8+ T cells, NI (**d**); highly infected cells, flu (**e**); infected monocytes, flu (**f**); monocytes, flu (**g**); monocytes, NI (**h**); NK cells, flu (**i**); and NK cells, NI (**j**). QQ plots for CD4+ T cells are in **Fig. 5c**. Color indicates method, including identity donor-level GRM (iden), GRM estimated from genotypes (grm), identity donor-level GRM with 1 top genotype PC as covariate (iden-gpc1), and GRM estimated from genotypes with 1 top genotype PC as covariate (grm-gpc1).



Supplementary Figure 7: **Inferred cGRN for CD4+ T cells under flu condition.** Only the largest connected component is shown. Node: gene. Edge: gene regulation. Node color: logFC in expression flu v.s. NI (red: positive; blue: negative; white: zero). Edge color: gene regulation effect size under flu condition (red: positive; blue: negative). See **Supplementary Data 4** for cGRN data.



Supplementary Figure 8: **Sub-cGRN between top master regulators in CD4+ T cells under NI condition.**



Supplementary Figure 9: **Master regulators had unique target genes not shared with other master regulators.** A large fraction (∼20%, Y) of "defense response to virus" target genes of master regulators were unique to one master regulator in CD4+ T cells under flu condition.