RESEARCH ARTICLE

Discovery and characterization of variance QTLs in human induced pluripotent stem cells

Abhishek K. Sarkar[1☯]*, **Po-Yuan Tung**[1☯], **John D. Blischak**[1], **Jonathan E. Burnett**[1], **Yang I. Li**[1,2], **Matthew Stephens**[1,3], **Yoav Gilad**[1,2]*

1 Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America, **2** Department of Medicine, University of Chicago, Chicago, Illinois, United States of America, **3** Department of Statistics, University of Chicago, Chicago, Illinois, United States of America

☯ These authors contributed equally to this work.
* aksarkar@uchicago.edu (AKS); gilad@uchicago.edu (YG)

Abstract

Quantification of gene expression levels at the single cell level has revealed that gene expression can vary substantially even across a population of homogeneous cells. However, it is currently unclear what genomic features control variation in gene expression levels, and whether common genetic variants may impact gene expression variation. Here, we take a genome-wide approach to identify expression variance quantitative trait loci (vQTLs). To this end, we generated single cell RNA-seq (scRNA-seq) data from induced pluripotent stem cells (iPSCs) derived from 53 Yoruba individuals. We collected data for a median of 95 cells per individual and a total of 5,447 single cells, and identified 235 mean expression QTLs (eQTLs) at 10% FDR, of which 79% replicate in bulk RNA-seq data from the same individuals. We further identified 5 vQTLs at 10% FDR, but demonstrate that these can also be explained as effects on mean expression. Our study suggests that dispersion QTLs (dQTLs) which could alter the variance of expression independently of the mean can have larger fold changes, but explain less phenotypic variance than eQTLs. We estimate 4,015 individuals as a lower bound to achieve 80% power to detect the strongest dQTLs in iPSCs. These results will guide the design of future studies on understanding the genetic control of gene expression variance.

Author summary

Common genetic variation can alter the level of average gene expression in human tissues, and through changes in gene expression have downstream consequences on cell function, human development, and human disease. However, human tissues are composed of many cells, each with its own level of gene expression. With advances in single cell sequencing technologies, we can now go beyond simply measuring the average level of gene expression in a tissue sample and directly measure cell-to-cell variance in gene expression. We hypothesized that genetic variation could also alter gene expression variance, potentially revealing new insights into human development and disease. To test this hypothesis, we used single cell RNA sequencing to directly measure gene expression variance in multiple

OPEN ACCESS

Citation: Sarkar AK, Tung P-Y, Blischak JD, Burnett JE, Li YI, Stephens M, et al. (2019) Discovery and characterization of variance QTLs in human induced pluripotent stem cells. PLoS Genet 15(4): e1008045. https://doi.org/10.1371/journal.pgen.1008045

Editor: Chris Cotsapas, Yale School of Medicine, UNITED STATES

Received: September 21, 2018

Accepted: February 22, 2019

Published: April 19, 2019

Data Availability Statement: The RNA-seq data have been deposited in Gene Expression Omnibus under accession number GSE118723.

Funding: This work was supported by NIH grant R01GM122930 awarded to Y.G. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

individuals, and then associated the gene expression variance with genetic variation in those same individuals. Our results suggest that effects on gene expression variance are smaller than effects on mean expression, relative to how much the phenotypes vary between individuals, and will require much larger studies than previously thought to detect.

Introduction

Robustness, or the ability to maintain a stable phenotype despite genetic mutations and environmental perturbations, is an important property of many key biological processes, such as those underlying embryogenesis and development [1, 2]. Conversely, evolvability, or the ability to generate heritable phenotypic variation, is a fundamental requirement of evolutionary processes [3]. A long-standing question in genetics, therefore, is how the balance between these two seemingly opposite processes has been fine-tuned [4].

To make progress in understanding the balance between robustness and evolvability, we need to characterize the mechanisms that underlie robustness. Robustness can arise through a number of different mechanisms: for example, redundancy of regulatory elements or feedback loops in regulatory circuits. In these different scenarios, we hypothesize evolvability could be maintained through different selective pressures. If we are able to characterize gene-specific robustness to expression variability, we can begin to ask about the balance between natural selection of gene function and the ability to maintain evolvability.

In model organisms, robustness and evolvability can be studied using experimental evolution approaches. These approaches typically quantify robustness as the change in trait variation after applying an experimental perturbation [5, 6]. However, in such experiments the phenotypic outcomes, rather than the underlying mechanisms of robustness, are measured. Moreover, experimental evolution studies have almost always considered population-average measurements of phenotypes using entire organisms, tissues, or cell cultures, with few exceptions [7, 8]. To truly understand how robustness and evolvability are established and encoded in the genome, we need to consider phenotypic variation across individual cells [9], and connect it to genetic variation, an approach termed “noise genetics” [10].

Using the yeast *Saccharomyces cerevisiae* as a model system, studies have shown that heterogeneity in the expression of certain genes across cells is highly heritable and placed under complex genetic control, suggesting that the level of noise in gene regulation may also differ between individuals of multicellular organisms depending on their genetic background [11]. Follow-up studies further demonstrated that gene expression noise mediated by promoter variants can provide a fitness benefit at times of environmental stress in yeast, highlighting the direct role of genetically controlled stochastic cell-cell variation in evolutionary robustness [12]. However, the genetic and molecular circuits that lead to robustness remain largely uncharacterized in mammals.

Here, we take an unbiased, genome-wide approach to identify quantitative trait loci associated with gene expression variance across cells (vQTLs). We study human induced pluripotent stem cells (iPSCs), which offer a homogeneous population of cells allowing a relatively simple statistical model. Investigating iPSCs also provides the possibility to study gene expression variance across cells during differentiation in follow-up studies. To directly measure the mean and variance of gene expression within cell populations as phenotypes, we generated single cell RNA-seq (scRNA-seq) data from cells derived from multiple individuals.

Results

Sample collection and quality control

Using the Fluidigm C1 platform, we isolated and collected scRNA-seq from 7,585 single cells from iPSC lines of 54 Yoruba in Ibadan, Nigeria (YRI) individuals. We used unique molecular identifiers (UMIs) to tag RNA molecules and account for amplification bias in the single cell data [13]. To estimate technical confounding effects without requiring separate technical replicates, we used a mixed-individual plate study design (Fig 1A). The key idea of this approach is that having observations from the same individual under different confounding effects and observations from different individuals under the same confounding effect allows us to distinguish the two sources of variation [14].

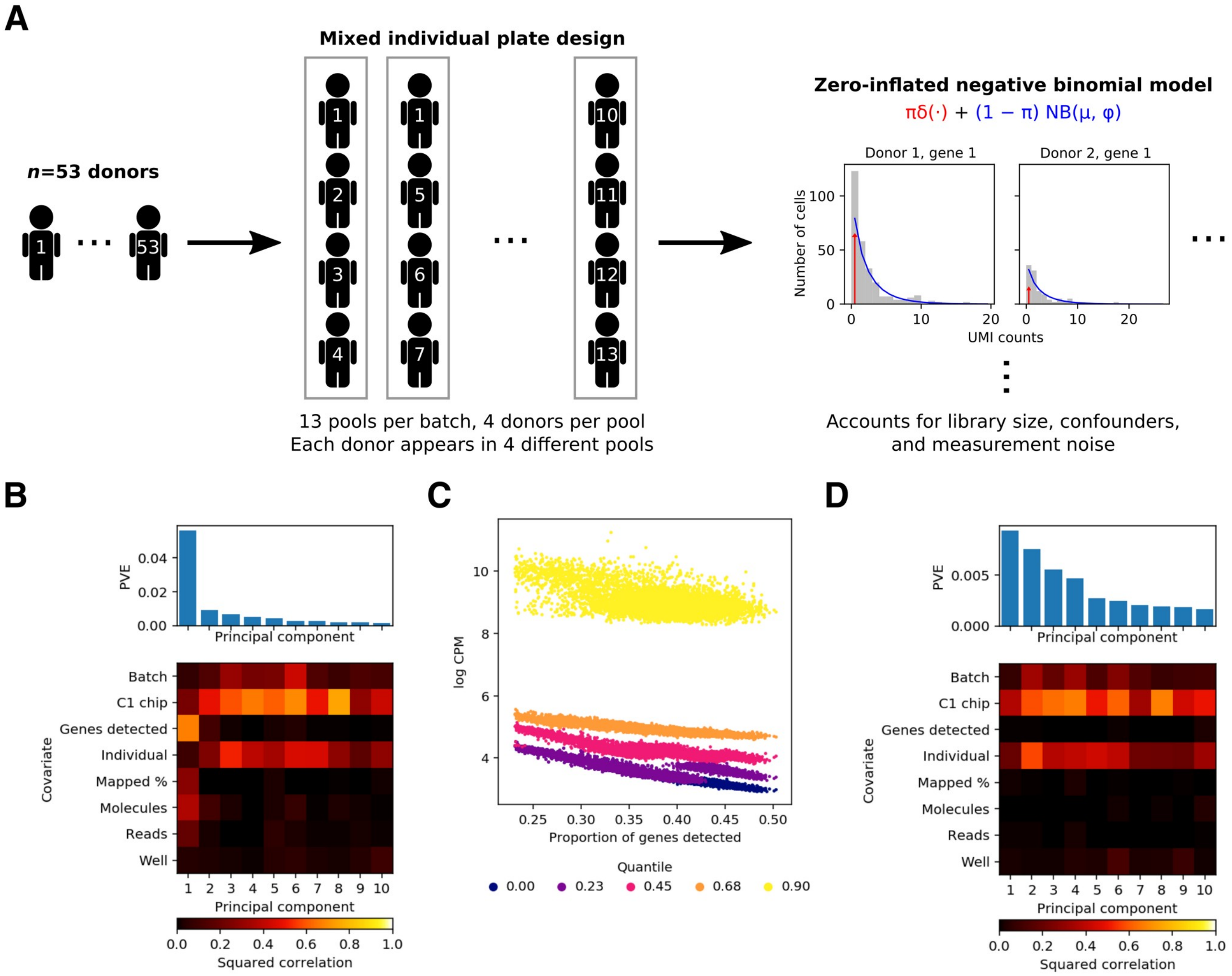

Fig 1. Study design and quality control. (A) We used a mixed-individual plate design to be able to distinguish technical effects from biological effects of interest, and used a zero-inflated negative binomial model to fit the distribution of the data, accounting for technical confounders. (B) Proportion of variance explained (PVE; top) and a heatmap of the correlations between the top 10 principal components of gene expression and observed technical covariates (bottom). (C) Dependence of the distribution of gene expression against gene detection rate (proportion of genes with at least one molecule detected) for each sample. Each vertical slice is a single cell (according to the gene detection rate). For each cell, there are 5 points, corresponding to the (0, 0.25, 0.5, 0.75, 1) quantiles of non-zero log CPM values observed for that cell. (D) PVE and correlation between principal components and observed covariates after correcting for gene detection rate.

https://doi.org/10.1371/journal.pgen.1008045.g001

We excluded data from one individual (NA18498) with evidence of contamination, then filtered poor quality samples as previously described [14]. After quality control, we analyzed the expression of 9,957 protein-coding genes in a median of 95 cells per individual in 53 individuals (total of 5,597 cells; S1 Fig).

To ensure that our measurements are comparable across samples, we first sought to assess the impact of observed technical variation on the data and to identify unobserved technical confounders. To this end, we performed principal components analysis (PCA) on the matrix of log counts per million (log CPM).

We found that across samples, the loading on the top principal component (PC) was correlated with gene detection rate (the proportion of genes with at least one molecule detected), but not with the biological variable of interest (individual) or the expected technical confounders (batch and C1 chip; Fig 1B). Indeed, as previously reported [15], the entire distribution of observed log CPM (over all genes) varies across samples, and appears to be associated with the gene detection rate (Fig 1C). After accounting for gene detection rate (Methods), the top PCs were correlated with individual, batch, and C1 chip, as expected (Fig 1D).

Estimating gene expression mean and variance

We developed a method to estimate the mean and variance of gene expression across cells for each gene in each individual (Fig 1A; Methods, S1 Text). Briefly, for each individual and each gene, our method uses maximum likelihood to fit a zero-inflated negative binomial distribution (ZINB) to the observed UMI counts across cells, and derives the mean and variance of gene expression from the estimated model parameters. When fitting the ZINB model the method controls for technical confounders (e.g. C1 chip) and library size, and when deriving the mean and variance it accounts for Poisson measurement noise in the UMI counts [16, 17]. These desirable, and arguably crucial features would not be achieved by directly computing the sample mean and variance of either the UMI counts or log CPM.

To evaluate the accuracy of the method, we first simulated data from the model and compared the estimated parameters, as well as the derived mean and variance, to the true values used to generate the data. We fixed the number of cells and number of molecules detected per cell to the median of those values in our observed data, and varied the ZINB parameters. Assuming that mean expression is high enough, we found the method produces accurate estimates of the underlying negative binomial parameters, but not the zero inflation parameter (S2 Fig). Despite not accurately estimating the zero inflation parameter, the method still produces accurate estimates of the derived mean and variance for genes that are expressed at intermediate to high levels.

Next, we tested for goodness of fit on each simulated data set (Methods). The key idea underlying the test is that if the data are truly distributed according to some cumulative distribution function *F*, then the values of *F* evaluated at the data should be uniformly distributed between 0 and 1. Applying the test to the simulated data, we rejected the null that the model fit the data for zero of 2,451 simulation trials after Bonferroni correction ($p < 2 \times 10^{-5}$; S3 Fig). The results suggest the method is successfully able to fit the observed data, and also suggest that inaccuracy in the estimated parameters is likely explained by noise due to small sample sizes.

We then applied our method to the observed data, correcting for batch and C1 chip. Importantly, we did not correct for gene detection rate, reasoning that the dependence on gene detection rate is only an artifact introduced by analyzing log CPM. We tested the goodness of fit for each individual and each gene, and rejected the null that the model fit the data for only 60 of 537,658 individual-gene combinations (0.01%) after Bonferroni correction ($p < 9 \times 10^{-8}$;

S4 Fig). Our results emphasize that careful experimental design as well as careful statistical modeling are required to robustly map effects on gene expression variance across cells.

Quantitative trait locus mapping

Previous studies have shown a clear relationship between the mean and variance of gene expression [18, 19]; therefore, apparent genetic effects on the variance could potentially be explained by effects on the mean. In our model, the mean-variance relationship is controlled by a single dispersion parameter per gene per individual. We sought to directly map QTLs which could alter the variance independently of altering the mean by using the estimated dispersion parameter as a quantitative phenotype. However, we found zero dispersion QTLs (dQTLs) using this approach (FDR 10%). Further, we found the QQ plot of association p-values did not show deviation from the null (S5 Fig).

Alternative approaches to decouple the mean-variance relationship include using the coefficient of variance (CV; ratio of standard deviation to mean) or Fano factor (ratio of variance to mean) as quantitative phenotypes. However, prior work shows these quantities have predictable relationships with the mean, and therefore effects could still be explained away [14, 19]. Therefore, we proceeded to map eQTLs, variance QTLs (vQTLs), CV-QTLs, and Fano-QTLs, and then asked whether we could discover variance effects which could not be explained as effects on mean expression.

We found 235 eQTLs, 5 vQTLs, 0 CV-QTLs, and 0 Fano-QTLs (FDR 10%; S5 Fig). To validate the eQTLs, we estimated the replication rate against eQTLs discovered in bulk RNA-seq from the same iPSC lines [20]. We found that 79% of the single cell eQTLs replicate in the matched bulk data (Fig 2A), and 80% of bulk eQTLs replicate in the single cell data. However, we found 1,390 eQTLs (FDR 10%) using all of the individuals in the bulk RNA-seq study ($n = 58$), and still recovered 1,136 eQTLs (FDR 10%) after subsampling to $n = 53$. Our results therefore suggest that eQTL discovery in scRNA-seq (as opposed to replication of previously discovered eQTLs) loses power compared to equal-sized studies in bulk RNA-seq, likely due to increased experimental noise.

We found 85% of the eQTLs were also discovered as vQTLs (when restricting to testing only at the eQTL SNP), and 100% of vQTLs were discovered as eQTLs (Fig 2B). We then sought to directly explain away vQTLs as eQTLs by regressing out the mean from the variance. Treating the residuals from the regression as the phenotype, we recovered zero vQTLs. These results suggest the significant variance effects detected in this study are all likely to be explained as effects on mean expression.

Power analysis

Our goal in this study was to find QTLs which alter the variance of gene expression independently of altering the mean expression. Under our model, these QTLs should explain variation in the dispersion parameter across individuals; however, we failed to find dQTLs. Further, all of the vQTLs we were able to identify could be explained by mean effects. In contrast, we were able to discover eQTLs, but fewer than expected based on bulk RNA-seq in matched samples.

To understand why we failed to discover dQTLs, and why we discovered fewer eQTLs than expected, we first derived the power function in terms of effect size (log fold change), sample size, noise ratio (ratio of measurement error variance to phenotypic residual variance), and significance level (Methods). We then sought to estimate the distribution of QTL effect sizes and the typical noise ratio, for both mean expression and dispersion.

To estimate the distribution of QTL effect sizes, we fit a flexible unimodal distribution for the true effect sizes which maximizes the likelihood of the observed effect sizes and standard

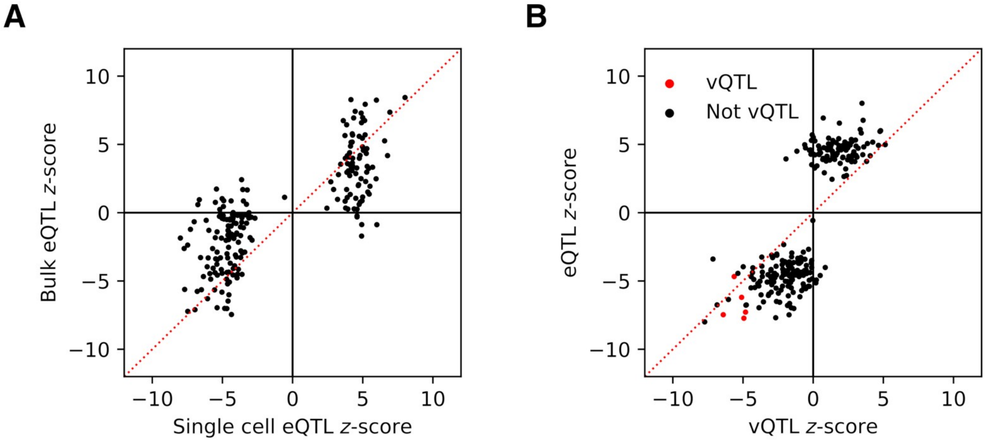

Fig 2. Discovery and overlap of expression QTLs and variance QTLs. (A) *z*-scores for expression QTL (eQTL) SNP-gene pairs discovered in pooled single cell RNA-seq data against *z*-scores of the same SNP-gene pairs in matched bulk RNA-seq data. (B) In the single cell data, *z*-scores for eQTL SNP-gene pairs (FDR 10%) against variance QTL (vQTL) *z*-scores of the same SNP-gene pairs. vQTL *z*-scores are stratified based on whether the gene was discovered as a vQTL at FDR 10%.

https://doi.org/10.1371/journal.pgen.1008045.g002

errors [21]. Surprisingly, we found that dQTL effects could be larger than eQTL effects (S6 Fig). For example, we estimate that the 99th percentile eQTL effect size is 0.022, but is 0.090 for dQTLs. Given this result and the power function we derived, there are two possible explanations for why we still failed to find dQTLs: (1) the noise ratio of dispersion is large (measurement error reduced power), or (2) the residual variance of dispersion is large (genetic variation explains little phenotypic variance).

To estimate the typical noise ratio, we developed a two-step procedure to estimate the measurement error variance and residual variance per gene (Methods). Briefly, in our approach we have one measurement error variance per individual, per gene, which equals the sampling variance of our ZINB model. To estimate each error variance, we used non-parametric bootstrapping. To estimate the measurement error variance for each gene, we took the median of the estimated measurement error variances across individuals. To estimate the residual variance for each gene, we fit a flexible unimodal distribution for the true phenotypes which maximizes the likelihood of the observed phenotypes and measurement errors, and estimated the variance of the posterior mean true phenotypes.

Using our approach, we estimated that the typical noise ratio of the dispersion is 2.99, compared to 4.18 for the mean (S7 Fig). This result suggests that we did not fail to find dQTLs only due to measurement error, because the noise ratio was lower for dQTLs than for eQTLs. As a reference point, a noise ratio equal to 1 has the same impact on power to detect a QTL as cutting the sample size in half, explaining why our study lost power to detect eQTLs. We found that the typical phenotypic standard deviation of dispersion is 7.2 fold larger than that of the mean expression, suggesting we failed to find dQTLs because the effect sizes of dQTLs (relative to phenotypic standard deviation) are smaller than the effect sizes of eQTLs.

We finally asked how much power our current study had to detect the 99th percentile dQTL effect size, assuming the typical noise ratio estimated above. We found that our study had only 0.001% power to detect that effect size at Bonferroni–corrected level $\alpha = 5 \times 10^{-6}$ (Fig 3). Fixing the typical noise ratio (a function of the number of cells per individual and sequencing depth), we estimate 16,015 individuals would be required to achieve 80% power. As a lower bound (setting the noise ratio to zero), we estimate 4,015 individuals would be required

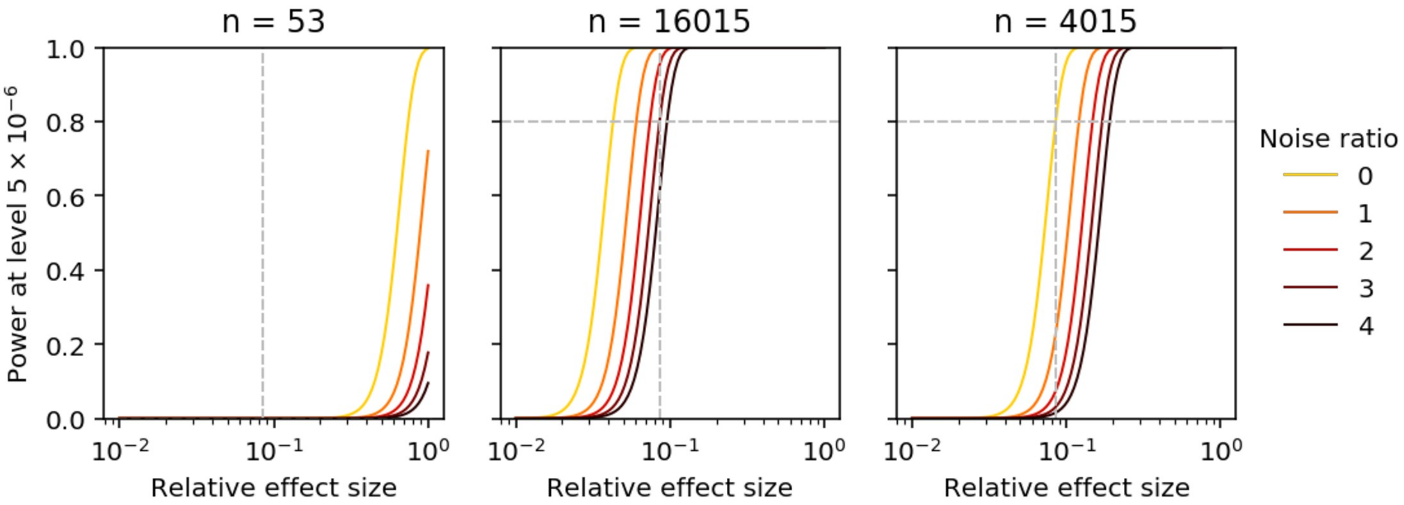

Fig 3. Power to detect dispersion QTLs. Power is a function of effect size (relative to phenotypic standard deviation), sample size, noise ratio, and significance level. Gray lines indicate the 99th percentile of dispersion effect sizes relative to the typical phenotypic standard deviation, and the power achieved to detect an effect of that size at the typical noise ratio. Power curves are computed for the current sample size (left), the sample size required to achieve 80% power for that effect size fixing the number of cells per individual (center), and the minimum sample size assuming no measurement error (right).

https://doi.org/10.1371/journal.pgen.1008045.g003

regardless of the number of cells per individual. Overall, our results suggest a much larger study, both in terms of number of individuals and number of cells per individual, would be required to detect the strongest dQTLs in iPSCs.

Discussion

Individual cells must tolerate both external and internal perturbations arising from the environment or mutations. It has long been argued that this outcome of robustness is an inherent property of biological systems [22], and arises from natural selection [23, 24]. Robustness is especially critical in the context of cell fate transitions during differentiation [25]. Other dynamic physiological processes must also be robust, and as a result, loss of robustness is associated with clinically relevant phenotypes and complex genetic disease [26, 27].

Cells maintain their identity and other phenotypes despite perturbations because of the robust regulation of key sets of genes [28]. We hypothesized that QTLs could disrupt the mechanisms underlying robust regulation, and therefore reveal new insights into the genetic regulation of differentiation and disease.

To investigate this hypothesis, we directly observed gene expression variance across multiple individuals using scRNA-seq, and sought to identify QTLs which could alter the variance of gene expression across cells within a single individual, independently of altering the mean expression. However, we failed to discover such QTLs, and demonstrated that QTLs which are associated with the variance of gene expression can be explained by effects on mean expression. We found that relative to the phenotypic standard deviation, effects on the dispersion are smaller than effects on the mean, partially explaining why this study failed to find them.

Our results do not rule out genetic effects on variance independent of mean effects, due to limitations of our analysis. First, our estimated distributions of effect sizes are based on an empirical Bayes estimate of the underlying effect sizes, given the observed effect sizes. Our results in simulation and observed data suggest the observed effect sizes may be not be accurately estimated given the size of the current study. Therefore, the empirical Bayes estimate may not accurately reflect the true distribution of effect sizes. However, we chose to bias the estimation procedure towards putting prior mass on zero, so our estimates of effect sizes are conservative. Additionally, our estimates may not generalize beyond iPSCs, because the distribution of dispersion effect sizes could vary across cell types and conditions.

Second, we made a strong assumption that latent gene expression is point-Gamma distributed. In this study, we directly assessed whether or not this was true using a simple statistical diagnostic, and did not find any gross violations of this assumption in the data. However, it is likely that this assumption will be violated in heterogenous populations of cells. One possible extension of our method to this case would be to assume there are K homogeneous subpopulations of cells, each described by a (possibly different) point-Gamma distribution. This mixture of ZINB model suggests an expectation-maximization approach where each cell is assigned to a subpopulation, and then the distributions of the subpopulations are re-estimated.

Finally, we took a modular approach to map QTLs in this study: (1) we estimated parameters for each individual using only the scRNA-seq data, and then (2) we mapped QTLs using phenotypes derived from the estimated parameters. An alternative approach would be to include genotype in the count model for the data, and jointly learn the mean, dispersion, proportion of excess zeros, and genetic effect sizes for mean and dispersion. Such an approach could borrow information across cells with common genotypes to improve power, holding the experiment size fixed. However, further development will be needed to efficiently fit the models at QTL mapping scale.

We stress that our power calculation is only a rough guideline for designing QTL mapping studies using scRNA-Seq. Intuitively, some minimum number of cells per individual is required to adequately estimate means and variances. However, having achieved that lower bound, the most important quantity to maximize is the number of individuals. In support of this argument, we estimate thousands of individuals would be required to detect a dQTL no matter how many cells were collected per individual.

We based our power calculations on typical values of the noise ratio for the mean expression and dispersion, and chose a conservative significance level. However, we found considerable variation in the noise ratio across genes, suggesting that our results may not generalize even across genes. Overall, our results suggest that the technical noise introduced by scRNA-seq greatly reduces the power to discover eQTLs. Our results also suggest that, for iPSC lines, dramatically larger studies will be required to map both eQTLs and dQTLs from scRNA-seq.

Materials and methods

Ethics statement

The cell lines used in this study were obtained from the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research. All samples were collected by the Coriell Institute for Medical with written informed consent and with IRB approval.

Sample collection and quality control

We cultured YRI iPSCs [20] in feeder-free conditions for at least ten passages in E8 medium (Life Technologies) [29]. We collected cells using the C1 Single-Cell Auto Prep IFC microfluidic chip (Fluidigm). We used a balanced block-incomplete design to randomize individuals across chips. For each chip, we freshly prepared a mixture of cell suspensions from four individuals. We measured live cell number via trypan blue staining (ThermoFisher), to ensure equal cell numbers across individuals per mixture. We performed single cell capture and library preparation as previously described using 6 bp Unique Molecular Identifiers [14]. We pooled the 96 samples on each C1 chip and sequenced them on an Illumina HiSeq 2500 using the TruSeq SBS Kit v3-HS (FC-401-3002).

We mapped the reads to human genome GRCh37 (including the ERCC spike-ins) with *Subjunc* [30], deduplicated the UMIs with *UMI-tools* [31], and counted molecules per protein-coding gene (Ensembl 75) with *featureCounts* [32]. We then matched single cells back to YRI individuals using *verifyBamID* [33].

We filtered samples on the following criteria, derived as previously described [14]:

- Only one cell observed per well
- Valid identification
- At least 1,011,612 mapped reads
- Less than 49% ERCC reads
- At least 4,730 genes with at least one read
- Linear discriminant analysis predicts one cell

We filtered genes for QTL mapping on the following criteria:

- Number of molecules less than $4^6 = 4096$
- Median log CPM at least 3

We applied principal component analysis (PCA) to the matrix **X** of log counts per million (log CPM), using the pseudocount proposed in *edgeR* [34].

We corrected for gene detection rate by simultaneously regressing out quantiles of gene expression, correcting for sample-specific and gene-specific means, and performing PCA. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be observed p-vectors, and let $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ be latent k-vectors where $k \ll p$. Then, PCA corresponds to maximum likelihood estimation in the following latent variable model [35]:

$$\mathbf{x}_i \sim \mathcal{N}(\cdot; \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{1}$$

In this parameterization, $\boldsymbol{\mu}$ denotes a per-coordinate mean (in our application, per-gene). However, as previously reported [15], we additionally have to account for the per-sample mean.

Our approach is based on the latent variable model:

$$x_{ij} \sim \mathcal{N}(\mathbf{W}_j\mathbf{z}_i + \mathbf{q}_i'\boldsymbol{\beta}_j + u_i + v_j, \sigma^2\mathbf{I}) \tag{2}$$

where **u** is an n-vector of per-sample means, **v** is a p-vector of per-gene means, and $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_n)$ is a $n \times k$ matrix of expression quantiles.

We fit the model as follows:

1. Estimate $\boldsymbol{\beta}_j$ via least squares estimation of the following linear model:

$$\mathbf{X}_j = \mu_j + \mathbf{Q}\boldsymbol{\beta}_j + \epsilon \tag{3}$$

2. Construct the residual matrix $x_{ij} := x_{ij} - \mathbf{q}_i' \boldsymbol{\beta}_j$, then estimate $\mathbf{u}$, $\mathbf{v}$ via coordinate descent:

$$u_i := \frac{1}{p} \sum_j x_{ij} - v_i \tag{4}$$

$$v_j := \frac{1}{n} \sum_j x_{ij} - u_i \tag{5}$$

3. Construct the residual matrix $x_{ij} := x_{ij} - u_i - v_j$, then estimate $\mathbf{W}$ via maximum likelihood. The MLE $\hat{\mathbf{W}}$ equals the top k singular vectors of residual matrix $\mathbf{X}$ [35].

We estimated the squared correlation between each PC and categorical covariates (batch, C1 chip, individual, well) by recoding each category as a binary indicator, fitting a multiple linear regression of the PC loadings against the binary indicators, and then estimating the coefficient of determination of the model.

Estimating gene expression mean and variance

We assume the count data are generated by a zero-inflated negative binomial (ZINB) distribution (S1 Text). Let:

- r_{ijk} be the number of molecules for individual i, cell j, gene k
- R_{ij} be a size factor for each cell
- μ_{ik} be proportional to relative abundance
- ϕ_{ik} be the variance of expression noise
- π_{ik} be the proportion of excess zeros
- $\mathbf{x}_{ij}$ be a q-vector of confounders per cell
- $\boldsymbol{\beta}_k$ be a q-vector of confounding effects on gene k

Then, we assume:

$$r_{ijk} \sim \text{Poisson}(\cdot; R_{ij} \exp(\mathbf{x}_{ij}' \boldsymbol{\beta}_k) \lambda_{ijk}) \tag{6}$$

$$\lambda_{ijk} \sim \pi_{ik} \delta_0(\cdot) + (1 - \pi_{ik}) \text{Gamma}(\cdot; \mu_{ik}, \phi_{ik}) \tag{7}$$

Under this model, the mean and variance of gene expression are:

$$\mathbb{E}[\lambda_{ijk}] = (1 - \pi_{ik}) \mu_{ik} \tag{8}$$

$$\mathbb{V}[\lambda_{ijk}] = (1 - \pi_{ik}) \mu_{ik}^2 \phi_{ik} + \pi_{ik} (1 - \pi_{ik}) \mu_{ik}^2 \tag{9}$$

Considering just the non-zero component, marginalizing out λ yields the negative binomial (NB) log likelihood, weighted by $1 - \pi_{ik}$:

$$\begin{aligned} l(\cdot) = {} & \ln(1 - \pi_{ik}) + r_{ijk} \ln\left(\frac{R_{ij}\exp(\mathbf{x}'_{ij}\boldsymbol{\beta}_k)\mu_{ik}}{R_{ij}\exp(\mathbf{x}'_{ij}\boldsymbol{\beta}_k)\mu_{ik} + \phi_{ik}^{-1}}\right) \\ & + \phi_{ik}^{-1} \ln\left(\frac{\phi_{ik}^{-1}}{R_{ij}\exp(\mathbf{x}'_{ij}\boldsymbol{\beta}_k)\mu_{ik} + \phi_{ik}^{-1}}\right) + \ln\Gamma(r_{ijk} + \phi_{ik}^{-1}) - \ln\Gamma(r_{ijk} + 1) - \ln\Gamma(\phi_{ik}^{-1}) \end{aligned} \tag{10}$$

Then, marginalizing over the mixture yields the ZINB log likelihood:

$$\ln p(r_{ijk} \mid \cdot) = \ln(\pi_{ik} + \exp(l(\cdot))) \text{ if } r_{ijk} = 0 \tag{11}$$

$$\ln p(r_{ijk} \mid \cdot) = l(\cdot) \text{ otherwise} \tag{12}$$

To estimate the model parameters, we maximized the ZINB log likelihood. The parameters must satisfy the constraints $\mu_{ik} > 0$, $\phi_{ik} > 0$, $0 \leq \pi_{ik} \leq 1$. To make the problem easier, we re-parameterized in terms of $\ln \mu_{ik}$, $\ln \phi_{ik}$, logit(π_{ik}) and performed unconstrained optimization.

The ZINB log likelihood is nonconvex; therefore, we used a two stage optimization procedure. In the first stage, we optimized the NB log likelihood with respect to $\ln \mu_{ik}$, $\ln \phi_{ik}$, initializing from zero. In the second stage, we used the NB solution and logit(π_{ik}) = −8 (corresponding to a suitably small value of π_{ik}) as the initialization and optimized the ZINB log likelihood. In both stages, we used batch gradient descent for 30,000 iterations with fixed learning rate 10^{-3}, accelerated by RMSProp [36]. We implemented the method in Tensorflow [37].

We defined the size factor of each cell as the total number of molecules detected (before excluding genes in QC). To correct for technical confounders, we included C1 chip as an observed confounder, recoded as binary indicator variables and centered. This approach is sufficient to also correct for batch, because in our experimental design, batch is a linear combination of C1 chip. Intuitively, if there were a batch effect independent of C1 chip, then we could add the batch effect to each chip effect and set the batch effect to 0.

To assess the goodness of fit of the method, we used a diagnostic test based on the following simple fact: if the data $x_1, \ldots, x_n$ are continuous random variables generated from a continuous CDF F, then $F(x_i) \sim \text{Uniform}(0, 1)$. Then, to test for goodness of fit of an estimated $\hat{F}$ to the data $x_1, \ldots, x_n$, we apply the Kolmogorov-Smirnov (KS) test to test whether the values $\hat{F}(x_1), \ldots, \hat{F}(x_n)$ are uniformly distributed. (This test is slightly conservative because it uses the data to estimate $\hat{F}$).

Here, we have to modify this simple procedure to account for the fact that our data are discrete counts, so F is not continuous. To address this issue, we used randomized quantiles [38]: we sample one random value per observation $u_i \mid x_i \sim \text{Uniform}(\hat{F}(x_i - 1), \hat{F}(x_i))$. These have the property that if $x_i \sim F$ then $u_i \sim \text{Uniform}(0, 1)$.

In our model, each observed UMI count x_{ijk} comes from a different distribution F_{ijk}, because it depends on the library size which is cell-specific. We therefore draw $u_{ijk} \mid x_{ijk} \sim \text{Uniform}(\hat{F}_{ijk}(x_{ijk} - 1), \hat{F}_{ijk}(x_{ijk}))$. Then, for each individual i and gene k, we apply the KS test to whether the randomized quantiles u_{ijk} across cells j are uniformly distributed.

Quantitative trait locus mapping

We imputed dosages for 120 Yoruba individuals from the HapMap project (Phase 3, hg19) as previously described [39]. We restricted our analysis to 8,472,478 variants with minor allele frequency at least 0.05.

For each single cell expression phenotype tested, we standardized and quantile-normalized the phenotype matrix to a standard normal as previously described [40]. We called QTLs within 100 kilobases of the transcription start site of each gene and controlled the gene-level false discovery rate using *QTLtools* [41]. We included principal components (PCs) of the normalized expression matrix as covariates for QTL mapping, and selected the number of PCs for each phenotype by greedily searching for the number of PCs which maximized the number of QTLs discovered on even chromosomes only at FDR 10%. We did not include genotype PCs as covariates. We additionally recalled eQTLs in the matched bulk RNA-seq data [20] using the re-processed dosage matrix.

We performed replication testing by taking each SNP-gene pair from the discovery cohort, and testing that pair in the replication cohort. We defined a hit as replicating if it passed the Benjamini–Hochberg procedure at level 10% (restricted to the set of SNP-gene pairs tested) and had the same effect size direction.

Power analysis

For individual i and gene k, we assume the generative model:

$$y_{ik} = x_i b + e_{ik} \tag{13}$$

$$\tilde{y}_{ik} = y_{ik} + \tilde{e}_{ik} \tag{14}$$

where $\tilde{y}_{ik}$ is the observed phenotype, y_{ik} is the true phenotype, x_i is the genotype at the SNP of interest, $\tilde{e}_{ik} \sim \mathcal{N}(0, \sigma_m^2)$, and $e_{ik} \sim \mathcal{N}(0, \sigma_r^2)$.

To perform QTL mapping, we fit a working model which ignores measurement error:

$$\tilde{y}_{ik} = x_i \beta + \epsilon_{ik} \tag{15}$$

where $\epsilon_{ik} \sim N(0, \sigma^2)$. From this model, we estimate $\hat{\beta}$. Assuming $\mathbb{V}[x] = 1$, we have $\sigma^2 = \sigma_r^2 + \sigma_m^2$ and:

$$\hat{\beta} \sim \mathcal{N}\left(b, \frac{\sigma_r^2 + \sigma_m^2}{n}\right) \tag{16}$$

where n is the number of individuals. Under the working model, the power function is:

$$\text{Pow}(\cdot) = \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{b}{\text{SE}(\hat{\beta})}\right) \tag{17}$$

where α denotes the significance level, SE($\cdot$) denotes standard error, and $\Phi(\cdot)$ denotes the standard Gaussian CDF. Under the assumed generative model, the power function equals:

$$\text{Pow}(\lambda, n, \delta, \alpha) = \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \lambda\sqrt{\frac{n}{1+\delta}}\right) \tag{18}$$

where $\lambda = b/\sigma_r$, and $\delta = \sigma_m^2/\sigma_r^2$. Parameterized in terms of δ, the power function implies useful reference points; for example, $\delta = 1$ is equivalent to cutting the sample size in half.

To determine the effect size b, we estimate the distribution of true effect sizes b given observed effect sizes $\hat{\beta}_j$ and associated standard errors $\hat{s}_j$. We assume the hierarchical model:

$$\hat{\beta}_j \mid b_j, \hat{s}_j \sim \mathcal{N}(b_j, \hat{s}_j^2) \tag{19}$$

$$b_j \mid \hat{s}_j \sim g(\cdot) \tag{20}$$

where g is a unimodal mixture of Gaussians. We estimate g using adaptive shrinkage (*ash*) [21]. We took b to be the 99th percentile of the fitted distribution.

Although we assumed a single measurement error variance σ_m^2, we actually have measurement errors for each individual and gene σ_{mik}^2. To estimate σ_{mik}^2, we used non-parametric bootstrapping. For each individual and gene, we resampled the counts (matched with the library size and technical confounders) with replacement, and refit the ZINB model. To reduce computational burden, we restricted our analysis to 200 randomly chosen genes, warm-started the optimization from the optimal parameters for the original data, and ran gradient descent for 30,000 iterations.

To estimate the typical noise ratio δ, we estimate a measurement error variance per gene σ_{mk}^2 and a residual variance per gene σ_{rk}^2. We take $\hat{\sigma}_{mk}^2 = \text{median}(\sigma_{mik}^2)$. To estimate σ_{rk}^2, we solve a deconvolution problem [42]:

$$\tilde{y}_{ik} \mid y_{ik}, \hat{\sigma}_{mik}^2 \sim \mathcal{N}(y_{ik}, \hat{\sigma}_{mik}^2) \tag{21}$$

$$y_{ik} \mid \hat{\sigma}_{mik}^2 \sim g(\cdot) \tag{22}$$

where g is a unimodal mixture of uniforms, estimated using *ash*. To fit the model, we centered the $\tilde{y}_{ik}$ for each gene k, concatenated them across genes, and assumed a common prior.

Then, the required estimates are:

$$\hat{\sigma}_{rk}^2 = \hat{\mathbb{V}}[\mathbb{E}[y_{ik} \mid \cdot]] \tag{23}$$

$$\hat{\delta} = \text{median}\left(\frac{\hat{\sigma}_{mk}^2}{\hat{\sigma}_{rk}^2}\right) \tag{24}$$

$$\lambda = \frac{b}{\text{median}(\hat{\sigma}_{rk}^2)} \tag{25}$$

where $\hat{\mathbb{V}}$ denotes sample variance.

Supporting information

S1 Text. Derivation of ZINB model.
(PDF)

S1 Fig. Descriptive statistics of the experiment. Number of cells per individual, and number of molecules per cell after applying quality control filters.
(PDF)

S2 Fig. Estimated ZINB parameters and latent mean and variance in idealized simulation. Estimates of $\ln(\mu)$ and latent mean are displayed for $\text{logit}(\pi) < 0$. Estimates of $\ln(\phi)$ and latent variance are displayed for $\ln(\mu) > -10$, $\text{logit}(\pi) < 0$. Estimates of $\text{logit}(\pi)$ are displayed over the entire range of parameter values. In each trial, simulated molecule counts for 95 cells are

drawn from the model assuming 114,026 molecules per cell, matching the median number of cells, and molecules per cell in the observed data.
(PDF)

S3 Fig. Histogram of diagnostic test *p*-values for goodness of fit on simulated data. For each simulated data set, we use Kolmogorov-Smirnov test to test for departure of randomized quantiles of the data (based on the fitted ZINB distribution) from the uniform distribution.
(PDF)

S4 Fig. Histogram of diagnostic test *p*-values for goodness of fit on real data. For the set of observed UMI counts for each individual, for each gene, we use Kolmogorov-Smirnov test to test for departure of randomized quantiles of the data (based on the fitted ZINB distribution) from the uniform distribution.
(PDF)

S5 Fig. Quantile-quantile plots for QTL discovery. QQ plots are shown for dispersion, mean, variance, coefficient of variation (CV), and Fano factor.
(PDF)

S6 Fig. Estimated distribution of QTL effect sizes. We fit a unimodal mixture of Gaussians to the distribution of observed eQTL (QTL) effect sizes (in terms of log fold change) using Empirical Bayes.
(PDF)

S7 Fig. Distribution of estimated noise ratios. Noise ratios (ratio of measurement error variance to phenotypic variance) are estimated for 200 randomly chosen genes using a two-step empirical Bayes procedure.
(PDF)

Acknowledgments

We thank Joyce Hsiao for helpful discussions. Computing resources were provided by the University of Chicago Research Computing Center.

Author Contributions

Conceptualization: Po-Yuan Tung, Yoav Gilad.

Data curation: John D. Blischak.

Formal analysis: Abhishek K. Sarkar.

Funding acquisition: Yoav Gilad.

Investigation: Po-Yuan Tung, Jonathan E. Burnett.

Methodology: Abhishek K. Sarkar, Yang I. Li, Matthew Stephens.

Software: John D. Blischak.

Supervision: Matthew Stephens, Yoav Gilad.

Writing – original draft: Abhishek K. Sarkar.

Writing – review & editing: Abhishek K. Sarkar, Po-Yuan Tung, John D. Blischak, Yang I. Li, Matthew Stephens, Yoav Gilad.

References

1. Kitano H. Biological robustness. Nature Reviews Genetics. 2004; 5:826. https://doi.org/10.1038/nrg1471 PMID: 15520792
2. Masel J, Siegal ML. Robustness: mechanisms and consequences. 2009; 25(9):395–403. https://doi.org/10.1016/j.tig.2009.07.005
3. Kirschner M, Gerhart J. Evolvability. Proceedings of the National Academy of Sciences. 1998; 95 (15):8420–8427. https://doi.org/10.1073/pnas.95.15.8420
4. Lenski RE, Barrick JE, Ofria C. Balancing Robustness and Evolvability. PLOS Biology. 2006; 4(12):1–3. https://doi.org/10.1371/journal.pbio.0040428
5. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. 2013; 14(12):827–839. https://doi.org/10.1038/nrg3564
6. de Visser J Arjan G M, Joachim H, P WG, Ancel ML, Homayoun BC, L BJ, et al. PERSPECTIVE: EVOLUTION AND DETECTION OF GENETIC ROBUSTNESS. Evolution; 57(9):1959–1972. https://doi.org/10.1554/02-750R PMID: 14575319
7. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. Science. 2014; 343(6167):193–196. https://doi.org/10.1126/science.1245316 PMID: 24408435
8. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing; 24(3):496–510. https://doi.org/10.1101/gr.161034.113
9. Raser JM, O'Shea EK. Control of Stochasticity in Eukaryotic Gene Expression. Science. 2004; 304 (5678):1811–1814. https://doi.org/10.1126/science.1098641 PMID: 15166317
10. Farkash-Amar S, Zimmer A, Eden E, Cohen A, Geva-Zatorsky N, Cohen L, et al. Noise Genetics: Inferring Protein Function by Correlating Phenotype with Protein Levels and Localization in Individual Human Cells. PLOS Genetics. 2014; 10(3):1–10. https://doi.org/10.1371/journal.pgen.1004176
11. Ansel J, Bottin H, Rodriguez-Beltran C, Damon C, Nagarajan M, Fehrmann S, et al. Cell-to-Cell Stochastic Variation in Gene Expression Is a Complex Genetic Trait; 4(4):1–10. https://doi.org/10.1371/journal.pgen.1000049
12. Liu J, Martin-Yken H, Bigey F, Dequin S, François JM, Capp JP. Natural Yeast Promoter Variants Reveal Epistasis in the Generation of Transcriptional-Mediated Noise and Its Potential Benefit in Stressful Conditions. 2015; 7(4):969–984. https://doi.org/10.1093/gbe/evv047
13. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 2013; 11:163. https://doi.org/10.1038/nmeth.2772 PMID: 24363023
14. Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies; 7:39921. https://doi.org/10.1038/srep39921
15. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments; p. kxx053. https://doi.org/10.1093/biostatistics/kxx053
16. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data; 14(1):R7. https://doi.org/10.1186/gb-2013-14-1-r7
17. Wang J, Huang M, Torre E, Dueck H, Shaffer S, Murray J, et al. Gene expression distribution deconvolution in single-cell RNA sequencing. https://doi.org/10.1073/pnas.1721085115
18. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics; 11:637. https://doi.org/10.1038/nmeth.2930
19. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments; 31:748. https://doi.org/10.1038/nbt.2642
20. Banovich NE, Li YI, Raj A, Ward MC, Greenside P, Calderon D, et al. Impact of regulatory variation across human iPSCs and differentiated cells; 28(1):122–131. https://doi.org/10.1101/gr.224436.117
21. Stephens M. False discovery rates: a new deal; 18(2):275–294. https://doi.org/10.1093/biostatistics/kxw041
22. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. 2010; 467(7312):167–173. https://doi.org/10.1038/nature09326
23. Waddington CH. Canalization of Development and Genetic Assimilation of Acquired Characters. 1959; 183:1654. https://doi.org/10.1038/1831654a0
24. Waddington CH. Evolutionary Systems–Animal and Human. 1959; 183:1634. https://doi.org/10.1038/1831634a0

25. Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J. Robustness of Cellular Functions. Cell. 2004; 118 (6):675–685. https://doi.org/10.1016/j.cell.2004.09.008. PMID: 15369668

26. Gibson G. Decanalization and the origin of complex disease. 2009; 10:134. https://doi.org/10.1038/nrg2502

27. Ogbunugafor CB, Pease JB, Turner PE. On the possible role of robustness in the evolution of infectious diseases. 2010; 20(2):026108. https://doi.org/10.1063/1.3455189

28. Garfield DA, Runcie DE, Babbitt CC, Haygood R, Nielsen WJ, Wray GA. The Impact of Gene Expression Variation on the Robustness and Evolvability of a Developmental Gene Regulatory Network. PLOS Biology. 2013; 11(10):1–16. https://doi.org/10.1371/journal.pbio.1001696

29. Chen G, Gulbranson DR, Hou Z, Bolin JM, Ruotti V, Probasco MD, et al. Chemically defined conditions for human iPS cell derivation and culture; 8(5):424–429. https://doi.org/10.1038/nmeth.1593

30. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. 2013; 41(10):e108–e108. https://doi.org/10.1093/nar/gkt214

31. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research. 2017; 27(3):491–499. https://doi.org/10.1101/gr.209601.116 PMID: 28100584

32. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677

33. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data; 91(5):839–848. https://doi.org/10.1016/j.ajhg.2012.09.004.

34. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation; 40(10):4288–4297. https://doi.org/10.1093/nar/gks042

35. Tipping ME, Bishop CM. Probabilistic Principal Component Analysis; 61(3):611–622.

36. Tieleman T, Hinton G. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude; 2012. COURSERA: Neural Networks for Machine Learning.

37. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Available from: https://www.tensorflow.org/.

38. Dunn PK, Smyth GK. Randomized Quantile Residuals. Journal of Computational and Graphical Statistics. 1996; 5(3):236–244. https://doi.org/10.1080/10618600.1996.10474708

39. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of Genetic Variants That Affect Histone Modifications in Human Cells; 342(6159):747–749. https://doi.org/10.1126/science.1242429

40. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, et al. DNase I sensitivity QTLs are a major determinant of human expression variation; 482:390. https://doi.org/10.1038/nature10808

41. Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool set for molecular QTL discovery and analysis; 8:15452. https://doi.org/10.1038/ncomms15452

42. Cordy CB, Thomas DR. Deconvolution of a Distribution Function. Journal of the American Statistical Association. 1997; 92(440):1459–1465. https://doi.org/10.1080/01621459.1997.10473667