# distQTL: Distribution Quantitative Trait Loci Identification by Population-Scale Single-Cell Data

Alexander Coulter,[1] Chun Yip Tong,[2] Yang Ni[1,3,*] and Yuchao Jiang[1,2,4,*]

[1]Department of Statistics, College of Arts and Sciences, Texas A&M University, College Station, TX 77843, United States, [2]Department of Biology, College of Arts and Sciences, Texas A&M University, College Station, TX 77843, United States, [3]Department of Statistics and Data Sciences, College of Natural Sciences, The University of Texas at Austin, Austin, TX 78705, United States and [4]Department of Biomedical Engineering, College of Engineering, Texas A&M University, College Station, TX 77843, United States.
*Co-corresponding authors: yangni87@gmail.com, yuchaojiang@tamu.edu.

## Abstract

Mapping expression quantitative trait loci (eQTLs) is a powerful method to study how genetic variation influences gene expression. Traditional bulk eQTL methods rely on averaged gene expression across a possibly heterogeneous mixture of cells, which can obscure underlying regulatory heterogeneity. Single-cell eQTL methods circumvent the averaging artifacts, providing an immense opportunity to interrogate transcriptional regulation at a much finer resolution. Recent developments in metric space regression methods allow the use of full empirical distributions as response objects instead of simple summary statistics such as mean. Here, we leverage Fréchet regression to identify distribution QTLs (distQTLs) using population-scale single-cell RNA sequencing data. We apply distQTL to the OneK1K cohort, consisting of scRNA-seq data of peripheral blood mononuclear cells from 982 donors, and compare results to various eQTL approaches based on summary statistics and mixed effects modeling. We demonstrate the superior performance of distQTL across different gene expression contexts compared to other methods and benchmark our results against findings from the Genotype-Tissue Expression Project. Finally, we orthogonally validate calls from distQTL using cell-type-specific epigenomic profiles.

**Key words:** quantitative trait loci, single-cell RNA sequencing, Fréchet regression, gene expression distribution.

## Introduction

Expression quantitative trait loci (eQTLs) link single-nucleotide polymorphisms (SNPs) to gene expressions, offering insights into the molecular mechanisms of genetic variants that are associated with complex traits (Mackay et al., 2009). High-throughput RNA sequencing (RNA-seq) data, paired with DNA genotype information, has achieved huge successes at eQTL-mapping, such as the identification of tissue-specific eQTLs from the Genotype-Tissue Expression (GTEx) Project (The GTEx Consortium, 2020). These studies, however, were at the bulk-tissue level, where one could only observe averaged expression across a possibly heterogeneous mixture of cells, and the true underlying cellular heterogeneity was attenuated.

Recent advancements in single-cell RNA-seq (scRNA-seq), along with its increasing accessibility and decreasing cost, have facilitated population-scale transcriptomic profiling at the cellular level, enabling the discovery of cell-type-specific eQTLs (van der Wijst et al., 2018; Neavin et al., 2021; Jerber et al., 2021; Nathan et al., 2022). For instance, the OneK1K cohort (Yazar et al., 2022) consists of scRNA-seq data of ∼1.27 million peripheral blood mononuclear cells (PBMCs) collected from 981 donors, resulting in the discovery of novel cell-type-specific eQTLs. See recent reviews (Maria et al., 2022; Luo et al., 2023) and database (Ding et al., 2024) for existing scRNA-seq data utilized for eQTL discovery.

Existing eQTL studies for scRNA-seq, while revelatory, most commonly identify cell clusters and annotate cell types from the single-cell data first, and then calculate average expression per individual for each cell type by constructing pseudo-bulk samples (Neavin et al., 2021; Elorbany et al., 2022). This partially forfeits the purpose of single-cell sequencing, which goes beyond the mean measurements to characterize the distribution of gene expression. Although methods for analyzing higher-order distribution features of gene expression have been developed, including variance QTL (vQTL) detection (Corty and Valdar, 2018; Hong et al., 2017), they have been largely confined to bulk eQTL-mapping. One notable exception is Sarkar et al. (2019), where zero-inflated negative binomial distributions were fit to donor-level scRNA-seq data, giving mean, variance, and over-dispersion estimates. However, in applied data analysis they found no vQTLs independent of mean QTLs, owing to the well established mean-variance dependence in scRNA-seq data (Eling et al., 2018).

Beyond pseudo-bulk approaches, attempts to account for cell-cell correlations have been made with mixed-effects Poisson regression models (Nathan et al., 2022). However, its high computational cost and non-convergence issue make it practically infeasible to scan every gene-SNP pair for data sets as large as the OneK1K cohort. Additionally, introducing an individual-specific random effect has been shown to result in a lack of power (Neavin et al., 2021). Finally, the class of mixed-effects models are only concerned with mean effects, missing out on higher-order distribution features.

In this paper, we propose a statistical method based on Fréchet regression (Petersen and Müller, 2019) and associated asymptotic partial F-test (Shen and Faraway, 2004; Petersen et al., 2021) to identify distribution QTLs (distQTLs). We directly use the empirical distribution of library-size corrected gene expression at the donor level as the response. Characterizing gene expression distribution broadens inference beyond the first and second moments (i.e. mean and variance) from scRNA-seq, while capturing expression level, variability, and stochasticity in a unified manner (Jiang et al., 2017;

Larsson et al., 2019). distQTL-mapping also uses a fraction of the time required for Poisson mixed-effects regression (less than one-tenth of a second per model, contrasted to minutes), enabling its use in large-scale application.

We compare distQTLs with baseline eQTL-mapping using mean ($\mu$QTL), variance (vQTL), coefficient of variation (cvQTL), and bursting fraction (percentage of cells with nonzero expression, bQTL) as outcomes. We apply a correction to distQTL (F-test) and baseline eQTL (Wald $t$-test) p-values to account for misspecified null distribution. We demonstrate our proposed framework in the OneK1K cohort, identifying more cell-type-specific eQTLs than the pseudo-bulk methods. We benchmark against existing reports to assess performance and further evaluate type I error control via permutation. Altogether, this work provides an atlas of how genetic variation influences gene expression across cell types with various statistical testing schemes that corroborate and complement existing studies and approaches. distQTL is an open-source R package available at `https://github.com/alexandercoulter/distQTL`.

## Materials and Methods

### QTL Identification

For each gene-SNP pair, we regress cell-type-specific gene expression against SNP genotype while controlling for a set of fixed donor-level covariates. We index donors by $1 \leqslant n \leqslant N$, cell-type groups by $1 \leqslant k \leqslant K$, individual donor and group level cells by $1 \leqslant c_{n,k} \leqslant C_{n,k}$, genes by $1 \leqslant g \leqslant G$, and *cis*-SNPs by $1 \leqslant s_g \leqslant S_g$. Denote by $Y_{g,c_{n,k}}$ the raw gene expression for cell $c_{n,k}$, i.e., from donor $n$ and cell-type group $k$. Similarly, denote by $\widetilde{Y}_{g,c_{n,k}}$ the library size corrected gene expression for cell $c_{n,k}$. Let $\mathbf{z}_n$ be the fixed donor-level covariates, including quadratic orthogonal polynomials of age (`age` and `age`$^2$), sex (`male` and `female`), and the top leading genotype principal components; see Discussion for details on additional covariates such as the probabilistic estimation of expression residuals (PEER) factors. Let $x_{n,s_g}$ be the individual SNP genotype, also included as a covariate in all regression models. All regression methods are applied within cell-type group; in the following equations, the $k$ subscript is dropped for simplicity.

#### *Baseline and univariate eQTL*

We calculate donor-level univariate summaries $f(\,\cdot\,)$ for library size corrected gene expression, fitting linear models

$$f(\{\widetilde{Y}_{g,c_n}\}_{c_n=1}^{C_n}) = \mathbf{z}_n^\top \boldsymbol{\alpha}_{g,s_g} + x_{n,s_g}\beta_{g,s_g} + \varepsilon_{n,s_g} \tag{1}$$

where $\mathbf{z}_n$ is the vector of fixed donor-level covariates, $x_{n,s_g}$ is the SNP genotype, and $\varepsilon_{n,s_g} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$. For each donor and gene, the univariate expression summaries include the sample mean, sample variance, proportion of cells with positive gene expression ("bursting"), and sample coefficient of variation. For each expression summary, we test null hypothesis $H_0 : \beta_{g,s_g} = 0$ with a Wald $t$-test. We permute responses for each model and recalculate the p-value, and aggregate these p-values under permutation to perform empirical null-correction on the original p-values; see Appendix. Hereafter, we refer to these four baseline eQTL-mapping methods as $\mu$QTL, vQTL, bQTL, and cvQTL, respectively.

#### *Poisson mixed-effects regression (PMER)*

We model cell-level raw gene expression with PMER (Nathan et al., 2022), fitting a random intercept per donor, utilizing the `glmer` function from the `lme4` R package (Bates et al., 2015). Library correction factors are used as regression offsets to account for differences in gene expression measurement exposure. We use the same fixed effects structure as (1) and test null hypothesis $H_0 : \beta_{g,s_g} = 0$ with Wald $t$-test. Because the computation time of PMER is very high, we implement this procedure only on a subset of gene-SNP pairs, obtained by selecting the top eQTL identified by distQTL within B cells. We do not perform p-value correction by permutation for PMER results since our fixed effects design does not differ within-donor, and there are no meaningful permutations which preserve donor-level cell-cell correlation structure.

#### *Distribution QTL (distQTL)*

We implement Fréchet regression (Petersen and Müller, 2019) with gene- and donor-specific empirical distributions of $\{\widetilde{Y}_{g,c_n}\}_{c_n=1}^{C_n}$. Unlike any of the aforementioned univariate summaries, Fréchet regression aims to model the entire distribution of cell-type-specific gene expression as a function of covariates and *cis*-SNP genotype. Fréchet regression generalizes the Euclidean regression problem, which takes covariate vector $X \in \mathbb{R}^p$ (where $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) = \boldsymbol{\Sigma}$) and response $Y \in \mathbb{R}$ and finds conditional mean

$$\mathbb{E}(Y|X = \mathbf{x}) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}, \tag{2}$$

where $\beta_0$ and $\boldsymbol{\beta}_1$ are respectively intercept and slope parameters. The Euclidean regression can be equivalently formulated without explicit dependence on these parameters

$$\mathbb{E}(Y|X = \mathbf{x}) = \arg\min_{y\in\mathbb{R}} \mathbb{E}\left[\left(1 + X^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\right)(Y-y)^2\right], \tag{3}$$

which is a weighted Euclidean mean problem. In the sample setting $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with plug-in estimator $\widehat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{X}^\top\mathbf{X}$, the sample predictor $\widehat{\mathbf{y}}$ is the familiar orthogonal projection of $\mathbf{y}$ onto the column space of $(1, \mathbf{X})$. Importantly, equation (3) can be generalized to accommodate non-Euclidean responses from a general metric space, such as distributions from the space of univariate quantile functions $\mathcal{Q}$ equipped with 2-Wasserstein metric $d_W(\mathbf{q},\mathbf{p}) := \|\mathbf{q}-\mathbf{p}\|_{L^2(0,1)}$,

$$Q_\oplus(\mathbf{x}) := \arg\min_{\mathbf{q}\in\mathcal{Q}} \mathbb{E}\left[\left(1 + X^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\right) d_W^2(\mathbf{q}, Q)\right].$$

In the sample setting $\{(\mathbf{x}_i, Q_i)\}_{i=1}^n$ with plug-in estimators, the $\mathbf{x}_i$-conditional response estimate is the solution to

$$\widehat{Q}_\oplus(\mathbf{x}_i) := \arg\min_{\mathbf{q}\in\mathcal{Q}} \frac{1}{n}\sum_{j=1}^n \left[1 + \mathbf{x}_j^\top (\mathbf{X}^\top\mathbf{X})^+ \mathbf{x}_i\right] d_W^2(\mathbf{q}, Q_j),$$

where $\mathbf{A}^+$ is the generalized inverse of $\mathbf{A}$. This is a weighted Fréchet mean problem; we refer to Petersen and Müller (2019) for details.

We calculate donor-specific empirical quantile functions (EQFs) per gene and cell-type group, using R function `quantile` (`type = 1` option), on a common $m$-grid in $(0,1)$, with a default $m = 200$. We use the same regressor set as in (1), and test null hypothesis $H_0 : Q_\oplus(\mathbf{z}, x) = Q_\oplus(\mathbf{z})$ with asymptotic partial F-test (Shen and Faraway, 2004; Petersen et al., 2021); see Supplement. We permute responses and refit once for each gene-SNP pair to perform p-value correction as described in the Supplement (Fig. S1). Hereafter, we refer to this eQTL-mapping method as *distQTL*.

## Simulations

We show the capacity of Fréchet regression to identify mean and non-mean effects through different zero-inflated negative binomial (zinbinom) simulation settings. We generate covariate vectors $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_5)$ for $i = 1, \ldots, 100$ samples, and from each generate a `zinbinom` quantile function $Q_i$ as a function only of the first covariate, $x_{i,1}$.

In setting (**A**), we generate each $Q$ such that the zero-inflation is fixed at zero (no inflation), the mean $E(Q) := \int_0^1 Q(p)dp$ is a log-linear function of $x_1$ with Gaussian errors, and the variance $V(Q) := \int_0^1 \{q - E(Q)\}^2 dq$ is held constant. In setting (**B**), the zero-inflation is fixed at zero, $V(Q)$ is a log-linear function of $x_1$ with Gaussian errors, and $E(Q)$ is held constant. In setting (**C**), $E(Q)$ and $V(Q)$ are generated randomly independent of $x_1$, the total positive proportion $\mathrm{P}_+(Q) := \int_0^1 \mathbb{1}\{Q(p) > 0\}dp$ is a logistic function of $x_1$ with Gaussian errors, and the size and probability parameters $(r, p)$ of the `nbinom` distribution component are chosen to satisfy the system of equations relating $E(Q)$, $V(Q)$, and $\mathrm{P}_+(Q)$.

We take univariate summaries (mean, variance, % positive, and CV) to be exactly known for each distribution, and evaluate $Q_i$'s on a common uniform $m = 150$ grid in $(0, 1)$. In each setting, across 1000 simulations and 6 effect sizes of increasing magnitude (starting at zero), we perform Wald $t$-tests (for univariate responses) and asymptotic partial F-tests (for EQF responses, i.e. distQTL) for the first covariate, and record corresponding $-\log_{10}(p)$-values.

## OneK1K Data and Quality Control

We use the OneK1K cohort (Yazar et al., 2022) data set, consisting of scRNA-seq data of ∼1.27 million PBMCs collected from 981 donors of European ancestry, along with corresponding donor-level genotype data. Integer-valued scRNA-seq read count data across 36,571 genes and 1,248,980 cells are downloaded from the Human Cell Atlas (HCA) and corrected for library-size factor. Processed genotype data are provided by the authors of the OneK1K paper. For quality control, we remove: (i) doublets and cells with cell-type prediction scores $< 0.6$ or extreme total coverages (number RNA reads $> 3$ M.A.D. from median); (ii) donors with extreme cell numbers ($> 3$ M.A.D. from median); and (iii) SNPs with low rare allele presence (fewer than five donors with the minor allele). We further filter out each gene with no SNPs within 200Kbp from the gene's transcription start site.

Cell-type labels are obtained from the HCA, and we focus on four major cell types: 102,179 B cells (naïve, memory, and transitional), 42,634 monocytes (CD14-positive, CD14-low, and CD16-positive), 221,802 CD4-positive T cells (central memory and effector memory), and 144,030 CD8-positive T cells (central memory and effector memory). Within each cell-type group, we filter out donors with fewer than 10 cells, and filter out genes with donor-level cell-type-specific nonzero expression rates below 1%.

## Benchmarking and Validation

eQTL findings are validated against the whole blood bulk eQTL results from the GTEx Project (The GTEx Consortium, 2020). eQTL-mapping overlap is illustrated in aggregated cell groups – to make the single-cell data comparable to the bulk data from GTEx – by contour plots, of proportion of OneK1K eQTL matches in GTEx against proportion of *cis*-SNPs that GTEx labels as eQTLs. distQTL is further directly compared with $\mu$QTL by filtering gene-SNP pairs coidentified in GTEx and OneK1K on (a) significant (using $p < 10^{-4}$) distQTL and n.s. $\mu$QTL, and (b) significant $\mu$QTL and n.s. distQTL, showing that distQTL has strongly significant results with a mixture of replication in GTEx. To further assess the performance of distQTL and to benchmark against the other univariate eQTL methods, we adopted cell-type-specific chromatin immunoprecipitation sequencing (ChIP-seq) data of histone modification in human blood cells (Bernstein et al., 2010). Specifically, we compared the ChIP-seq signals for activation and enhancer markers (H3K27ac, H3K4me1, H3K4me3) and repression marker (H3K27me3) at eQTLs identified by distQTL, at eQTLs uniquely detected by distQTL but not by the other methods, and at randomly selected SNPs serving as negative controls.

## Model Fit Run Time

We calculate total time to gene-SNP models in each cell-type group for distQTL and the competing eQTL-mapping methods. For all eQTL-mapping methods except PMER, we fit models to expression data from highly expressed genes in original ordering and with permuted responses, for approximately 5.8 million models across cell-type groups. For PMER, due to long per-model fit times, we select randomly 1000 genes with high expression in each cell type group and fit one model per gene using the top eQTL identified by distQTL in B cells. All model fits were performed splitting genes in parallel over nine cores, and we estimate fit time per model using aggregate run time divided by number of models fit, multiplied by 9.

# Results

## Simulations

We simulate zinbinom distributions in different covariate-dependent settings to demonstrate efficacy of distribution-based Fréchet regression. For $i = 1, \ldots, n$, we sample covariate vectors $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_5)$ and generate response distributions dependent on the first covariate $x_{i,1}$. For baseline eQTL comparison, we take distribution summaries like mean and variance to be known exactly; for distribution responses to apply distQTL, we evaluate distributions as empirical quantile functions on a uniform discrete $m = 150$-grid in $(0, 1)$.

In the first setting (Fig. 1**A**), we let the distribution mean be a log-linear function of $x_1$, with variance held constant. As expected, vQTL does not have statistical power in this setting. On the other hand, for non-vQTL methods we observe the distribution of $-\log_{10}(p)$-values shifts more positive as effect size increases, indicating increasing power with effect magnitude. In the second setting (Fig. 1**B**), we let variance be a log-linear function of $x_1$, with mean held constant. $\mu$QTL does not have statistical power in this setting, while we observe non-$\mu$QTL methods, including distQTL, exhibit increasing power as effect size increases. In the third setting (Fig. 1**C**), we generate random mean and variance parameters independent of covariates, and let the distribution mass placed on positive values be a logistic function of $x_1$ so that small $x_1$ values correspond to more zeros. Neither $\mu$QTL, vQTL, nor cvQTL exhibit statistical power in these settings, as expected, however distQTL exhibits higher statistical power than bQTL. Across all settings, distribution based regression has comparable or superior power to univariate response regression methods, even in specific cases where certain univariate methods exhibit no sensitivity.

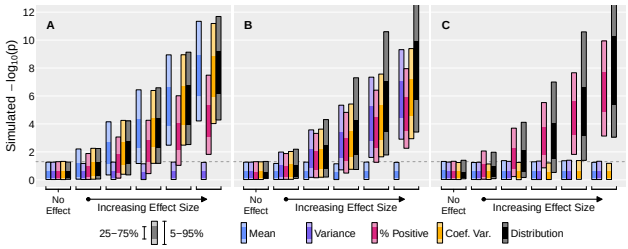


**Fig. 1.** Empirical type I error and power of partial F-tests in simulated covariate-dependent, zero-inflated negative binomial distributions, using different response targets. Sample size fixed at $n = 100$; distributions represented by EQFs evaluated on fixed $m = 150$ grid; covariates fixed at $p = 5$, where only one covariate has varying effect size on (**A**) mean but not variance, (**B**) variance but not mean, and (**C**) positive proportion but not mean or variance. Results from 1000 iterations per setting.

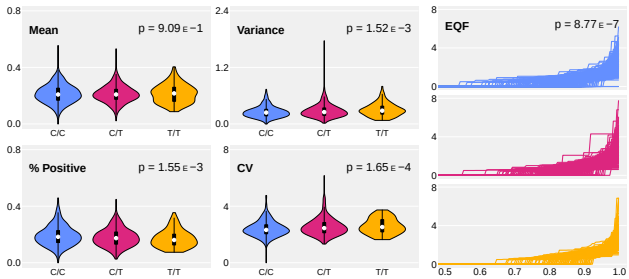


**Fig. 2.** Donor-level B-cell POLE4 gene expression summaries and partial F test corrected p-values over levels of rs1861457 SNP.

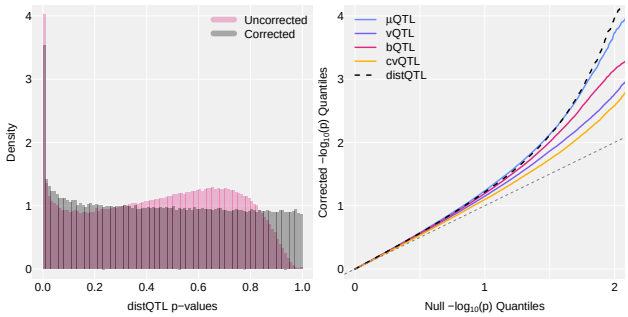


**Fig. 3.** P-value correction and power calculation. (**A**) Empirical histograms of raw and null-corrected distQTL p-values from gene-SNP pair models in B cells. (**B**) QQ plot of null-corrected $-\log_{10}(p)$-values for gene-SNP pairs in B cells.

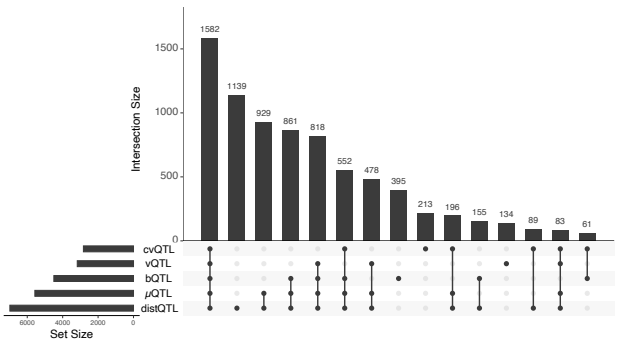


**Fig. 4.** Upset plot of eQTL-labeled gene-SNP pairs across B cells, using corrected p-values and significance threshold $\alpha = 10^{-4}$. Truncated to 15 intersections.

## OneK1K eQTL Identification

### *distQTL v.s. pseudo-bulk and univariate eQTL*

Fig. 2 illustrates donor-level univariate targets and EQFs for POLE4 gene across levels of rs1861457 SNP genotype (ref./rare alleles C/T) in B cells, with associated corrected t-test (baseline eQTL) and partial F-test (distQTL) p-values. Mean expression is not significantly associated with SNP genotype, but non-mean univariate responses are significantly associated with genotype, such as coefficient of variation ($p = 1.65\text{E}{-4}$). distQTL, which evaluates the whole conditional distribution of POLE4 expression, identifies the most significant association ($p = 8.77\text{E}{-7}$) for SNP genotype.

Fig. 3A shows observed distQTL p-value distribution before and after p-value correction. The estimated null p-value distribution indicates uncorrected distQTL p-values exhibit inflated Type I error, which can be ameliorated by permutation-based correction (see Supplement for details). The distribution of corrected p-values is nearly flat over the support while retaining a slightly reduced spike toward lower values. This demonstrates that our correction method properly adjusts for the mis-specified null distributions. We observe similar correction patterns in other cell-type groups (Fig. S2). Quantile-quantile (QQ) plot Fig. 3B illustrate that, post p-value correction, distQTL is more powerful in detecting eQTLs in B cells than univariate summary regression methods, most comparable with $\mu$QTL. The results can be reproduced in other cell-type groups (Fig. S2).

On the genome-wide scale, distQTL identifies more eQTLs when compared to the other methods, exemplified by the upset plots for the different cell types (Fig. 4, Fig. S3). In particular, distQTL identifies more gene-SNP pairs as strongly significant ($p < 10^{-4}$) in B cells while other methods label them as insignificant ($p > 0.1$), than vice versa — 1/0 compared to $\mu$QTL, 94/0 compared to vQTL, 333/137 compared to bQTL, and 1435/23 compared to cvQTL. This pattern holds across all cell-type groups (Table S2).

### *distQTL v.s. PMER*

We compare distQTL to PMER (Nathan et al., 2022) on the top eQTLs identified by distQTL in B cells. Fig. 5 illustrates $\log_{10}(-\log_{10} p)$ transformed nominal $p$-values from distQTL and PMER, using only highly expressed genes in a cell-type-specific manner. The transformation visually separates significant $p$-values while reducing skew: large positive values indicate highly significant test results, e.g. a value of 1 corresponds to $p = 1\text{E}{-10}$, and a value of 0 corresponds to $p = 0.1$. In total, we fit 35,035 cell-type-specific PMER models, using the Texas A&M High Performance Research Computing (TAMU-HPRC) resource.

distQTL and PMER exhibit strong correlation in eQTL-mapping results and attain very similar p-values when identifying significant eQTLs. However, on occasion the PMER model fitting procedure does not converge. The most common convergence issues are violation of the stationarity condition, and non-convexity or near-singularity of the Hessian, indicating poor identification of some parameters. These convergence issues occurred despite scaling covariates vectors prior to model fit. PMER convergence problems do not necessarily mean poor estimated coefficients, but can lead to unrealistically significant eQTL identification. Spurious results are only identifiable by comparison to other methods, or otherwise by resolving the convergence problems through manual tuning of the optimization algorithm on a case-by-case basis.

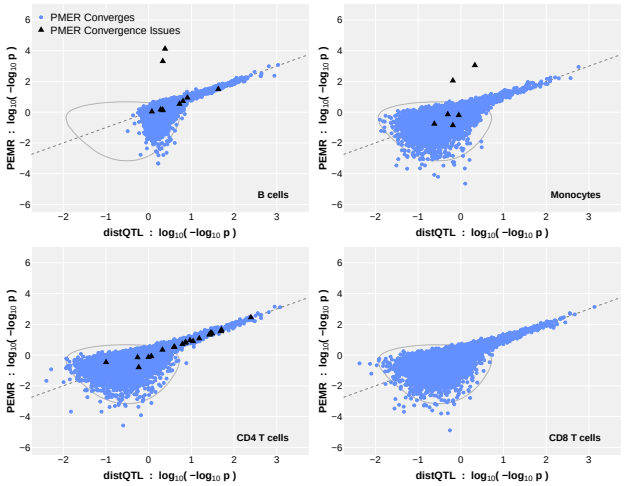


**Fig. 5.** eQTL-mapping $\log_{10}(-\log_{10} p)$-value comparison between PMER and distQTL for each cell-type group. Each point corresponds to one highly expressed gene and its top eQTL identified by distQTL. Contours are visual aids for joint p-value distribution under null of no SNP effect. PMER fits accompanied by convergence flags are labeled with black triangles. For reference, PMER model fit time was approximately 33 CPU hours; distQTL model fit time was approximately 1.6 CPU minutes.

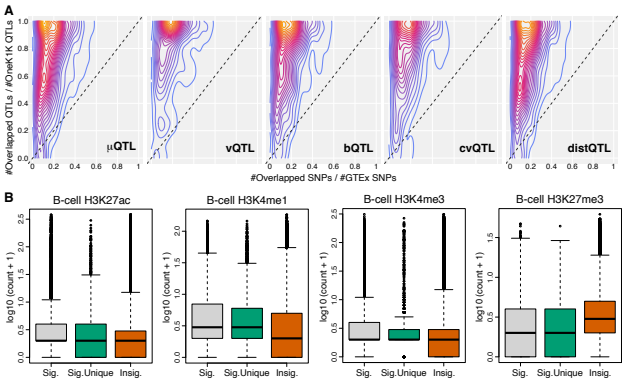


**Fig. 6.** Benchmark and validation results. (**A**) Gene-SNP OneK1K/GTEx enrichment contour plots, conditioned on OneK1K eQTL $\geqslant 4$ and GTEx eQTL $\geqslant 4$. eQTLs identified by $p \leqslant 10^{-4}$; matches are eQTL identification in both data sets. Values above the dashed $x = y$ line indicate eQTL matching better than chance. (**B**) B-cell eQTL validation by ChIP-seq of histone modifications. Sig.: significant QLTs detected by distQTL. Sig.Unique: Significant eQTLs detected exclusively by distQTL (not identified by other methods). Insig.: insignificant eQTLs. $\log_{10}(y+1)$ of ChIP-seq read counts is shown as y-axis. eQTLs detected (uniquely) by distQTL show stronger ChIP-seq signals for enhancer markers (H3K27ac, H3K4me1, and H3K4me3) and weaker signals for repression marker (H3K27me3), demonstrating the accuracy and outperformance of distQTL.

| Method | Total (min) B | Mono | CD4 T | CD8 T | Num. Models | Ave. (ms) |
|---|---|---|---|---|---|---|
| $\mu$QTL | 2.1 | 1.8 | 2.2 | 2.1 | 5.8 M | 7.6E−1 |
| PMER | 44.6 | 18.9 | 93.1 | 61.4 | 4000 | 2.9E+4 |
| distQTL | 63.6 | 100.4 | 36.9 | 59.9 | 5.8 M | 2.4E+1 |

**Table 1.** Model fit times (in minutes) for high expression genes, within each cell-type group; total number of models fit per method across the four cell-type groups; and average model fit time (in milliseconds).

## Benchmarking and Validation

Baseline eQTL methods and distQTL applied to OneK1K data set for cell-type group exhibit varying levels of enrichment against eQTL findings from whole blood bulk tissue samples in GTEx. We compare methods individually to GTEx findings by identifying *cis*-SNPs as eQTLs if the corresponding eQTL-mapping method's corrected (as applicable) p-value is smaller than $10^{-4}$, and counting matches of eQTL labeling between the two data sets. To reduce noise for illustrative purposes, we narrow focus to genes with at least four eQTLs found in both GTEx and OneK1K, on a method-by-method basis.

Fig. 6A shows contour plots of within-gene proportion of OneK1K eQTL labels which are also labeled eQTLs in GTEx, against proportion of GTEx gene-SNP pairs which were eQTL labeled. Values above the diagonal line indicate better than chance matching of GTEx-labeled eQTLs by methods applied to the OneK1K data set. Baseline $\mu$QTL and distQTL outperform the other methods both in identifying eQTLs in the aggregated cell groups and in identifying eQTLs that are also eQTLs in GTEx. Across all gene-SNP pairs that were identified as significant by GTEx, $\mu$QTL in OneK1K recovered 21.9%, and distQTL recovered 21.4%.

To further benchmark distQTL against the other univariate methods, we adopted cell-type-specific epigenomic profiles, encompassing ChIP-seq data of histone modifications. Compared to insignificant calls, eQTLs detected by distQTL overall and eQTLs exclusively detected by distQTL (but not by the other methods) exemplify stronger signals for active enhancer markers and weaker signal for repression marker (Fig. 6B). This orthogonal high-throughput validation from a different modality demonstrates the accuracy and outperformance of distQTL.

## Computation Time

Table 1 displays model fit times and total number of models fit, in each cell type specific group, by competing eQTL-mapping methods. Pseudo-bulk methods reduce scRNA-seq data set size by collapsing donor-level gene expression data into single numeric summaries. For simplicity, we present results only for $\mu$QTL; computation results for other baseline eQTL-mapping methods are nearly identical. Estimated time to fit each model with $\mu$QTL is on the order of 0.5 ms, owing to closed form solutions. Total $\mu$QTL run time across 5.8 million models was approximately 8 minutes.

scRNA-seq specific methods try to leverage the complete data set, which can increase computational burden. For PMER method, each individual cell is present as a row in the response and covariate objects, with a random intercept per donor accounting for cell-cell correlations. Such model structure requires very large data object sizes, with a large number of non-trivial parameters to estimate per model. To comparably evaluate PMER to other methods on the same machine, we fit a reduced set of PMER models as baseline eQTL and distQTL, using parallel computing. We randomly select 1000 genes which have high expression in all cell-type groups, and fit the top eQTL per gene as identified by distQTL in B cells. Average model fit time was approximately 29 seconds, upward of one minute for cell-type groups with larger numbers of cells; total time to fit all 4000 models was 33 CPU hours. (Fitting the larger set of 35,035 models on HPRC took approximately 589 CPU hours.) Notably, model fit time was weakly negatively correlated with eQTL significance level, indicating our reported average model fit times are likely optimistic lower bounds if

PMER is applied to all possible pairs of genes and *cis*-SNPs in OneK1K, some 2.9 million total models. This presents a daunting computational task, even with access to high power computational resources like TAMU-HPRC, and is infeasible with a typical personal computer.

distQTL is an scRNA-seq method that replaces complicated covariance structures with donor specific distribution responses. Hypothesis testing involves matrix calculations which have closed form solutions provided a set of covariate-dependent distribution responses can be estimated. The bulk of distQTL computation burden lies in estimating those responses, a convex quadratic programming problem (Petersen and Müller, 2019) which can be quickly solved to machine precision (Coulter et al., 2025). Individual distQTL model fit times are on the order of tens of milliseconds, upward of 1000× reduction in computation time compared to PMER, and obtaining more numerically stable results. On a personal laptop, total time to fit all highly expressed gene-SNP models, including with permuted responses, was a little over four hours. We believe this makes distQTL an attractive and computationally feasible scRNA-seq method which sidesteps the cell-cell correlation issue at the donor level.

## Discussion

Here, we propose distQTL for eQTL-mapping using population-scale scRNA-seq data, demonstrating its high power to detect mean effects while also identifying non-mean associations. Simulations and application to the OneK1K data set suggest that distQTL can identify more eQTLs that other methods lack the power to find (Table S1). Aggregating p-value results from permutation-refitted models indicates that type-I errors remain well-controlled. As with other eQTL-mapping methods, the permutation correction methods are a computational bottleneck and cannot adjust very significant p-values. Visual inspection of the uncorrected p-value distribution in Fig. 3B suggests that beta distribution correction methods (Ongen et al., 2016) will not be appropriate. Nonetheless, fast implementation of Fréchet regression enables efficient model computation under permutation, reducing computation time to just a few hours on a laptop. As a unified method that extends beyond the first and second moments of the gene expression distribution, distQTL results can be seamlessly integrated with those from other approaches (e.g., $\mu$QTL, vQTL, bQTL, and cvQTL) using heavy-tailed combination tests to integrate dependent p-values (Gui et al., 2023).

Tissue-specific eQTL-mapping by the GTEx Project includes more sophisticated regressors such as the PEER factors (The GTEx Consortium, 2020), which have been shown to increase power of eQTL-mapping by modeling latent batch effects (Stegle et al., 2010). While PEER factors can be easily incorporated as covariates in distQTL as well, it is unclear how PEER factors can be estimated using the population-scale scRNA-seq data exhibiting nested structures, except for generating averaged pseudo-bulk samples, which has recently been shown to be nontrivial (Xue et al., 2023). Another recent study suggest principal component analysis is orders of magnitude faster, better-performing, and much easier to interpret and use than the PEER factors for eQTL-mapping (Zhou et al., 2022). Additionally, at the time of writing, we were not able to install the `R peer` package which implements the PEER factor analysis through or its GitHub repository. Despite utilizing a simpler regression model, distQTL identifies more gene-SNP pairs in OneK1K which are significant at the $10^{-4}$ level while the other methods retain insignificance at the 0.10 level (Table S2).

For scRNA-seq data, transcriptional bursting that results in stochastic gene expression is a widespread phenomenon observed at the single-cell level, and a two-state Poisson-Beta model has been successfully adopted to infer bursting kinetics (Jiang et al., 2017; Vu et al., 2016). However, data from 5' or 3' end tagging techniques (e.g., 10x Genomics) is sparser and noisier than full-transcript scRNA-seq, leading to poor model fittings and identifiability issues (Jiang et al., 2017; Weideman et al., 2025). The inherent heterogeneity across cell types and/or cell states further exacerbates the problem, necessitating a large number of cells for accurate parameter estimation. While we included and benchmarked bursting QTLs – using the proportion of cells with non-zero expression as a surrogate of the bursting frequency that cannot be uniquely and accurately estimated – distQTL characterizes gene expression distributions beyond mean-based bulk or pseudo-bulk approaches, linking them to genotypes across individuals to reveal the underlying regulatory mechanism.

With rapidly increasing sequencing capacity and technological advancement, distQTL – alongside univariate summaries such as variance, bursting kinetics, and coefficient of variation when applied to more cells sequenced at a higher depth – can uncover regulatory relationships at a finer resolution. Finally, distQTL is not limited to using expression as the quantitative trait; it can be readily extended to other omics measures such as chromatin accessibility, alternative splicing, and allele-specific expression. Mapping eQTLs using single-cell multiomics data, where multiple modalities are simultaneously profiled within the same cells (Jiang et al., 2022), holds the potential to provide deeper insights into into the intricate transcriptional regulatory network and warrants additional methodological developments as a future direction.

## Data availability

OneK1K scRNA-seq and genotype data were downloaded from the HCA (`https://explore.data.humancellatlas.org/projects/f2078d5f-2e7d-4844-8552-f7c41a231e52`). H3K27ac, H3K4me1, H3K4me3, and H3K27me3 ChIP-seq data of B cells (CD19) were downloaded from GEO with accession numbers GSM1027287, GSM1027296, GSM1027300, and GSM1160194, respectively. Tissue-specic *cis*-QLTs from GTEx Analysis V8 were downloaded from the GTEx Portal (`https://www.gtexportal.org/home/downloads/adult-gtex/qtl`).

## Competing interests

The authors declare no competing interests.

## Author contributions statement

A.C., Y.N., and Y.J. initiated the study and conceived the methods. A.C. and Y.J. performed data quality control; all authors performed data analysis and results interpretation. Y.N. and Y.J. supervised the work and provided funding support. A.C. and Y.J. wrote the manuscript, which was reviewed and approved by all authors. A.C. wrote the accompanying R package.

## Acknowledgments

This work was supported by the National Institute of Health (NIH) grant R01 GM148974 (to Y.N.) and R35 GM138342 (to Y.J.). The authors thank Dr. Joseph Powell for providing support on the OneK1K genotype calls. Portions of this research were conducted with the advanced computing resources and consultation provided by Texas A&M High Performance Research Computing.

## References

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software 67*(1), 1–48.

Bernstein, B. E., J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology 28*(10), 1045–1048.

Corty, R. W. and W. Valdar (2018). vqtl: An R Package for Mean-Variance QTL Mapping. *G3 Genes|Genomes|Genetics 8*(12), 3757–3766.

Coulter, A., R. Lee, and I. Gaynanova (2025). fastfrechet: An R package for fast implementation of Fréchet regression with distributional responses. arXiv:2503.06401.

Ding, R., Q. Wang, L. Gong, T. Zhang, X. Zou, K. Xiong, Q. Liao, M. Plass, and L. Li (2024). scQTLbase: an integrated human single-cell eQTL database. *Nucleic Acids Research 52*(D1), D1010–D1017.

Eling, N., A. C. Richard, S. Richardson, J. C. Marioni, and C. A. Vallejos (2018). Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Systems 7*(3), 284–294.e12.

Elorbany, R., J. M. Popp, K. Rhodes, B. J. Strober, K. Barr, G. Qi, Y. Gilad, and A. Battle (2022). Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. *PLOS Genetics 18*(1), e1009666.

Gui, L., Y. Jiang, and J. Wang (2023). Aggregating Dependent Signals with Heavy-Tailed Combination Tests. arXiv:2310.20460.

Hong, C., Y. Ning, P. Wei, Y. Cao, and Y. Chen (2017). A Semiparametric Model for VQTL Mapping. *Biometrics 73*(2), 571–581.

Jerber, J., D. D. Seaton, A. S. E. Cuomo, N. Kumasaka, J. Haldane, J. Steer, M. Patel, D. Pearce, M. Andersson, M. J. Bonder, E. Mountjoy, M. Ghoussaini, M. A. Lancaster, J. C. Marioni, F. T. Merkle, D. J. Gaffney, and O. Stegle (2021). Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature Genetics 53*(3), 304–312.

Jiang, Y., Y. Harigaya, Z. Zhang, H. Zhang, C. Zang, and N. R. Zhang (2022). Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions. *Cell systems 13*(9), 737–751.

Jiang, Y., N. R. Zhang, and M. Li (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biology 18*(1), 74.

Larsson, A. J. M., P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, A. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg (2019). Genomic encoding of transcriptional burst kinetics. *Nature 565*(7738), 251–254.

Luo, J., X. Wu, Y. Cheng, G. Chen, J. Wang, and X. Song (2023). Expression quantitative trait locus studies in the era of single-cell omics. *Frontiers in Genetics 14*, 1182579.

Mackay, T. F. C., E. A. Stone, and J. F. Ayroles (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics 10*(8), 565–577.

Maria, M., N. Pouyanfar, T. Örd, and M. U. Kaikkonen (2022). The Power of Single-Cell RNA Sequencing in eQTL Discovery. *Genes 13*(3), 502.

Nathan, A., S. Asgari, K. Ishigaki, C. Valencia, T. Amariuta, Y. Luo, J. I. Beynor, Y. Baglaenko, S. Suliman, A. L. Price, L. Lecca, M. B. Murray, D. B. Moody, and S. Raychaudhuri (2022). Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature 606*(7912), 120–128.

Neavin, D., Q. Nguyen, M. S. Daniszewski, H. H. Liang, H. S. Chiu, Y. K. Wee, A. Senabouth, S. W. Lukowski, D. E. Crombie, G. E. Lidgerwood, D. Hernández, J. C. Vickers, A. L. Cook, N. J. Palpant, A. Pébay, A. W. Hewitt, and J. E. Powell (2021). Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biology 22*(1), 76.

Ongen, H., A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics 32*(10), 1479–1485.

Petersen, A., X. Liu, and A. A. Divani (2021). Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics 49*(1), 590–611.

Petersen, A. and H.-G. Müller (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics 47*(2), 691–719.

Sarkar, A. K., P.-Y. Tung, J. D. Blischak, J. E. Burnett, Y. I. Li, M. Stephens, and Y. Gilad (2019). Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLOS Genetics 15*(4), e1008045.

Shen, Q. and J. Faraway (2004). An F Test for Linear Models with Functional Responses. *Statistica Sinica 14*(4), 1239–1257.

Stegle, O., L. Parts, R. Durbin, and J. Winn (2010). A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLOS Computational Biology 6*(5), e1000770.

The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science 369*(6509), 1318–1330.

van der Wijst, M. G. P., H. Brugge, D. H. de Vries, P. Deelen, M. A. Swertz, and L. Franke (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics 50*(4), 493–497.

Vu, T. N., Q. F. Wills, K. R. Kalari, N. Niu, L. Wang, M. Rantalainen, and Y. Pawitan (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics 32*(14), 2128–2135.

Weideman, A. M. K., R. Wang, J. G. Ibrahim, and Y. Jiang (2025). Canopy2: tumor phylogeny inference by bulk dna and single-cell rna sequencing. *Statistics in Biosciences*, 1–43.

Xue, A., S. Yazar, D. Neavin, and J. E. Powell (2023). Pitfalls and opportunities for applying latent variables in single-cell eqtl analyses. *Genome Biology 24*(1), 33.

Yazar, S., J. Alquicira-Hernandez, K. Wing, A. Senabouth, M. G. Gordon, S. Andersen, Q. Lu, A. Rowson, T. R. P. Taylor, L. Clarke, K. Maccora, C. Chen, A. L. Cook, C. J. Ye, K. A. Fairfax, A. W. Hewitt, and J. E. Powell (2022). Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science 376*(6589), eabf3041.

Zhou, H. J., L. Li, Y. Li, W. Li, and J. J. Li (2022). Pca outperforms popular hidden variable inference methods for molecular qtl mapping. *Genome biology 23*(1), 210.