METHOD

Open Access

SURGE: uncovering context-specific genetic-regulation of gene expression from single-cell RNA sequencing using latent-factor models

Benjamin J. Strober[1], Karl Tayeb[2], Joshua Popp[3], Guanghao Qi[3], M. Grace Gordon[4,5,6,7], Richard Perez[6], Chun Jimmie Ye[5,6,7,8,9] and Alexis Battle[3,10,11*]

*Correspondence: ajbattle@jhu.edu

[3] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
Full list of author information is available at the end of the article

Abstract

Genetic regulation of gene expression is a complex process, with genetic effects known to vary across cellular contexts such as cell types and environmental conditions. We developed SURGE, a method for unsupervised discovery of context-specific expression quantitative trait loci (eQTLs) from single-cell transcriptomic data. This allows discovery of the contexts or cell types modulating genetic regulation without prior knowledge. Applied to peripheral blood single-cell eQTL data, SURGE contexts capture continuous representations of distinct cell types and groupings of biologically related cell types. We demonstrate the disease-relevance of SURGE context-specific eQTLs using colocalization analysis and stratified LD-score regression.

Keywords: eQTL, Single-cell transcriptomics

Background

A complete, mechanistic understanding of the genetic basis of complex traits could provide insights into the biological basis of human health and disease. A powerful approach to filling in the missing links between genetics and complex traits is to use molecular measurements, such as gene expression levels, as an intermediate phenotype. Genetic variants significantly associated with gene expression are known as expression quantitative trait loci (eQTLs) [1–5]. Although eQTL studies have now been performed in large cohorts and numerous tissues [5, 6], characterizing the impact of regulatory genetic variants is far from complete. This complexity arises in part because the effects of genetic variation on gene expression vary considerably between different cellular contexts, such as cell types, developmental stage, or condition (Fig. 1A) [7–14].

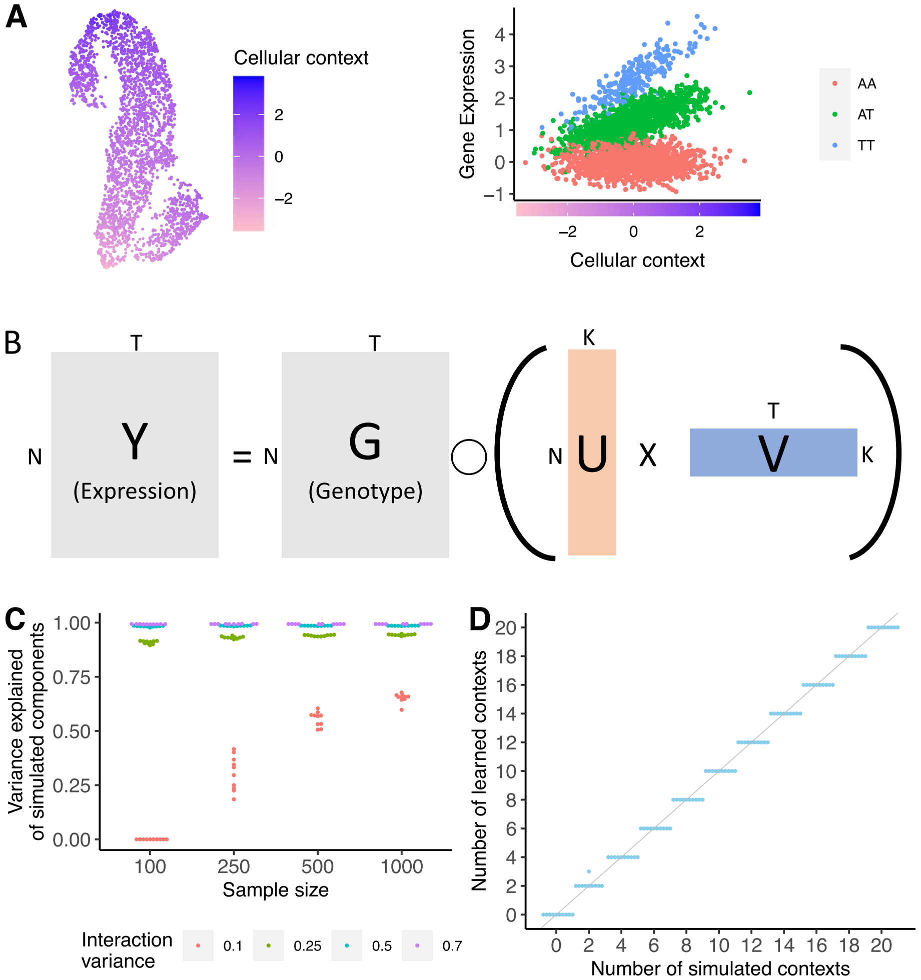

Fig. 1 SURGE model overview and simulation. **A** Schematic example of an interaction eQTL where the eQTL effect size (right) changes as a function of cellular context (depicted in UMAP embedding, left). **B** SURGE is a novel probabilistic model that uses matrix factorization to jointly learn a continuous representation of the cellular contexts defining each measurement (U) and the corresponding eQTL effect sizes specific to each learned context (V) based on observed expression (Y) and genotype (G) data. SURGE additional accounts for the effects of known covariates and sample repeat structure on gene expression (not shown in figure; see the "Methods" section). Assume there are N samples, T genome-wide independent variant-gene pairs, and K latent contexts. **C** Based on simulated data, we evaluated SURGE's ability to reconstruct simulated latent contexts as measured by the average variance explained of the simulated latent contexts by the learned latent contexts (y-axis). We simulate 5 latent contexts and vary the sample size (*x*-axis) and the strength (variance; see the "Methods" section) of the interaction terms (colors). We fix the fraction of tests that are context-specific eQTLs for each context to .3 (see the "Methods" section). For each parameter setting, we run 10 independent simulations. Each dot is an independent simulation. **D** Based on simulated data, we evaluate SURGE's ability to identify the number of simulated latent contexts across 10 independent simulations. The sample size was fixed to 250, the strength (variance) of the simulated interaction terms was fixed to .25, and the fraction of tests that are context-specific eQTLs for a particular context (see the "Methods" section) was fixed to .3. For each parameter setting, we run 10 independent simulations. Each dot is an independent simulation

Indeed, eQTLs from adult bulk tissue samples fail to explain the majority of known disease loci [11, 15–17]. It is therefore critical to identify eQTLs from diverse contexts in order to more fully characterize the molecular mechanisms underlying disease associated loci. Recent work has shown single-cell RNA sequencing (scRNA-seq) provides unique data to uncover cell type- and context-specific eQTLs; such higher-resolution data will naturally better reflect diverse cell types and cellular states, many of which would not be detectable from bulk RNA-seq [9, 10, 12–14, 18, 19].

However, the relevant contexts, such as cell type or state, that actually modulate genetic effects may not be known a priori, for example, genetic regulatory effects that are only present in a rare cell type, during intermediate stages of cellular differentiation [8, 9] or in response to environmental stimuli [7] that may not already be known to be disease relevant. Furthermore, an individual cell may be defined by multiple, overlapping contexts, such as both cell type and a perturbation response affecting partially overlapping sets of cells [9, 20, 21]. Contexts, such as differentiation progress or time, may manifest as continuous effects rather than discrete clusters. We developed SURGE (Single-cell Unsupervised Regulation of Gene Expression), a novel probabilistic model that uses matrix factorization to learn a continuous representation of the cellular contexts that modulate genetic effects. This includes the extent of relevance of each context to each cell or sample, and the corresponding eQTL effect sizes specific to each learned context, allowing for discovery of context-specific eQTLs without pre-specifying subsets of cells or samples.

First, we evaluate the statistical power of SURGE to identify latent contexts that modulate genetic effects on gene expression using simulated data. Next in a proof-of-concept experiment, we apply SURGE to bulk gene expression measurements from ten GTEx version 8 tissues [5] to uncover the relevant contexts underlying eQTL regulatory patterns in bulk RNA-seq data. We then use SURGE to identify context-specific eQTLs in a single-cell data set consisting of 1.2 million peripheral blood mononuclear cells (PBMC) spanning 224 genotyped individuals [18]. Finally, we demonstrate the disease-relevance of SURGE context-specific eQTLs using colocalization analysis and stratified LD-score regression (S-LDSC) [22, 23].

Results

A standard approach to identify context-specific eQTLs is to quantify the effect of the interaction between genotype and a pre-specified cellular context on gene expression levels using a linear model (interaction-eQTLs) [8, 9, 20, 24]. However, this approach requires pre-specifying which contexts, such as known cell types, to test for interaction, therefore inhibiting eQTL discovery in previously unstudied cellular contexts or uncharacterized cell types. Relatedly, others have tested the interaction between genotype and reduced-dimension expression features (such as gene expression principal components or MOFA factors [25]) on gene expression [13, 21], but this imposes the limiting assumption that all contexts that modulate the genetic regulation of gene expression can be explained by reduced-dimension features of gene expression and limits eQTL discovery to contexts with large effects on broad gene expression patterns.

To address these issues, we developed SURGE, which uses a matrix factorization approach to uncover context-specific eQTLs without requiring pre-specification of the contexts of interest. SURGE achieves this goal by leveraging information across genome-wide variant-gene pairs to jointly learn both a continuous representation of the latent cellular contexts defining each measurement (henceforth referred to as SURGE latent contexts) and the interaction eQTL effect sizes corresponding to each

SURGE latent context (Fig. 1B; see the "Methods" section). Importantly, SURGE allows for any individual measurement (such as a single cell) to be defined by multiple, overlapping contexts. From an alternative but equivalent lens, SURGE discovers the latent contexts whose linear interaction with genotype explains the most variation in gene expression levels. From this perspective, SURGE enables unsupervised discovery of the principal axes of genetic regulation of gene expression defining an eQTL data set. To more accurately infer SURGE latent contexts and increase power to detect context-specific eQTLs, SURGE jointly controls for the effects of known covariates as well as sample repeat structure induced by assaying multiple measurements (such as many cells) from the same individual on gene expression when inferring latent contexts (see the "Methods" section). Finally, SURGE automatically selects the number of relevant latent contexts by placing Automatic Relevance Determination prior distributions [26] on the inferred latent contexts (see the "Methods" section). The user simply would initialize the number of latent contexts to be large and greater than the likely number of underlying latent contexts present in the eQTL data set, and SURGE will prune unnecessary contexts during optimization (see the "Methods" section).

SURGE can be applied any eQTL data set using publicly available software (https://github.com/BennyStrobes/surge) and can be scaled to tens of thousands of measurements (Additional file 1: Fig. S1). After inference, SURGE latent contexts can be associated with available measurements and annotations to help interpret their biological meaning. In addition, latent contexts can be used identify SURGE interaction eQTLs or variants whose effect on genes expression significantly changes as a function of the SURGE latent context (see the "Methods" section).

Recently, there have been two methods proposed to identify contexts related to genetic regulation of gene expression from eQTL data sets [27–29]. SURGE is unique from these methods in that it identifies contexts whose linear interaction with genotype explain the most variation in single-cell gene expression levels. Vochteloo et al. [27] identifies contexts using an iterative algorithm such that the interaction between the context and genotype maximize expression variation in only the previous iteration's genome-wide significant context-interaction eQTLs. Contexts identified by this approach may not directly correspond to contexts whose interaction with genotype maximally explains variation in gene expression. Furthermore, this approach does not model sample-repeat structure and was not developed for application on single-cell eQTL data. Gewirtz et al. [28, 29] propose a method to identify shared latent topics present in both expression and genotype data. Topics identified by this approach will not directly correspond to contexts whose interaction with genotype maximally explains variation in gene expression. Therefore, the goals of each method are distinct, and SURGE uniquely identifies contexts where interaction between genotype and context drive variation in gene expression.

We utilize a simulation framework to statistically quantify SURGE's ability to accurately infer the latent contexts that alter genetic regulation of gene expression (see the "Methods" section). As expected, reconstruction of the simulated contexts depends on the sample size of the eQTL data set as well as the true effect size and number of context-specific eQTLs present in the simulated eQTL data set (Fig. 1C, Additional file 1: Fig. S2). However, given a realistic eQTL data set containing 100 modest effect context-specific eQTLs (simulated realistic interaction variance 0.25 [8]; see the "Methods" section) and sample size ($n = 250$), SURGE accurately infers the simulated latent contexts (Fig. 1C, Additional file 1: Fig. S2) as well as the number of simulated latent contexts (Fig. 1D, Additional file 1: Fig. S3).

As a proof of concept in real sequencing data, we apply SURGE to model RNA sequencing samples from 10 GTEx version 8 tissues (4169 individual-tissue pairs; Adrenal Gland, Colon-Sigmoid, Esophagus Mucosa, Muscle-Skeletal, Pituitary, Skin [not sun exposed suprapubic], Skin [sun exposed lower leg], Small Intestine terminal ileum, Stomach, and Thyroid), selected to be largely diverse with a small number of related tissues. SURGE identifies 15 latent contexts, resulting in hundreds of genes with at least one SURGE interaction eQTL (eFDR < = 0.05, see the "Methods" section) (Additional file 1: Figs. S4, S5, S6, and S7, Tables S1, S2). In this dataset, each RNA sample was extracted from a specific tissue, and while tissue identity information is not provided to SURGE, all 15 of the SURGE latent contexts are associated with differences in tissue type

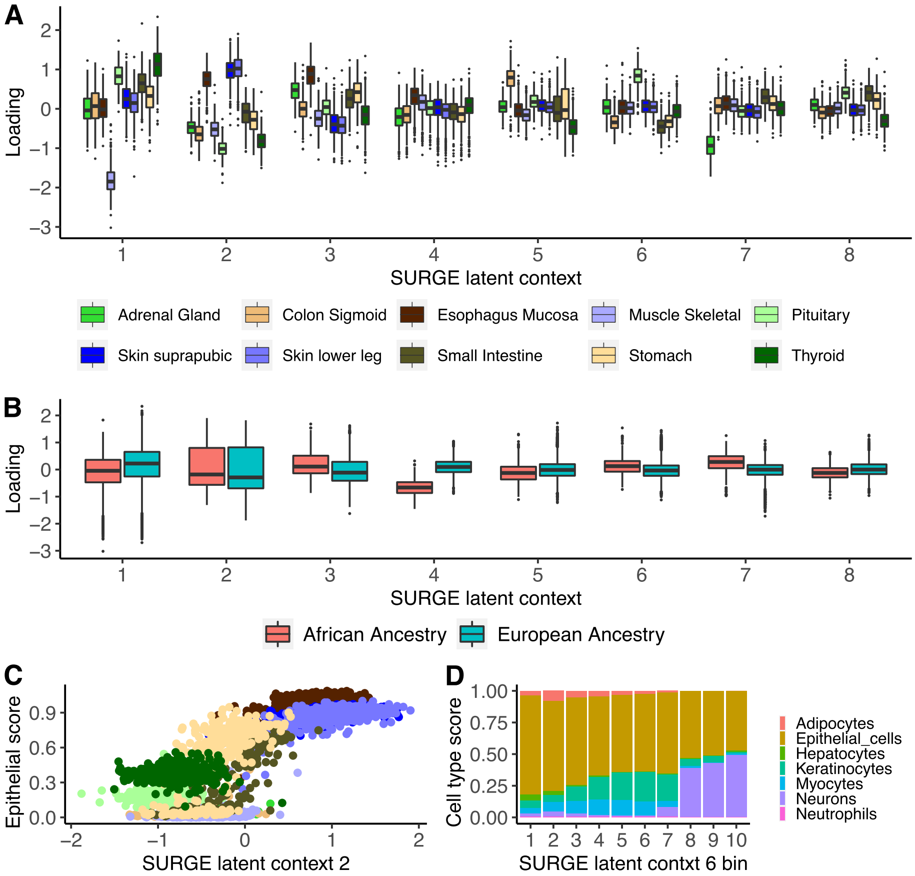

Fig. 2 SURGE applied to GTEx v8 bulk RNA-seq samples. **A**, **B** SURGE latent context loadings of GTEx v8 RNA-seq samples (*y*-axis) stratified by **A** known tissue identity and **B** known ancestry for top 8 inferred SURGE latent contexts. **C** Scatter plot of SURGE latent context 2 loadings (*x*-axis) and xCell Epithelial cell type enrichment score (*y*-axis) for GTEx v8 RNA-seq samples colored by known tissue identity (same color palette as **A**). **D** GTEx v8 RNA-seq samples are separated into 10 quantile bins according to their value on SURGE latent context 6. The stacked bar plot depicts the average xCell cell type enrichment scores across all samples normalized to sum to 1 (*y*-axis) in each of the 10 bins (*x*-axis)

between the samples (Additional file 1: Table S3) with more than 30% of the variation in 7 of the top 8 SURGE latent contexts (latent contexts ordered by PVE, see the "Methods" section) explained by tissue identity (Fig. 2A, Additional file 1: Table S3). SURGE latent context 1, for example, isolates RNA samples from Muscle-Skeletal tissue ($p < = 2.2$ e − 16, Wilcoxon rank sum test); RNA samples derived from Muscle-Skeletal tissue have an average latent context 1 value of − 1.82 (sdev 0.342), while RNA samples from other tissues have an average latent context 1 value of − 0.011 (sdev 0.456). Furthermore, SURGE latent context 4 and 7 cluster samples according to their known ancestry (Additional file 1: Table S4); samples from African Ancestry donors are strongly loaded on both latent context 4 and 7 (Fig. 2B, Additional file 1: Fig. S8).

Next, we intersect the learned SURGE latent contexts with previously computed computational estimates of each RNA sample's cell type composition according to xCell (xCell infers cell type enrichment scores that reflect cell type composition based on external cell type-specific gene expression data) [24, 30]. We find that the SURGE latent contexts are not simply identifying differences in tissue identity between the samples but learning differences in cell type composition of samples both across tissues and within a single tissue (Fig. 2C, Fig. 2D, Additional file 1: Fig. S9-S11). SURGE latent context 2, for example, is highly correlated with epithelial cell enrichment score across samples from all ten tissues (Fig. 2C; Spearman's rho 0.724, $p < = 2.2$ e − 16). Moreover, many of the SURGE latent contexts capture complex multi-cell type composition continuums, not simply the change in proportions of a single cell type (Fig. 2D, Additional file 1: Fig. S9, Table S5). SURGE identifies latent contexts underlying cell type composition continuums even when applied to RNA samples from only a single tissue (see the "Methods" section, Additional file 1: Fig. S12), demonstrating the importance of cell type composition differences across samples extracted from the same tissue. Importantly, we observe greater power to detect context-specific eQTLs with SURGE latent contexts than with the previously studied approach [30] of testing genetic interactions with cell type enrichment score estimates from xCell (see the "Methods" section, Additional file 1: Fig. S13). In summary, SURGE identifies tissue type, cell type, and ancestry as the primary axes of genetic regulation of gene expression within GTEx eQTL data.

Next, we apply SURGE to a recently generated single-cell eQTL data set consisting of 1.2 million PBMCs from 224 genotyped individuals [18]. One hundred forty-one of these individuals have systemic lulus erythematosus (SLE), while the remainder are healthy. To mitigate the sparsity characteristic of 10X sequencing data, we aggregate cell level expression data across highly correlated cells to generate 22,774 pseudocells (see the "Methods" section, Additional file 1: Fig. S14) [21, 31], aggregating on average 22 cells per pseudocell. Here, SURGE identifies 6 latent contexts, resulting in hundreds of genes with at least one SURGE interaction eQTL (eFDR < 0.1, see the "Methods" section) (Additional file 1: Figs. S15, S16, S17, Tables S6, S7).

SURGE latent contexts 1, 2, and 4 capture continuous representations of distinct blood cell types while integrating biologically related cell types along a gradient within a single latent context (Fig. 3A, Additional file 1: Figs. S18-S20, Tables S8, S9). SURGE latent context 2, for example, is strongly loaded on natural killer (NK) cells ($p < = 2.2$ e − 16, Wilcoxon rank sum test), while still identifying fine-resolution differences distinguishing bright NK cells from dim NK cells (Additional file 1: Fig. S18; $p < = 2.2$ e − 16, Wilcoxon

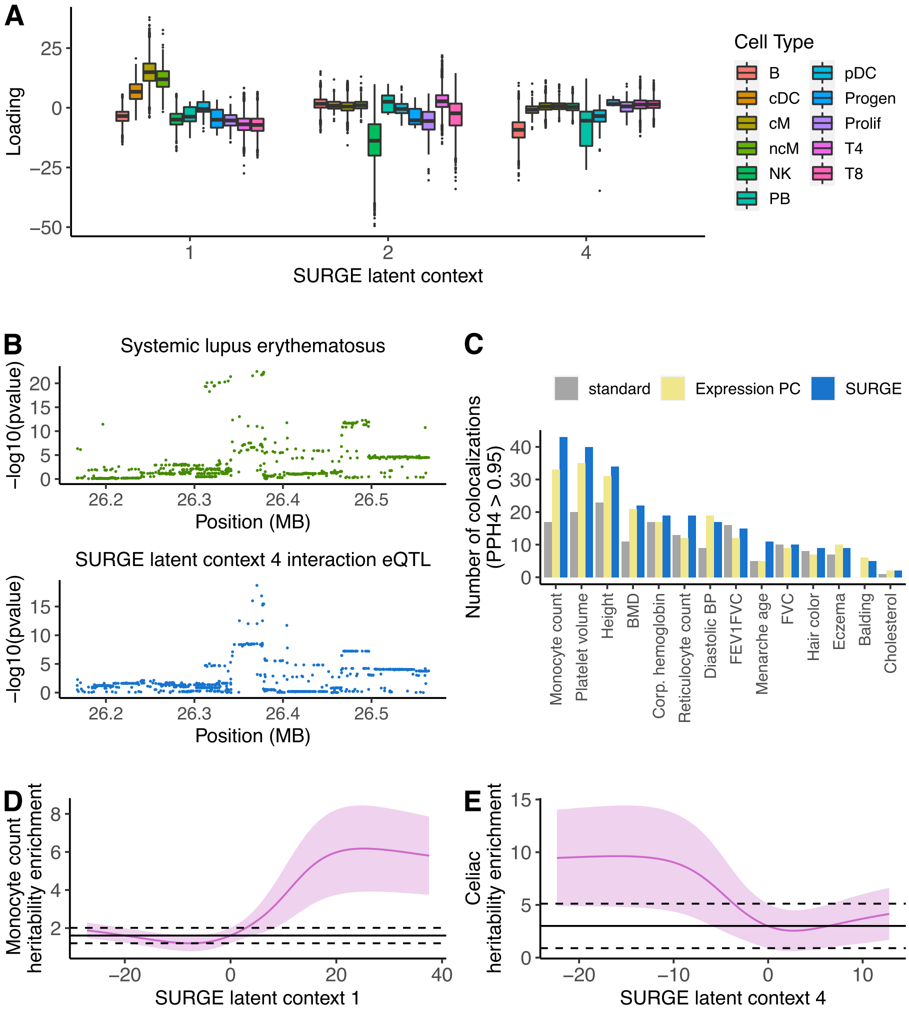

Fig. 3 SURGE applied to PBMC single-cell eQTL data. **A** SURGE latent context loadings of pseudocells (*y*-axis) stratified by cell type (color) according to marker gene expression profiles for each of the SURGE latent context 1, 2, and 4 (*x*-axis). **B** Colocalization between SURGE latent context 4 interaction eQTL variant chr6:26370572:C:T for BTN3A2 and GWAS signal for SLE. **C** Number of colocalizations identified (PPH4 > .95; *y*-axis) between various 14 independent GWAS studies (*x*-axis) and eQTLs identified from pseudocells. The number of colocalizations using standard eQTLs shown in grey, the number of unique colocalizations using expression PC interaction eQTLs aggregated across the top 6 expression PC shown in yellow, and the number of unique colocalizations using SURGE interaction eQTLs, aggregated across the 6 SURGE latent contexts, shown in blue. **D**, **E** S-LDSC enrichment (*y*-axis) of squared standard eQTL effect sizes (black line) and SURGE predicted squared eQTL effect size at a specific SURGE latent context value (pink line at a specific *x*-axis position) within **D** monocyte count and **E** celiac disease heritability. SURGE predicted eQTL effect sizes at a particular SURGE latent context value was calculated at 200 equally spaced positions along the range of SURGE latent context values. Black dashed line represents 95% confidence on the standard eQTL S-LDSC enrichment. Light pink region depicts 95% confidence on the SURGE predicted eQTL S-LDSC enrichment

rank sum test). Additionally, SURGE latent context 1 identifies subtle differences isolating monocytes derived from healthy individuals from monocytes derived from SLE individuals (Additional file 1: Fig. S21; $p <= 2e-16$, Wilcoxon rank sum test). Interestingly, SURGE latent contexts 3, 5, and 6 are only modestly explained by known cell types (Additional file 1: Table S8, Table S9) and are not explained by broad expression trends related to cell types defining the top gene expression principal components (Additional file 1: Figs. S22-S23). SURGE latent contexts 5 and 6 instead show strong correlation with expression of genes involved in specific biological processes (see the "Methods" section, Additional file 1: Tables S10, S11). For example, SURGE latent context 4 is correlated with genes that are extremely enriched in the Hallmark interferon-gamma response (odds ratio: 28.52, $p < 4.2e-10$) [32]. The interferon gamma response is a well-studied immune-related pathway shown to be involved in regulating SLE [18, 33, 34].

Finally, we evaluate the relationship between SURGE interaction eQTLs and disease-associated loci across diverse traits with genome-wide association studies (GWAS) available. Using coloc [22], we identify hundreds of colocalizations between SURGE interaction eQTLs and GWAS loci (Fig. 3B, C, Additional file 1: Fig. S24). For example, a SURGE context 4 interaction eQTL for BTN3A2 colocalized with a GWAS signal for SLE (Fig. 3B). We identify significantly more trait colocalizations with SURGE interaction eQTLs relative to using standard eQTLs (Fig. 3C; $p < 0.0026$, paired Wilcoxon rank sum test across 14 independent traits). In addition, we compared the number of trait colocalizations identified using SURGE interaction eQTLs with expression PC interaction eQTLs or significant interactions between gene expression principal components (PCs) and genotype on gene expression (see the "Methods" section). Despite both methods identifying comparable numbers of interaction eQTLs (Additional file 1: Fig. S17), SURGE interaction eQTLs colocalized (PPH4 > 0.95) with 2.4 more loci per trait on average across 15 independent traits (Fig. 3C; $p < 0.017$, paired Wilcoxon rank sum test). SURGE identified 66 trait colocalizations (PPH4 > 0.95) that could not be replicated (PPH4 > 0.1) by the expression PC interaction eQTL analysis, whereas the expression PC interaction eQTL only identified 31 trait colocalizations not replicated by SURGE interaction eQTLs. Notably, 50 of the 66 trait colocalizations unique to SURGE were discovered in the SURGE latent contexts not well explained by expression PCs (latent contexts 3, 5, and 6; Additional file 1: Figs. S22, S25) demonstrating the importance of not relying exclusively on expression PCs for interaction eQTL calling.

Next, we assess how eQTL enrichment in complex trait and disease heritability varied along the SURGE latent contexts using S-LDSC [23, 35]. Briefly, we used SURGE to estimate eQTL effect sizes at multiple positions along each SURGE latent context and then use S-LDSC to quantify the heritability enrichment of eQTLs identified at each position (see the "Methods" section). We note that this approach is not limited to SURGE interaction eQTLs and could be applied to any analysis that infers interaction eQTLs. We observed that eQTL enrichment in complex trait and disease heritability significantly varies along the SURGE latent contexts for many diseases and complex traits (Fig. 3D, E, Additional file 1: Fig. S26). For example, predicted eQTL effects in cells negatively loaded on SURGE latent context 4 (corresponding to a B cell continuum) are approximately four times more enriched in celiac disease heritability than static eQTLs (Fig. 3E). This result coincides with the previously reported role of B cells in celiac disease [36, 37]. Ultimately, this analysis highlights the importance of assessing eQTLs in disease-relevant contexts as well as SURGE's capacity for identifying disease-relevant contexts.

Discussion

Here, we presented SURGE, a novel probabilistic model that identifies context-specific eQTLs from single-cell data without pre-specifying context, such as cell types or subsets of samples. SURGE leverages information from variant-gene pairs across the entire genome to learn a continuous representation of the cellular contexts defining each measurement and the corresponding eQTL effect sizes specific to each learned context. Importantly, SURGE allows for unsupervised discovery of the principal axes of genetic regulation of gene expression within an eQTL data set, identifying latent contexts associated with cell type, tissue type, and ancestry when applied to GTEx tissue samples and highly resolved blood cell types and gene programs when applied to blood-derived single cells. We demonstrated that eQTL enrichment in complex trait and disease heritability significantly varied along the SURGE latent contexts and, ultimately, SURGE identified many trait-relevant loci that could not be detected through traditional eQTL approaches.

Although SURGE allows for identification of context-specific eQTLs without pre-specifying contexts, it has several limitations. First, SURGE is limited to discovery of interaction eQTLs with small effect sizes or those from contexts whose effects explain a large fraction of gene expression variation aggregated across many genes in the genome (Fig. 1C) and will be underpowered to detect interaction eQTLs from contexts that interact with the genetic regulation gene expression in a small number of genes. Second and like many unsupervised algorithms, it can be challenging to interpret the biological meaning of some SURGE latent contexts, particularly those that capture signals independent of observed measurement annotations such as cell type or sample covariates. Third, SURGE latent context inference and SURGE interaction eQTL calling is limited to the most significant variant for each eGene, hindering discovery of multiple independent context specific eQTLs regulating a single gene. Fourth, SURGE makes the simplifying assumption that the residual variance in gene expression is distributed normally. Recent work has demonstrated the benefits of non-normal distributions in single-cell interaction eQTL calling [13]. We leave the extension of SURGE with alternative residual distributions to future work. Fifth, SURGE cannot be reasonably scaled to more than 50,000 expression samples (Additional file 1: Fig. S1). This does not present a concern for existing eQTL data sets; however, SURGE optimization could be extended in the future, if needed, to scale to arbitrarily large eQTL data sets using parallel programming or stochastic optimization [25].

In conclusion, SURGE provides a statistically principled approach to uncover the dominant axes of genetic regulation of gene expression in an eQTL data set. It will become increasingly useful as large single-cell eQTL data sets containing cells spanning increasingly diverse cellular contexts are generated.

Methods

SURGE model overview

The SURGE model is defined according to the following probability distributions:

$$y_{nt} \sim N(\mu_t + \sum_l X_{nl} W_{lt} + \sum_i I[n \in i]\alpha_{it} + G_{nt} F_t + G_{nt}(\sum_k U_{nk} V_{kt}), \sigma_t^2)$$

$$U_{nk} \sim N\left(0, \gamma_k^2\right)$$

$$V_{kt} \sim N(0,1)$$

$$1/\gamma_k^2 \sim Gamma(\alpha_0, \beta_0)$$

$$F_t \sim N(0,1)$$

$$\alpha_{it} \sim N(0, \psi_t^2)$$

$$1/\psi_t^2 \sim Gamma(\alpha_0, \beta_0)$$

$$1/\sigma_t^2 \sim Gamma(\alpha_0, \beta_0)$$

Here, n indexes RNA samples, t indexes representative variant-gene pairs being tested for eQTL analysis, and i indexes individuals. We use the notation $n \in i$ to represent the instance where RNA sample n is drawn from the individual$i$. y_{nt} is the observed standardized gene expression (mean 0 and variance 1 for each test$t$) level of the gene corresponding to test t in sample$n$. We assume the gene expression data has been properly normalized prior to standardization. G_{nt} is the observed, standardized (described in more detail below) genotype of the variant corresponding to test t in sample$n$. X_{nl} is the observed value of covariate l for sample n. SURGE infers the values of:

- F_t: the eQTL effect size of test t that is shared across samples
- V_{kt}: the eQTL effect size of test t for latent context k
- U_{nk}: the latent context value of sample n on factor k
- μ_t: the intercept of each test
- W_{lt}: The effect size of covariate l on the gene corresponding to test t
- α_{it}: the random effect intercept for each individual for each test
- γ_k^2: The variance of the values in latent context k
- ψ_t^2: The variance of intercept corresponding to each individual in test t
- σ_t^2: The residual variance in gene expression levels in test t

a_0, and β_0 are model hyper-parameters set to provide non-informative priors while stabilizing optimization. In practice, we set α_0 to 1e^{-3} and β_0 to 1e^{-3}.

To standardize the genotype of the variant corresponding to test t, we center the genotype vector to have mean 0 across samples and then we scale the genotype vector for test t (G_{*t}) by the standard deviation of Y_{*t}/G_{*t}. This scaling encourages the low-dimensional factorization (UV) to explain variance equally across tests instead of preferentially explaining variance in tests with small variance in Y_{*t}/G_{*t}.

It is worth highlighting that a mean-zero gaussian prior is placed on U_{nk} in order to produce interpretable assignments of samples to factors. The level of regularization of that prior is learned separately for each latent context (γ_k^2), allowing SURGE to zero-out (γ_k^2 approaches 0) irrelevant contexts and automatically learn the effective number of latent contexts. This approach has been used by others for inference of the number of effective components in more traditional matrix factorizations [25, 38] and is similar to Automatic Relevance Determination [26].

SURGE optimization

We approximate the posterior distribution of all latent variables $[Z = (F_t, V_{kt}, U_{nk}, \mu_t, W_{lt}, \alpha_{it}, \gamma_k^2, \psi_t^2, \sigma_t^2)]$ using mean-field variational inference [39]. The goal of variational inference is to minimize the KL-divergence between $q(Z)$ and $p(Z|Y,G,X)$, which can be written as $KL(q(Z)||p(Z|Y,G,X)$. Here, $q(Z)$ is a simple, tractable distribution that is used to approximate $p(Z|Y,G,X)$. We used the "mean-field approximation" for $q(Z)$ such that all latent variables are independent of one another. More specifically:

$$\log q(Z) =$$

$$\sum_t\sum_k logN(V_{kt}|\mu_{V_{kt}}, \sigma^2_{V_{kt}})+$$

$$\sum_t\sum_i logN(\alpha_{it}\middle|\mu_{\alpha_{it}}, \sigma^2_{\alpha_{it}}\Big)+$$

$$\sum_t\sum_l \log N\Big(W_{lt}|\mu_{W_{lt}}, \sigma^2_{W_{lt}}\Big)+$$

$$\sum_t [logN\Big(F_t|\mu_{F_t}, \sigma^2_{F_t}\Big)+logN\Big(\mu_t|\mu_{\mu_t}, \sigma^2_{\mu_t}\Big)+logG\Big(1/\psi_t^2|\alpha_{\psi_t}, \beta_{\psi_t}\Big)+logG\Big(1/\sigma_t^2|\alpha_{\sigma_t}, \beta_{\sigma_t}\Big)]+$$

$$\sum_k logG(1/\gamma_k^2|\alpha_{\gamma_k}, \beta_{\gamma_k})+$$

$$\sum_n\sum_k logN(U_{nk}|\mu_{U_{nk}}, \sigma^2_{U_{nk}})$$

where $N(x|\mu,\sigma^2)$ is a univariate normal distribution parameterized by mean μ and variance σ^2, and $G(X|\alpha,\beta)$ is a univariate gamma distribution parameterized by α and β.

It can be shown that minimizing the KL-divergence $KL(q(Z)||p(Z|Y,G,X)$ is equivalent to maximizing the evidence lower bound (ELBO):

$$E_q[logp(G,Y,X,Z)] - E_q[logq(Z)]$$

Therefore, we will frame SURGE optimization from the perspective of maximizing the ELBO with respect to the parameters defining $q(Z)$ or the variational parameters. Noteworthy is $p(G,Y,X,Z)$ is explicitly defined in the "SURGE model overview" section and can be easily computed. The approach we take to maximize the ELBO is through coordinate ascent [39], iteratively updating the variational distribution each latent variable, while holding the variational distributions of all other latent variables fixed. Accordingly, the ELBO is guaranteed to monotonically increase after each variational update. In the case of the SURGE model, each update is available in closed form (see the supplement materials).

Optimization of variational parameters is performed as follows: we randomly initialize all variational parameters (see the "Random initialization for SURGE optimization" section) and then iteratively loop through all latent variables in Z and update the variational parameters corresponding to that latent variable until we reach convergence.

To assess convergence, we assess the change in ELBO from one iteration to the next. We consider the model converged when the change in ELBO is less than $1\mathrm{e}-2$.

Random initialization for SURGE optimization

It is important to note that mean-field variational inference is not guaranteed to converge to the global optima of the ELBO. To mitigate the effects of local optima, we recommend optimizing multiple models with different random initializations and using the parameters learned from the model that achieves the largest ELBO.

Proportion variance explained of SURGE latent contexts

Following the approach taken by [40], we define the "proportion variance explained" (PVE) of the k^{th} latent context as:

$$pve_k = \frac{s_k}{(\sum_k s_k) + (N * \sum_t \sigma_t^2)}$$

$$s_k = \sum_n \sum_t G_{nt} U_{nk} V_{kt}$$

As stated in [40], this approach is a measure of the amount of signal in data set that is identified by the k^{th} latent context. However, the name "proportion variance explained" should be considered loosely as the factors are not orthogonal.

We also consider the "Proportion of SURGE-mediated variance explained" (PSMVE) for the k^{th} latent context as:

$$PSMVE_k = \frac{s_k}{\sum_k s_k}$$

PSMVE quantifies the fraction of total variance in gene expression that is mediated through context-specific eQTL effects attributed to each latent context.

Removing irrelevant latent contexts

Upon model convergence, we remove latent contexts with PVE $\leq 1e^{-5}$.

Simulation experiments

To assess SURGE's ability to accurately capture contexts underlying context-specific eQTLs, we performed the following simulation experiment.

We randomly generated genotype and expression matrices across 1000 variant-gene pairs and N RNA samples. For each simulated variant-gene pair, we simulated the genotype vector (G) across the N samples according to the following probability distributions:

$$G_n \sim Binomial(2, \text{allele\_frequency})$$

$$\text{allele\_frequency} \sim Uniform(.05, .95)$$

Then, we simulated the expression vector (y) across the N samples using that variant-gene pair's simulated genotype vector according to the following probability distributions:

$$y_n \sim N(\mu + \beta G_n + \sum_k G_n U_{nk} V_k \theta_k, 1)$$

$$\mu \sim N(0, 1)$$

$$\beta \sim N(0, 1)$$

$$U_{nk} \sim N(0, 1)$$

$$V_k \sim N(0, \gamma)$$

$$\theta_k \sim Bernoulli(p)$$

Using this simulation, we can evaluate SURGE's ability to re-capture the simulated latent contexts (U) (Fig. 1C, Additional file 1: Fig. S2) as a function of the simulation hyper-parameters:

- The number of latent contexts (K)
- The sample size (N)
- The strength of the interaction terms (γ)
- The fraction of tests that are context-specific eQTLs for a particular context (p)

We can also access SURGE's ability to accurately estimates the number of relevant contexts (K) (Fig. 1D, Additional file 1: Fig. S3) by only retaining contexts with PVE $> 1e^{-5}$.

Selection of representative variant-gene pairs used for optimization

SURGE optimization (i.e., learning the SURGE latent contexts) requires an input expression matrix and genotype matrix. As specified above, both matrices should be of dimension N X T, where N is the number of RNA samples and T is the number of genome-wide representative variant gene pairs. We desire each variant-gene pair to be independent of one another because we want SURGE to infer eQTL patterns that are persistent across the genome, not specific to a single gene or variant.

To encourage the expression and genotype data to consist of independent variant-gene pairs, we limit there to be a single variant-gene pair selected for each gene and limit there to be a single variant-gene pair selected for each variant.

Furthermore, it has been shown that context-specific eQTLs are more likely to be standard eQTLs than not. We therefor limit the representative variant-gene pairs used for SURGE optimization to those that are standard eQTLs within the data set (more details presented below). For computational efficiency, we recommend using a maximum of 2000 genome-wide representative variant-gene pairs for SURGE optimization.

SURGE interaction-eQTLs

SURGE optimization on a subset of genome-wide representative variant-gene pairs will result in approximations to the posterior distributions of the SURGE latent contexts (U) as well as eQTL effect sizes for each of the SURGE latent contexts for only the representative variant gene pairs (V). It is of interest, however, to call interaction eQTLs with respect to each of the SURGE latent contexts for *all* variant gene-pairs, not just the subset of representative variant-gene pairs used for SURGE optimization.

Therefore, to identify SURGE interaction-eQTL for an arbitrary variant-gene pair, we treat the expected value of the inferred posterior distribution on the SURGE latent contexts ($\widehat{U}$: dim NXK) as observed and optimize the following linear mixed model for each variant-gene pair. The linear mixed model is as follows:

$$y_n \sim N(\mu + \sum\nolimits_i \alpha_i I[n \in i] + \sum\nolimits_l W_l X_{nl} + \beta_g G_n + \sum\nolimits_k \beta_k \widehat{U}_{nk} + \sum\nolimits_k \beta_{gxk} G_n \widehat{U}_{nk}, \sigma^2)$$

$$\alpha_i \sim N(0, \psi^2)$$

Here:

- y_n is the observed expression level of the gene corresponding to the variant-gene pair in sample n
- g_n is the observed genotype of the variant corresponding to the variant-gene pair in sample n
- X_{nl} is the observed value of covariate l in sample n
- μ is the intercept
- α_i is the random effect intercept for individual i. We use the notation $n \in i$ to represent the case where sample n is drawn from individual i
- W_l is the fixed effect for covariate l
- β_g is the fixed effect for genotype
- β_k is the fixed effect of the k^{th} latent context
- β_{gxk} is the fixed effect of the interaction between the k^{th} latent context and genotype

We use the R package “lme4” to quantify the significance of all K interaction terms: $\beta_{gx1}, \ldots, \beta_{gxk}, \ldots, \beta_{gxK}$. Intuitively, if the k^{th} interaction term (β_{gxk}) is significant, it implies that the eQTL effect size of this variant-pairs significantly changes along latent context k.

Calibration of SURGE interaction eQTLs using permutation analysis

P-values resulting from SURGE interaction-eQTL analysis are potentially inflated due to SURGE interaction eQTLs being identified from the same data used to learn the SURGE latent contexts. This statistical phenomenon is known as “double-dipping,” and there exist well-studied approaches to ensure statistical calibration in the presence of “double-dipping” [41–44]. We used the empirical false discovery rate (eFDR) [45] to generate well-calibrated SURGE interaction eQTLs. eFDR establishes significance by comparing the observed SURGE interaction eQTL *p*-values with an empirical null distribution of *p*-values.

In more detail, we use a conservative, permutation analysis to generate an empirical null distribution of gene-level *p*-values that can be used to calibrate the observed gene-level *p*-values. The permutation analysis consisted of (1) permuting genotype of each individual, (2) re-optimizing SURGE latent contexts (U) using the permuted genotype data, and (3) calling SURGE interaction-eQTLs with the permuted genotype data and the SURGE latent contexts learned using the permuted genotype data.

To permute genotype, we generated a single permutation of genotype that was used across all analyzed variants to ensure we did not break the correlation structure across variant-gene pairs. In addition, we only permuted genotype across individuals, not RNA samples, to ensure we preserved sample repeat structure. This means that multiple RNA samples from the same individual will always have the same genotype values in the permutation run. Lastly, similar to previous permutation experiments on linear-interaction effects [8, 46], we only permuted the genotype variable in the interaction term while leaving the fixed effect of genotype un-permuted (for both SURGE optimization and SURGE interaction-eQTL calling).

Given both the observed and permuted SURGE interaction eQTL *p*-values, we generated gene-level *p*-values using Bonferroni correction for both observed and permuted interaction eQTLs. We then evaluated genome-wide significance of the observed gene-level *p*-values using empirical FDR (eFDR) [45] calibrated with the permuted gene-level *p*-values. This approach was performed independently for each SURGE latent context. We will refer to this approach as the "per-context eFDR."

In the real (un-permuted) data, we only called SURGE interaction-eQTLs for SURGE latent contexts with PVE > $1e^{-5}$. Unsurprisingly, permuted SURGE latent contexts consistently explained less PVE than un-permuted SURGE latent contexts (Additional file 1: Fig. S4, Fig. S15); there existed zero permuted SURGE latent contexts explaining PVE > $1e^{-5}$ across all experiments. Therefore, if Z SURGE latent contexts have PVE > $1e^{-5}$, we selected the top Z permuted SURGE latent contexts to be used in the permuted SURGE interaction eQTL analysis.

We consider an alternative multiple testing procedure where the eFDR is computed jointly across SURGE interaction eQTLs in all SURGE latent contexts instead of computing eFDR in each context independently (per-context eFDR). Specifically, we evaluated genome-wide significance of the observed (gene, SURGE context) pair *p* values using eFDR calibrated using the permuted (gene, SURGE permuted context) pair *p* values. We will refer to this approach as the "all-context eFDR."

However, in concordance with previous multiple testing-correction procedures for multi-context eQTL calling [5, 24], we recommend using the per-context eFDR.

Expression PC interaction-eQTLs

We benchmarked the SURGE interaction eQTLs against expression PC interaction eQTLs. Intuitively, expression PC interaction eQTLs identify statistically significant interactions between a gene expression principal component and the genotype of a particular snp on the expression levels of a particular gene. We set the number of expression PCs equal to the number of SURGE latent contexts.

To identify expression PC interaction-eQTLs for an arbitrary variant-gene pair, we first use PCA to calculate the expression principal components (E: dim NXK). Then, we optimize the following linear mixed model for each variant-gene pair. The linear mixed model is as follows:

$$y_n \sim N(\mu + \sum_i \alpha_i I[n \in i] + \sum_l W_l X_{nl} + \beta_g G_n + \sum_k \beta_k E_{nk} + \sum_k \beta_{gxk} G_n E_{nk}, \sigma^2)$$

$$\alpha_i \sim N(0, \psi^2)$$

Here:

- y_n is the observed expression level of the gene corresponding to the variant-gene pair in sample n
- g_n is the observed genotype of the variant corresponding to the variant-gene pair in sample n
- X_{nl} is the observed value of covariate l in sample n
- μ is the intercept
- α_i is the random effect intercept for individual i. We use the notation $n \in i$ to represent the case where sample n is drawn from individual i
- W_l is the fixed effect for covariate l
- β_g is the fixed effect for genotype
- β_k is the fixed effect of the k^{th} expression PC
- β_{gxk} is the fixed effect of the interaction between the k^{th} expression PC and genotype

We use the R package "lme4" to quantify the significance of all K interaction terms: $\beta_{gx1}, \ldots, \beta_{gxk}, \ldots, \beta_{gxK}$. Intuitively, if the k^{th} interaction term (β_{gxk}) is significant, it implies that the eQTL effect size of this variant-pairs significantly changes along expression PC k.

Expression PC interaction eQTLs are closely related to CellRegMap interaction eQTLs [21], except CellRegMap uses MOFA factors [25] instead of gene expression PCs, and CellRegMap treats the interaction eQTL effect sizes as random effects instead of fixed effects. It was shown in Appendix Figure S5 of the CellRegMap manuscript [21] that MOFA factors were highly correlated with gene expression PCs, and CellRegMap interaction eQTL p-values were highly concordant regardless of whether gene expression PCs or MOFA factors were used. Ultimately, we used our own implementation of expression PC interaction eQTL calling simply to minimize modeling assumption differences with SURGE and allow a focused comparison between the use of SURGE latent contexts and expression PCs in interaction eQTL calling.

Application of SURGE to GTEx samples from 10 tissues: expression quantification

To normalize expression from samples from 10 GTEx tissues (Adrenal gland, Colon-sigmoid, Esophagus-Mucosa, Muscle-Skeletal, Pituitary, Skin-not-sun-exposed, Skin-sun-exposed, small-intestine-terminal-ileum, Stomach, Thyroid), we concatenated log-TPM expression measurements across all samples used in the GTEx v8 eQTL analysis for one of those tissues [5]. We also limited to genes that were tested for eQTLs in the GTEx v8 analysis [5] in all 10 tissues. Next, we quantile normalized this matrix to ensure each sample had an equivalent distribution across genes and then standardized each gene (mean 0 and standard deviation 1). We excluded RNA samples that were outliers (Z-score > = 4) according to Mahalanobis distance computed on 80 expression PCs.

Application of SURGE to GTEx samples from 10 tissues: standard eQTL calling

We first tested for standard eQTLs, or association between genotype and the concatenated (across tissues) expression vector described in the "Application of SURGE to GTEx samples from 10 tissues: expression quantification" section. For this analysis, we limited to genes that passed filters described in the "Application of SURGE to GTEx samples from 10 tissues: expression quantification" section. We then limited to variants with MAF > = 0.05 that were less than 50 KB from the transcription start site of a gene. We controlled for the effects of 80 expression PCs and 4 genotype PCs (as recommended by [5] given the sample size). We assessed genome-wide significance according to a gene-level Bonferroni correction, followed by a genome-wide Benjamini–Hochberg correction.

Application of SURGE to GTEx samples from 10 tissues: SURGE optimization

To select a subset of variant-gene pairs to be used for SURGE model optimization, we first limited to variant-gene pairs that were standard eQTLs (FDR < = 0.05; see the "Application of SURGE to GTEx samples from 10 tissues: standard eQTL calling" section). This was done to ensure a higher fraction of the variant-gene pairs used for SURGE optimization were context-specific eQTLs as it is known standard eQTLs are more likely to be context-specific eQTLs than variant-gene pairs that are not standard eQTLs. Furthermore, we limited to the most significant variant per gene among the 2000 most significant genes and removed a variant-gene pair if the variant was already in the training set for its association with a more significant gene. This yielded 1996 genome-wide representative variant-gene pairs used for SURGE optimization. We then ran SURGE under default parameter settings over these representative variant-gene pairs. We included 80 expression PCs and 4 genotype PCs as covariates in SURGE optimization. The converged SURGE model resulted in 15 latent contexts with PVE $> 1e^{-5}$ and hundreds of genome-wide significant SURGE interaction eQTLs (eFDR < = 0.05) (Additional file 1: Fig. S5, Table S1).

We did not control for a random effects term related to sample repeat structure. While this analysis contained multiple RNA-seq samples from the same individual, there were only 5.1 samples per individual on average. This was too few repeated measurements to accurately estimate an additional variance parameter. As a secondary analysis, however, we ran SURGE on this data using a random effect intercept. There are high levels of correlation between the identified SURGE latent contexts between the two models (Additional file 1: Fig. S7).

Application of SURGE to GTEx samples from a single tissue

To run SURGE on GTEx samples from a single GTEx tissue, we took a very similar approach to that described in the "Application of SURGE to GTEx samples from 10 tissues: expression quantification," "Application of SURGE to GTEx samples from 10 tissues: standard eQTL calling," and "Application of SURGE to GTEx samples from 10 tissues: SURGE optimization" sections. The only difference is that we now limit to samples from the tissue of interest. Furthermore, we now only control for 60 expression PCs and 2 genotype PCs during standard eQTL calling and SURGE optimization. The converged model resulted in 1 latent contexts with PVE $> 1e^{-5}$ and 1287 genome-wide significant SURGE interaction eQTLs (eFDR $< = 0.05$).

Application of SURGE to PBMC single-cell eQTL data: pseudocell expression quantification

We imported raw, un-normalized UMI counts from [18]. We used SCRAN [47] to generate log-normalized counts for each cell. We removed genes that were expressed in fewer than 0.5% of cells. We then limited to the top 6000 highly variable genes via the Scanpy function "highly_variable_genes" [48]. We then removed the effects of sequencing batch using Combat [49] as implemented in Scanpy. We then scaled each gene to have mean 0 and variance 1, with a maximum absolute value of 10 to mitigate outlier effects as implemented by "scanpy.pp.scale."

Next, we sought to generate pseudocells that represented groupings of highly correlated cells within an individual. We first removed individuals from this analysis with fewer than 2500 cells. Next, we performed Leiden clustering as implemented by Scanpy [50] independently in each individual using all default parameters, except we used a fine-grained cluster resolution of 10. Here, each Leiden cluster corresponds to a pseudocell. We took the average expression across all cells assigned to the pseudocell to estimate the expression profile of the pseudocell. Finally, we standardized each gene (across pseudocells) to have mean 0 and standard deviation 1, again capping the absolute value of standardized scores to be 10 to mitigate outlier effects. We excluded RNA pseudocells that were outliers (Z-score $> = 4$) according to Mahalanobis distance computed on 30 expression PCs.

Application of SURGE to PBMC single-cell eQTL data: standard eQTL calling

We first tested for standard eQTLs or association between genotype and the expression vector across pseudocells described in the "Application of SURGE to Ye-lab generated single-cell eQTL data: pseudocell expression quantification" section. For this analysis, we limited to genes that passed filters described in the "Application of SURGE to Ye-lab generated single-cell eQTL data: pseudocell expression quantification" section. We then limited to variants with MAF $> = 0.05$ that were less than 200 KB from the transcription start site of a gene. We controlled for the effects of 30 expression PCs and 2 genotype PCs. We controlled for sample-repeat structure stemming from multiple pseudocells originating from the same individual using a random effects intercept for each individual. We assessed genome-wide significance according to a gene-level Bonferroni correction, followed by a genome-wide Benjamini–Hochberg correction.

Application of SURGE to PBMC single-cell eQTL data: SURGE optimization

To select a subset of variant-gene pairs to be used for SURGE model optimization, we first limited to variant-gene pairs that were standard eQTLs (FDR$<=0.05$; see the "Application of SURGE to Ye-lab generated single-cell eQTL data: standard eQTL calling" section). This was done to ensure a higher fraction of the variant-gene pairs used for SURGE optimization were context-specific eQTLs as it is known standard eQTLs are more likely to be context-specific eQTLs than variant-gene pairs that are not standard eQTLs. Furthermore, we limited to the most significant variant per gene among the 2000 most significant genes and removed a variant-gene pair if the variant was already in the training set for its association with a more significant gene. We than ran SURGE under default parameter settings over these representative variant-gene pairs. We included 30 expression PCs and 2 genotype PCs as covariates in SURGE as well as a random effect intercept term for each individual. The converged SURGE model resulted in 6 latent contexts with PVE $> 1e^{-5}$ and hundreds of genome-wide significant SURGE interaction eQTLs (eFDR$<=0.1$) (Additional file 1: Fig. S16, Table S6).

Gene set enrichment analysis

We tested enrichment of genes whose expression levels was highly correlated with SURGE latent contexts (identified when SURGE was applied to single-cell PBMC data) within known gene sets. Specifically for each SURGE latent context, we identified the 50 genes whose expression levels across pseudocells were most strongly correlated (absolute value of correlation coefficient) with the SURGE latent context. We then tested gene set enrichment of these 50 genes relative to all genes that passed filters described in the "Application of SURGE to PBMC single-cell eQTL data: pseudocell expression quantification" section. We tested enrichment of these strongly correlated genes in both the Hallmark gene set and the MSigDB Biological Process gene set [32] (Additional file 1: Table S10, Table S11).

Application of stratified LD score regression (S-LDSC)

Recall, SURGE interaction eQTLs for a specific variant-gene pair can be identified by evaluating the following likelihood (see the "Methods" section "Surge interaction eQTLs" for more details):

$$y_n \sim N\left(\mu + \sum\nolimits_i \alpha_i I[n \in i] + \sum\nolimits_l W_l X_{nl} + \beta_g G_n + \sum\nolimits_k \beta_k U_{nk} + \sum\nolimits_k \beta_{gxk} G_n U_{nk}, \sigma^2\right)$$

$$\alpha_i \sim N(0, \psi^2)$$

Upon maximizing this likelihood (assume $\widehat{\beta_g}$ and $\widehat{\beta_{gxk}}$ are the estimated values of β_g and β_{gxk} that maximize the likelihood), we can estimate the expected eQTL effect size for the variant-gene pair for a particular value of a latent context of U using the following function:

$$\beta^* = \widehat{\beta_g} + U_k^* \widehat{\beta_{gxk}}$$

Here, β^* is the expected eQTL effect size for the particular variant-gene pair when the k^{th} latent context value of U is equal to U_k^*. Ultimately, this enables us to compute the expected eQTL effect size for all variant-gene pairs when the k^{th} latent context value of U is equal to U_k^*.

We use the above expectation to assess how eQTL enrichment in complex trait and disease heritability varied along the SURGE latent contexts. Specifically, for each SURGE latent context, we generated 200 equally spaced positions along the range of SURGE latent context values. For each of those 200 positions, we computed the expected eQTL effect sizes (using the above expectation of β^*) for all variant-gene pairs. We then used the squared expected eQTL effect size across variant-gene pairs as annotation in S-LDSC [23, 35] along with all BaselineLD v2.2 annotations excluding four QTL related annotations ("GTEx_eQTL_MaxCPP," "BLUEPRINT_H3K27acQTL_MaxCPP," "BLUEPRINT_H3K4me1QTL_MaxCPP," "BLUEPRINT_DNA_methylation_MaxCPP"). If a given variant mapped to multiple genes, we used the sum of squared expected eQTL effect sizes across genes as the annotation similar to [16]. This analysis was done for each of the 200 equally spaced positions for each of the 5 SURGE latent contexts identified when SURGE was run on the single-cell PBMC eQTL data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03152-z.

Additional file 1: Supplementary methods. **Fig. S1.** Computational runtime of SURGE latent context inference and SURGE interaction eQTL calling. **Fig. S2.** Evaluation of SURGE's ability to re-capture simulated latent contexts simulations. **Fig. S3.** Evaluation of SURGE's ability to identify correct number of simulated latent contexts in simulations. **Fig. S4.** Proportion of expression variance explained by SURGE latent contexts when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Fig. S5.** Q-Q plot for SURGE interaction eQTLs identified when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Fig. S6.** Q-Q plot for SURGE interaction eQTLs relative to expression PC interaction eQTLs when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Fig. S7.** Comparison of inferred SURGE latent contexts with and without random effects included when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Fig. S8.** SURGE latent context 4 and SURGE latent context 7 are explained by genotype PC1 when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Fig. S9.** Relationship between SURGE latent contexts and xCell cell type enrichment score when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Fig. S10.** Relationship between SURGE latent context 5 and Epithelial cell type enrichment score, and SURGE latent context 6 and Neuron cell type enrichment score. **Fig. S11.** Correlation between SURGE latent contexts and gene expression principal components and genotype principal components when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Fig. S12.** Relationship between SURGE latent context 1 and xCell cell type enrichment score when SURGE was applied to samples from Colon-Sigmoid GTEx v8 tissue. **Fig. S13.** *P* values of SURGE latent context 1 interaction eQTLs when SURGE was applied to samples from only Colon-Sigmoid GTEx v8 tissue compared to pvalues of interaction eQTLs using xCell cell type enrichment score. **Fig. S14.** Distribution of cells per pseudocell and distribution of pseudocells per individual. **Fig. S15.** Proportion of expression variance explained by SURGE latent contexts when SURGE was applied to PBMC pseudocells. **Fig. S16.** Q-Q plot for SURGE interaction eQTLs identified when SURGE was applied to PBMC pseudocells. **Fig. S17.** Q-Q plot for SURGE interaction eQTLs relative to expression PC interaction eQTLs when SURGE was applied to PBMC pseudocells. **Fig. S18.** SURGE latent contexts capture known cell type when SURGE was applied to PBMC pseudocells. **Fig. S19.** UMAP-projected SURGE latent context loadings of pseudocells colored by cell type labels when SURGE was applied to PBMC pseudocells. **Fig. S20.** UMAP-projected SURGE latent context loadings of pseudocells colored by cell type marker genes when SURGE was applied to PBMC pseudocells. **Fig. S21.** SURGE latent context 1 captures differences in disease status of individuals when SURGE was applied to PBMC pseudocells. **Fig. S22.** Correlation between SURGE latent contexts and gene expression principal components when SURGE was applied to PBMC pseudocells. **Fig. S23.** Variance explained of SURGE latent contexts by various pseudocell sample characteristics when SURGE was applied to PBMC pseudocells. **Fig. S24.** Number of colocalizations identified between 15 GWAS studies and various types of eQTLs called on PBMC pseudocells. **Fig. S25.** Number of trait colocalizations with SURGE interaction eQTLs stratified by SURGE latent context when SURGE was applied to PBMC pseudocells. **Fig. S26.** Complex tratit S-LDSC enrichment along eQTLs of SURGE latent contexts. **Table S1.** Number of genes with a genome-wide significant SURGE interaction eQTL according to per "per context empirical FDR" when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Table S2.** Number of genes with a genome-wide significant SURGE interaction eQTL according to per "all context empirical FDR" when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Table S3.** Association significance between tissue identity and SURGE latent context when SURGE was applied to samples concatenated across 10 GTEx v8 tissues.

Table S4. Association significance between known ancestry and SURGE latent context when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Table S5.** *P*-value of association between SURGE latent context and xCell cell type enrichment scores when SURGE was applied to samples concatenated across 10 GTEx v8 tissues. **Table S6.** Number of genes with a genome-wide significant SURGE interaction eQTL according to per "per context empirical FDR" when SURGE was applied to PBMC pseudocells. **Table S7.** Number of genes with a genome-wide significant SURGE interaction eQTL according to per "all context empirical FDR" when SURGE was applied to PBMC pseudocells. **Table S8.** Association significance between known cell type and SURGE latent context when SURGE was applied to PBMC pseudocells. **Table S9.** Association significance between known fine-grained cell type and SURGE latent context when SURGE was applied to PBMC pseudocells. **Table S10.** Hallmark gene set enrichment analysis of genes strongly correlated with SURGE latent contexts when SURGE was applied to PBMC pseudocells. **Table S11.** MSigDB Biological Process gene set enrichment analysis of genes strongly correlated with SURGE latent contexts when SURGE was applied to PBMC pseudocells.

Additional file 2. Review history.

Acknowledgements

We thank Alkes L. Price for insights into the S-LDSC analysis and for providing comments on the manuscript.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 2.

Authors' contributions

AB and BJS proposed the idea for the project. BJS developed the SURGE model and performed the analysis. BJS and AB wrote the manuscript. KT aided in deriving the variational updates for SURGE. JP provided review of SURGE code. JP, GQ, and AB suggested relevant downstream analysis to run based on SURGE output. MGG, RP, and CJY provided guidance on analysis of single-cell eQTL data.

Funding

AB was supported by NIH/NIGMS R35GM139580, NIH/NIDDK R01DK122586, and the Chan Zuckerberg Initiative. BJS was funded in part by U01 HG012009.

Availability of data and materials

SURGE software is available on GitHub at https://github.com/BennyStrobes/surge under an MIT license. The GTEx v8 data [5] can be downloaded from the dbGaP website under phs000424.v8.p2 and on the GTEx portal (http://gtexportal.org/). PBMC single-cell eQTL [18] expression data are available in the Human Cell Atlas Data Coordination Platform and at GEO accession number GSE174188. PBMC single-cell eQTL [18] genotype data are available at dbGaP accession number phs002812.v1.p1. Additionally, the SURGE source code used in this study is also accessible on Zenodo (https://doi.org/https://doi.org/10.5281/zenodo.10383060) [51].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

AB is a shareholder of Alphabet, Inc., and a consultant for Third Rock Ventures.

Author details

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [2]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]Biological and Medical Informatics Graduate Program, University of California, San Francisco, CA, USA. [5]Division of Rheumatology, Department of Medicine, University of California, San Francisco, CA, USA. [6]Institute for Human Genetics, University of California, San Francisco, CA, USA. [7]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA. [8]Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, CA, USA. [9]Chan-Zuckerberg Biohub, San Francisco, CA, USA. [10]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [11]Department of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA.

Received: 22 December 2022 Accepted: 20 December 2023
Published online: 22 January 2024

References

1. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci. 2013;368:20120362.
2. Lappalainen T, The Geuvadis Consortium, Sammeth M, Friedländer MR, 'tHoen PAC, Monlong J, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
3. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24:14–24.
4. Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet. 2021;53:1290–9.
5. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
6. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10.
7. Knowles DA, Burrows CK, Blischak JD, Patterson KM, Serie DJ, Norton N, et al. Determining the genetic basis of anthracycline-cardiotoxicity by molecular response QTL mapping in induced cardiomyocytes. Elife. 2018;7: e33480.
8. Strober BJ, Elorbany R, Rhodes K, Krishnan N, Tayeb K, Battle A, et al. Dynamic genetic regulation of gene expression during cellular differentiation. Science. 2019;364:1287–90.
9. Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nat Commun. 2020;11:810. https://doi.org/10.1038/s41467-020-14457-z.
10. Jerber J, Seaton DD, Cuomo ASE, Kumasaka N, Haldane J, Steer J, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nat Genet. 2021;53:304–12.
11. Umans BD, Battle A, Gilad Y. Where are the disease-associated eQTLs? Trends Genet. 2021;37:109–24.
12. Elorbany R, Popp JM, Rhodes K, Strober BJ, Barr K, Qi G, et al. Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. PLoS Genet. 2022;18:e1009666.
13. Nathan A, Asgari S, Ishigaki K, Valencia C, Amariuta T, Luo Y, et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. Nature. 2022;606:120–8.
14 Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science. 2022;376:eabf3041.
15. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet. 2017;49:600–5.
16. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet. 2020;52:626–33.
17. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. bioRxiv. 2022;2022.05.07.491045. https://doi.org/10.1101/2022.05.07.491045
18 Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science. 2022;376:eabf1970.
19. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, LifeLines Cohort Study, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50:493–7.
20. Findley AS, Monziani A, Richards AL, Rhodes K, Ward MC, Kalita CA, et al. Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions. Elife. 2021;10:e67077. https://doi.org/10.7554/eLife.67077.
21. Cuomo ASE, Heinen T, Vagiaki D, Horta D, Marioni JC, Stegle O. Cell RegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. Mol Syst Biol. 2022;18:e10663.
22. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLos Genet. 2014;10:e1004383.
23. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228–35.
24. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type-specific genetic regulation of gene expression across human tissues. Science. 2020;369:eaaz8528. https://doi.org/10.1126/science.aaz8528.
25. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21:111.
26. Wipf D, Nagarajan S. A new view of automatic relevance determination. Available: https://papers.nips.cc/paper/2007/file/9c01802ddb981e6bcfbec0f0516b8e35-Paper.pdf.Cited 22 Nov 2022
27. Vochteloo M, Deelen P, Vink B, Tsai EA, Runz H, Andreu-Sánchez S, et al. Unbiased identification of unknown cellular and environmental factors that mediate eQTLs using principal interaction component analysis. bioRxiv. 2022. https://doi.org/10.1101/2022.07.28.501849
28. Gewirtz AD, Townes FW, Engelhardt BE. Telescoping bimodal latent Dirichlet allocation to identify expression QTLs across tissues. Life Sci Alliance. 2022;5:e202101297. https://doi.org/10.26508/lsa.202101297.
29. Gewirtz ADH, Townes FW, Engelhardt BE. Expression QTLs in single-cell sequencing data. bioRxiv. 2022. https://doi.org/10.1101/2022.08.14.503915
30. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18:220.
31. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019;20:206.
32. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.
33. Theofilopoulos AN, Koundouris S, Kono DH, Lawson BR. The role of IFN-gamma in systemic lupus erythematosus: a challenge to the Th1/Th2 paradigm in autoimmunity. Arthritis Res. 2001;3:136–41.

34. Schroder K, Hertzog PJ, Ravasi T, Hume DA. Interferon-gamma: an overview of signals, mechanisms and functions. J Leukoc Biol. 2004;75:163–89.
35. Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat Genet. 2017;49:1421–7.
36. du Pré MF, Sollid LM. T-cell and B-cell immunity in celiac disease. Best Pract Res Clin Gastroenterol. 2015;29:413–23.
37. Jagadeesh KA, Dey KK, Montoro DT, Mohan R, Gazal S, Engreitz JM, et al. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. Nat Genet. 2022;54:1479–92.
38. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14: e8124.
39. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. J Am Stat Assoc. 2017;112:859–77.
40. Wang W, Stephens M. Empirical Bayes matrix factorization. arXiv [stat.ME]. 2018. Available: http://arxiv.org/abs/1802.06931
41. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics. 2015;31:545–54.
42. Chung NC. Statistical significance of cluster membership for unsupervised evaluation of cell identities. Bioinformatics. 2020;36:3107–14.
43. Chen YT, Witten DM. Selective inference for k-means clustering. arXiv [stat.ME]. 2022. Available: http://arxiv.org/abs/2203.15267
44. Neufeld A, Gao LL, Popp J, Battle A, Witten D. Inference after latent variable estimation for single-cell RNA sequencing data. arXiv [stat.ME]. 2022. Available: http://arxiv.org/abs/2207.00554
45. Gamazon ER, Huang RS, Dolan ME, Cox NJ, Im HK. Integrative genomics: quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data. Front Genet. 2012;3:202.
46. Knowles DA, Davis JR, Edgington H, Raj A, Favé M-J, Zhu X, et al. Allele-specific expression reveals interactions between genetic variation and environment. Nat Methods. 2017;14:699–702.
47. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:75.
48 Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15. https://doi.org/10.1186/s13059-017-1382-0.
49 Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2:lqaa078.
50. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:5233.
51. Strober BJ, Tayeb K, Popp J, Qi G, Gordon M, Perez R, Ye C, Battle A. SURGE: uncovering context-specific genetic-regulation of gene expression from single-cell RNA-sequencing using latent-factor models. https://github.com/bennystrobes/surge https://doi.org/10.5281/zenodo.10383060 (2023).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.