Pseudoreplication bias in single-cell studies; a practical solution

Authors: Kip D. Zimmerman [1*], Mark A. Espeland [1], Carl D. Langefeld [1*]

Affiliations:[1]Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, NC, USA.

*Corresponding authors. Email: clangefe@wakehealth.edu or kdzimmer@wakehealth.edu

Cells from the same individual share a common genetic and environmental background and are not independent, therefore they are subsamples or pseudoreplicates. Empirically, we show this dependence across a range of cell types. Thus, single-cell data have a hierarchical structure that current single-cell methods do not address and subsequently the application of such tools leads to biased inference and reduced robustness and reproducibility. When properly simulating the hierarchical structure of single-cell data, commonly applied single-cell differential expression analysis tools exhibit highly inflated type I error rates, particularly when applied together with a batch effect correction for individual as a means of accounting for within sample correlation. As single-cell experiments increase in size and frequency, we propose applying generalized linear mixed models that include random effects for differences among persons to properly account for the correlation structure that exists among measures from cells within an individual.

The rapid evolution of single-cell technologies will enable novel interrogation of fundamental questions in biology, dramatically accelerating discoveries across many biological disciplines. Thus, researchers are developing methods that leverage or account for the unique properties of single-cell RNA sequencing (scRNA-seq) data, particularly its increased sparseness and heterogeneity compared to its bulk sequencing counterpart[1–3]. An important characteristic of single-cell experiments is that they result in many cells from the same individual, and therefore the same genetic and environmental background. Here we empirically document the correlation among measures from cells within an individual and demonstrate how differential expression analysis of scRNA-seq data without considering this correlation, the current common practice, violates fundamental assumptions and leads to false conclusions. Proper identification of the experimental unit (i.e., the smallest observation for which independence can be assumed) for the hypothesis is critical for proper inference. Observations nested within an experimental unit are referred to as subsamples, technical replicates, or pseudoreplicates. Pseudoreplication, or subsampling, is formally defined as "the use of inferential statistics where replicates are not statistically independent"[4]. There are two types of pseudoreplication commonly occurring in single-cell experiments: simple and sacrificial. Simple pseudoreplication occurs when "samples from a single experimental unit are treated as replicates representing multiple experimental units" [4,5]. Sacrificial pseudoreplication occurs when "the samples taken from each experimental unit are treated as independent replicates"[4,5]. Pseudoreplication has been addressed repeatedly in the fields of ecology, agriculture, psychology, and neuroscience[4–8] and has been acknowledged as one of the most common statistical mistakes in scientific literature[9]. New technologies are particular prone to this error. Thus, it is not surprising that pseudoreplication is ubiquitous in the single-cell literature. Properly identifying the right experimental unit in single-cell studies will greatly increase both robustness and reproducibility, thereby leveraging the very features that make single-cell methods powerful.

Measures from cells from the same individual should be more (positively) correlated with each other than cells from unrelated individuals. Empirically, this appears true across a range of cell types (Fig. 1). Thus, single-cell data have a hierarchical structure in which the single-cells may not be mutually independent and have a study-specific correlation (e.g., exchangeable

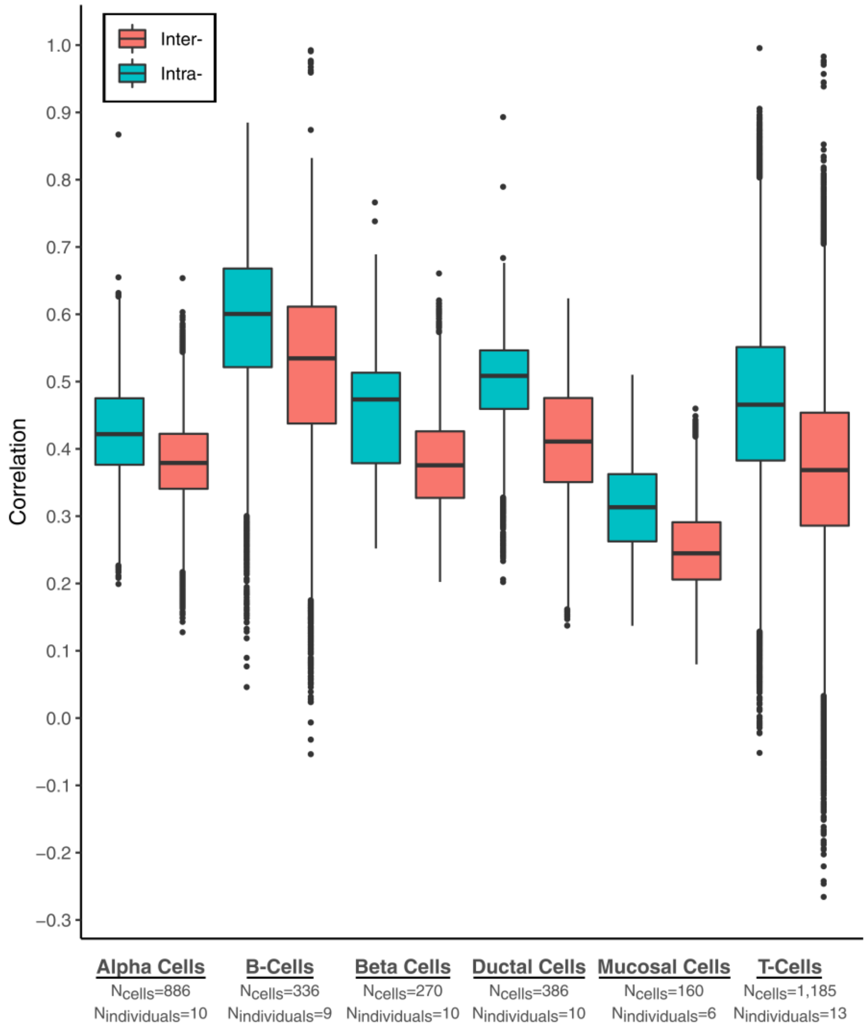

Fig. 1 | Intra-individual correlation. Box plot of the intra- and inter-individual Spearman's correlations for gene expression values across six different cell types. Cell types, along with their respective numbers of cells (Ncells) and individuals (Nindividuals), are labeled on the x axis. Mean correlation among a donor's own cells (intra-individual) is always greater than the mean correlation across individuals (inter-individual). Some cell types may be more correlated than others. We note that the population of B-cells is largely unbalanced. Over 80% of the cells are contributed by only three of ten individuals, which may partially explain the lack of difference between inter- and intra-individual correlation. We also note that cell types were designated by previous authors.

correlation within an individual). We note, that, within a cell type, cells appear to also exhibit some correlation across individuals (Fig. 1). We hypothesize this is due to zero-inflation and the stability in functional gene expression that is needed for a cell to classify as a specific cell type (e.g., T-cells need to have some consistent signals of gene expression related to their function as T-cells). As the denominator of most statistical tests (e.g., Wald test) is a function of the variance, not accounting for the positive correlation among sampling units underestimates the true standard error and leads to false positives[10,11]. In addition, treating each cell as independent inflates the test degrees of freedom, making it easier to falsely reject the null hypothesis (type 1 error). Too many false positives can mask true associations, especially when multiple comparison procedures such as false discovery rate are applied. In combination, this will adversely affect downstream analyses (pathway analysis), robustness, and reproducibility – increasing the cost of science.

Single-cell studies designed to identify differentially expressed genes rarely note or address the correlation among cells from the same individual or experimental unit. Excellent reviews of the field and methodological work have largely focused on challenges presented by properly classifying cell types, multimodality, dropout, and higher noise derived from biological and technical factors. However, they fail to highlight the effect of pseudoreplication and, furthermore, publications evaluating the performance single-cell specific tools all compute the simulations as if cells were independent[12–18]. The result is reduced reproducibility with real data, leading to the conclusion that tools built specifically to handle single-cell data do not appear to perform better than tools created for bulk data analysis[19–21]. We completed a simulation study that reproduces both the inter- and intra-individual variance structures estimated from real data and documented the effect of intra-individual correlation on

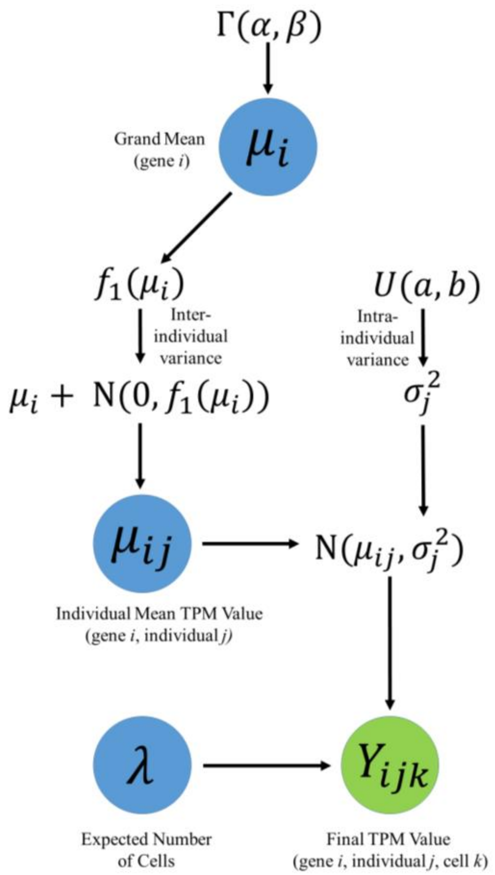
$\Gamma(\alpha,\beta)$
Grand Mean (gene *i*)
μ_i
$f_1(\mu_i)$
Inter-individual variance
$U(a,b)$
Intra-individual variance
$\mu_i + N(0, f_1(\mu_i))$
σ_j^2
μ_{ij}
$N(\mu_{ij}, \sigma_j^2)$
Individual Mean TPM Value (gene *i*, individual *j*)
λ
Y_{ijk}
Expected Number of Cells
Final TPM Value (gene *i*, individual *j*, cell *k*)

Fig. 2 | Simulation workflow. A gamma [$\boldsymbol{\Gamma(\alpha,\beta)}$] distribution was fit to the global mean transcripts per million read counts (TPM) of each gene to obtain a grand mean, $\boldsymbol{\mu_i}$. The variance of the individual-specific means (inter-individual variance) was modeled as a quadratic function of the grand mean, $\boldsymbol{f_1(\mu_i)}$ and the within-sample variance (intra-individual variance) was simulated using a uniform $\boldsymbol{U(a,b)}$ distribution. Using a normal $\boldsymbol{N(\mu,\sigma^2)}$ distribution with an expected value of zero and a variance computed by the first quadratic relationship, $\boldsymbol{f_1(\mu_i)}$, a difference in means was drawn for each individual in the simulation. This difference was summed with the grand mean to obtain an individual mean, $\boldsymbol{\mu_{ij}}$. A Poisson ($\boldsymbol{\lambda}$) distribution with a λ equal to the expected number of cells desired for each individual was then used to obtain the count of cells per individual. For each cell assigned to an individual, a TPM count, $\boldsymbol{Y_{ijk}}$, was drawn from a normal distribution with an expected value equal to the individual's assigned mean TPM value, $\boldsymbol{\mu_{ij}}$, and a variance, $\boldsymbol{\sigma_j^2}$, drawn from a uniform distribution.

the type 1 error rates of the most frequently used single-cell analysis tools (Fig. 2, fig. S1). Our simulation compared methods that do and do not account for the repeated observations within an experimental unit (see Methods). We varied the number of individuals and cells within an individual. All methods considered use asymptotic approximations and admit covariates.

We observed that the generalized linear mixed model (GLMM), either employing a tweedie distribution or a two-part hurdle model with a random effect (RE) for individual, outperformed other methods across a variety of conditions (Table 1, tables S1-S4)[22–28].

Table 1 | Type I error rates of some of the currently applied tools in single-cell analysis. Type I error rates of nine different methods under sixteen different conditions and a significance threshold of p<0.05. 250,000 iterations were computed to obtain an error rate for each method. The conservative type I error rates computed with mixed models at the lower numbers of individuals per group are a consequence of underpowered study designs. Type I error rates are well controlled for with mixed models, while type I error rates inflate with other methods as additional independent samples or more cells are added.

		Two-part Hurdle			Tweedie			Nested	Modified	
$N_{individuals}$	N_{cells}	Default	RE	Corrected	GLM	GLMM	GEE1	FE	*t*	Tobit
2	10	0.041	0.013	0.230	0.074	0.064	0.336	0.080	0.077	0.858
	25	0.064	0.034	0.551	0.089	0.071	0.310	0.114	0.105	0.865
	50	0.094	0.058	0.568	0.114	0.084	0.306	0.147	0.136	0.872
	100	0.138	0.081	0.573	0.154	0.101	0.298	0.195	0.178	0.883
5	10	0.040	0.029	0.247	0.063	0.048	0.133	0.083	0.077	0.856
	25	0.067	0.044	0.465	0.082	0.053	0.124	0.113	0.104	0.865
	50	0.101	0.056	0.628	0.112	0.059	0.120	0.150	0.134	0.873
	100	0.144	0.066	0.736	0.152	0.067	0.119	0.196	0.178	0.884
10	10	0.044	0.035	0.241	0.060	0.043	0.090	0.083	0.076	0.856
	25	0.073	0.046	0.453	0.083	0.047	0.083	0.114	0.101	0.863
	50	0.106	0.051	0.589	0.111	0.051	0.081	0.150	0.134	0.873
	100	0.151	0.060	0.718	0.151	0.055	0.079	0.195	0.175	0.883
25	10	0.049	0.043	0.238	0.057	0.042	0.063	0.084	0.076	0.854
	25	0.079	0.048	0.444	0.081	0.044	0.062	0.113	0.102	0.864
	50	0.113	0.052	0.582	0.111	0.046	0.061	0.150	0.133	0.873
	100	0.157	0.056	0.696	0.152	0.049	0.060	0.196	0.177	0.883

***Default denotes MAST was implemented without random-effects, RE denotes random-effects, Corrected denotes data was batch-corrected for individual prior to analysis without using individual as a random-effect, GLM denotes generalized linear model, GLMM denotes generalized linear mixed-effects model, and FE denotes fixed-effects.**

****Two-part Hurdle model as implemented in MAST, Tweedie distribution as implemented in ‘glmmTMB’, GEE1 as implemented in ‘geepack’, Modified t as implemented in ROTS, and Tobit as implemented in Monocle.**

Specifically, among the methods that explicitly model the correlation structure, GLMM consistently had more appropriate type 1 error rate control than both generalized estimating equations (GEE1) models and nested fixed-effects models, where the latter two perform poorly for any number of subsampling until the number of independent experimental units approached 25[29,30]. When the number of experimental units was small, the GEE1 sandwich estimator of the variance provided standard errors that were too small and therefore inflated the type 1 error rate. Similarly, for nested fixed-effects models, the standard errors were also underestimated with standard estimation techniques (i.e., REML). The models that explicitly model the correlation structure all outperformed the methods that do not account for the lack of independence among experimental units (Table 1, tables S1-S4). All methods that treat observations as independent perform increasingly worse as the number of correlated cells increases. We note that DESeq2 regularly failed to compute in scenarios where the numbers of cells and samples were large because the geometric mean normalization method implemented requiring at least one transcript to consist completely of all non-zero values (Tables S1-S4). A particularly noteworthy approach that has been suggested to account for the within-individual correlation is applying a batch effect correction method, for which the batches are individuals. This approach had markedly increased type 1 error rates (Table 1, tables S1-S4). This is primarily because regressing out the person-specific effect as a batch effect and subsequently analyzing each cell as an independent observation will underestimate the overall variance by removing inter-individual differences while maintaining an inappropriately large number of degrees of freedom when treating cells as if they are independent.

One of the most heavily cited single-cell analysis tools, Model-based Analysis of Single-cell Transcriptomics (MAST), is a two-part hurdle model built to handle sparse and bi-modally distributed single-cell data[22]. Although, to our knowledge, there are no publications that employ MAST to account for pseudoreplication as discussed here, Finak et al. note that MAST "can easily be extended to accommodate random effects"[22]. Here, we emphasize this tool as an already well-established tool in the field and demonstrate that MAST performs exceptionally well when adjusting for individual as a random effect (i.e., MAST with RE), but no different than other tools when not doing so. This specific evaluation of MAST's performance with and without a random effect for individual serves as a perfect example of why accounting for pseudoreplication is so important. While we do recommend computing differential expression analysis using MAST with RE, alternative methods include the tweedie GLMM or permutation testing. In order not to violate the exchangeability assumption, permutation methods must randomize at the independent experimental unit (e.g., individual) and properly account for covariates (i.e., conditional permutation). The tweedie GLMM method could be implemented using the 'glmmTMB' R-package[23], but neither of these alternative approaches explicitly incorporate some of the single-cell specific concepts implemented in MAST (e.g., cellular detection rate). As detailed in their original manuscript, MAST models a $\log_2$(TPM + 1) gene expression matrix as a two-part generalized regression model[22]. Using their same notation, the addition of random effects for differences among persons is as follows:

$$\boldsymbol{logit}\left(\mathbf{Pr}\left(\boldsymbol{Z_{ig}} = \mathbf{1} | \boldsymbol{X_i}\right)\right) = \boldsymbol{X_i \beta_g}$$

$$\mathbf{Pr}\left(\boldsymbol{Y_{ig}} = \boldsymbol{y} \middle| \boldsymbol{Z_{ig}} = \mathbf{1}\right) = \boldsymbol{N}(\boldsymbol{X_i \beta_g} + \boldsymbol{W_i \gamma_j}, \boldsymbol{\sigma_g^2})$$

where Y_{ig} is the expression level for gene g and cell i, Z_{ig} is an indicator for whether gene g is expressed in cell i, X_i contains the predictor variables for each cell i, and W_i is the design matrix for the random effects of each cell i belonging to each individual j (i.e., the random complement to the fixed X_i). β_g represents the vector of fixed-effects regression coefficients and γ_j represents the vector of random effects (i.e., the random complement to the fixed β_g). γ_j is distributed normally with a mean of zero and variance $\sigma_{\gamma_i}^2$. To obtain a single result for each gene, likelihood ratio or Wald test results from each of the two components are summed and the corresponding degrees of freedom for each component are added[22]. These tests have asymptotic χ^2 null distributions and they can be summed and remain asymptotically χ^2 because Z_g and Y_g are defined conditionally independent for each gene[22].

We computed an extensive simulation-based power analysis to provide researchers estimates across a wide range of experimental conditions. This was computed using a two-part hurdle model with random effects for individuals as implemented in MAST. Increasing the number of independent experimental units (e.g., individuals) in a study is the best way to increase power to detect true differences (Fig. 3, fig. S2-S31).

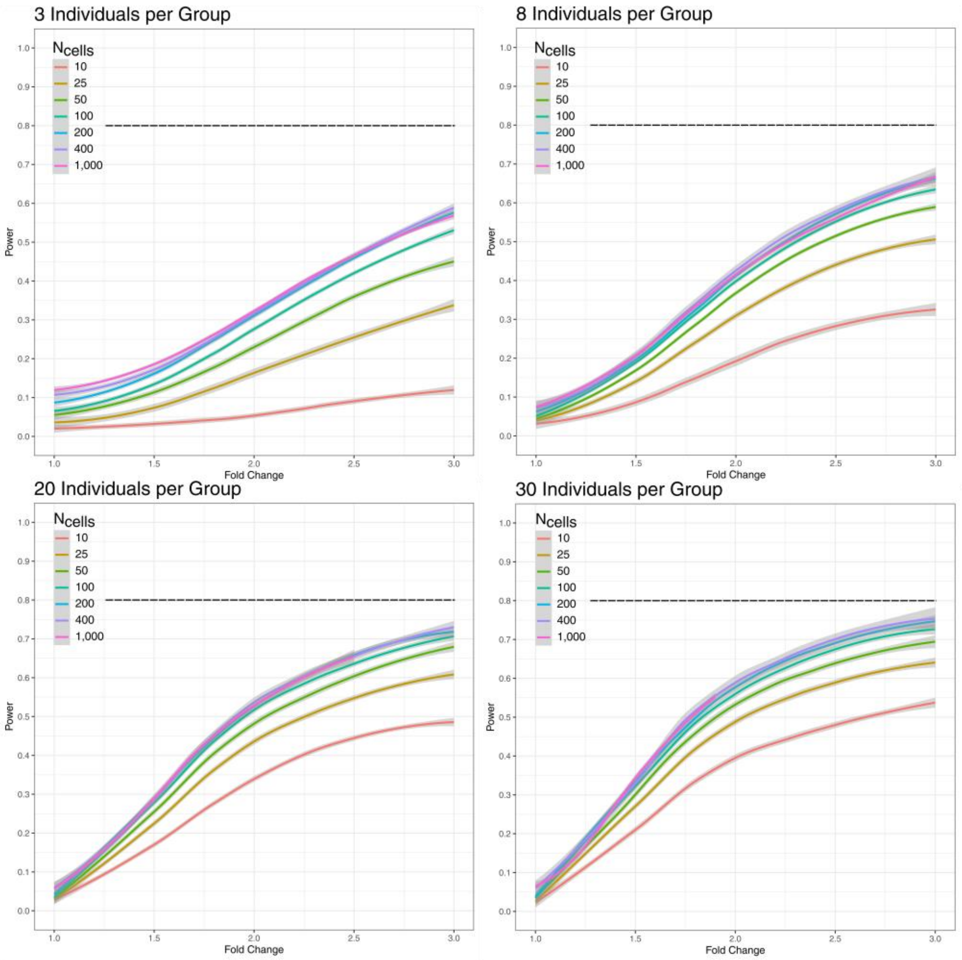

Fig. 3 | Power calculations using MAST with a random effect for individual. Power curves for twenty-eight different, but likely single-cell scenarios using MAST with a random effect for individual. Fold change is simulated by multiplying the global mean gene expression values by the fold change value for one group. Power is capped just above 0.8 because of high amounts of dropout in lowly expressed transcripts that cause either complete separation in the model or have too few observations to make inference. While increasing the expected number of cells per donor gives moderate gains in power, the greatest increase in power is achieved by increasing the number of donors.

Empirically, there are only marginal gains in power when more than twentyfive cells per individual are sampled. Increasing the number of cells per individual provides more precision in the estimate for an individual. However, it does not directly affect power for detecting differences across individuals such as treatment effects applied at the individual level (i.e., cases/control studies). We estimated negligible improvement power when increasing the expected number of cells per individual beyond 100 in a handful of situations (Fig. 3). We note that estimating power with more than 100 cells per individual was exceedingly slow and computationally expensive. Because 1000s of cells per individual is not atypical for single-cell experiments, tools that account for the correlation structure when analyzing these data need to be further developed to increase computational efficiency.

Most papers compare cells across very few individuals, sometimes even a single case and control (simple pseudoreplication); in the former case the estimate of the variance is possible but has wide bounds on parameter confidence intervals, and in the latter case the variance is not estimable. These power simulations indicate that the majority of published studies are underpowered (Fig. 3, fig. S2-S31). The majority of single-cell papers show a deep understanding of the underlying biology and conduct otherwise very informative experiments, appropriately landing in very high visibility journals. However, our type 1 error and power simulations document that many published studies are missing important true effects while reporting too many false positives generated via pseudoreplication. As single-cell technology continues to evolve and costs decrease, reviewers need to be aware of this issue to avoid proliferation of irreproducible results. We encourage the use of mixed models, such as the tweedie GLMM or the two-part hurdle model with a random effect (e.g., as implemented in MAST with RE), as ways of accounting for the repeated observations from an individual while being able to adjust for covariates at the individual level and, if appropriate, at the individual cell level. Finally, we note that although our focus here is on hypothesis testing for finding differentially expressed genes, the concept is applicable to all single-cell sequencing technologies such as proteomics, metabolomics, and epigenetics.

Materials and Methods

Literature Review

A PubMed search for the keywords "single-cell differential expression" returned 251 articles published in the last 3 years which were subsequently sorted and filtered by each of their abstracts. Many of the returned articles were associated with bulk RNA sequencing or completely irrelevant to differential expression analysis in single-cell and were therefore eliminated. Of the 251 original hits, 76 of them were deemed appropriate for further consideration. Of those, 10 of them were reviews, 36 of them were methods papers, and 30 of them were implementation papers. This method is not meant to be a perfect capture of all of the literature, but provides a clear snapshot of the current state of single-cell differential expression analyses. Each of the methods and implementation articles was thoroughly reviewed and tabled along with its number of citations, date of publication, and any other pertinent information such as number of independent samples, tools used, or number of cells captured.

Intra and Inter-correlation analyses

Pairwise comparisons between all cells of interest were made to compute intra- and inter-individual correlations. Genes were filtered if the average transcript-per-million (TPM) value was not greater than five. For intra-individual correlation, spearman's correlation was computed for all possible pairs of cells within an individual. For inter-individual correlation, spearman's correlation was computed for all possible pairs of cells from a random draw of one cell from each individual. 1,000 draws were computed. Correlations and their means were tested for differences (Fig. 1). The measures were compared in six different cell types across three different single-cell studies. These studies are publically available under the accession numbers GSE81861, GSE72056, and E-MTAB-5061. The cell type designations that were used were given by the authors of these studies.

Simulation

Means and variances were computed empirically from the transcript per million read count values previously reported in six different cell types across three different single-cell studies. Once consistent patterns were identified across cell types, alpha cells from the pancreas dataset, were used as the model data for our simulation. After removing the top percent of the most variant genes, we fit a gamma distribution to the global mean transcript per million read count values of each gene to obtain a grand mean, μ_i. The variance of the individual-specific means (inter-individual variance) was modeled as a quadratic function of the grand mean, $f_1(\mu_i)$ and the within-sample variance (intra-individual variance) was simulated using a uniform distribution, $U(a,b)$. (Fig. 2). The objective for obtaining the inter-individual variance as function of the grand mean was to simply capture the average associations between the variance and the grand mean - the uncertainty in this relationship was not of particular interest for this simulation. Using a normal distribution with an expected value of zero and a variance computed by the first quadratic relationship, $f_1(\mu_i)$, a difference in means was drawn for each individual in the simulation. This difference was summed with the grand mean to obtain an individual mean, μ_{ij}. A Poisson distribution with a λ equal to the expected number of cells desired for each individual was then used to obtain the count of cells per individual. For each cell assigned to an individual, a transcript per million read count value, Y_{ijk}, was drawn from a normal distribution with an expected value equal to the individual's assigned mean transcript per million read count value, μ_{ij}, and a variance, σ_j^2, drawn from a uniform distribution. The correlation structure between genes was not taken into account in this simulation and this simple process was replicated a selected number of times to obtain the desired number of genes. Due to their widespread use in the field, tSNE plots were made of the simulated data to assess how realistic the simulated data appeared and to assess the effects of altering intra-individual variance in these data (Fig S1).

Type I error

250,000 iterations of our simulation were computed for varying numbers of cells and varying numbers of individuals. The number of individuals per group was fixed at either 2, 5, 10, or 25. The number of cells per individual was drawn from a Poisson distribution with either a λ of 10, 25, 50, or 100. For each of the 250,000 iterations, the number of results that met our significance threshold were counted and the type I error was computed as the percentage of significant results. After primary analysis of the type I error using a tweedie mixed-effects model, type I error was computed with the following tools: MAST, MAST with random effects, MAST with a batch effect correction, DESeq2, Monocle, ROTS, Tweedie GLM, and a GEE1 with a Gaussian link and exchangeable correlation. All of these tools were selected because they could handle the transcript per million read count values being simulated – with the exception of DESeq2. DESeq2 requires integers and at least one gene without a zero value to compute its normalization, so as the number of samples and cells increased, the likelihood of if computing greatly decreased. We acknowledge DESeq2 is not appropriate for analyzing these data, but felt that where we could complete simulations, the tool must be addressed because of its frequent use in the field. MAST was implemented with and without the use of a random effect for individual and the remaining single-cell tools were implemented exactly as their vignettes instruct. GEE1 with exchangeable correlation was implemented to compare its performance to the mixed-effects model, particularly where the numbers of donors are low. Type I errors were computed using significance thresholds of 0.05, 0.01, 0.001, and 0.0001 (Table 1, tables S1-S4).

Power calculations

Using MAST with a random effect for individual we computed power curves to estimate how well this tool functions with varying numbers and ratios of cells and individuals. Computations were identical to the type I error analyses with exception of multiplying a constant, hereafter labeled fold change, with the global mean gene expression value of a gene to spike the expression values in one group. Power was computed at small increments between a fold change of 1 and 5, or until MAST with RE was unable to compute because of complete separation. For lowly expressed genes with high amounts of zero inflation, inference remained difficult, causing MAST with RE to asymptote out before reaching maximum power. This is just the nature of sparse single-cell data, and it cannot be avoided.

References and Notes:

1. Grün, D. & van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* **163**, 799–810 (2015).
2. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research* **42**, 8845–8860 (2014).
3. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257–272 (2019).

4. Hurlbert, S. H. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs* **54**, 187–211 (1984).
5. Heffner, R. A., Butler, M. J. & Reilly, C. K. Pseudoreplication Revisited. *Ecology* **77**, 2558–2562 (1996).
6. Freeberg, T. & Lucas, J. Pseudoreplication Is (Still) a Problem. *Journal of comparative psychology (Washington, D.C. : 1983)* **123**, 450–1 (2009).
7. Lazic, S. E. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neuroscience* **11**, (2010).
8. Millar, R. B. & Anderson, M. J. Remedies for pseudoreplication. *Fisheries Research* **70**, 397–407 (2004).
9. Makin, T. R. & Orban de Xivry, J.-J. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* **8**, e48175 (2019).
10. Maas, C. J. M. & Hox, J. J. Sufficient Sample Sizes for Multilevel Modeling. *Methodology* **1**, 86–92 (2005).
11. McNeish, D. Analyzing Clustered Data with OLS Regression: The Effect of a Hierarchical Data Structure. *Multiple Linear Regression Viewpoints* **40**, 11–16 (2014).
12. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun* **10**, 1–11 (2019).
13. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology* **18**, (2017).
14. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740–742 (2014).
15. Korthauer, K. D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology* **17**, (2016).

16. Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biology* **19**, (2018).

17. Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128–2135 (2016).

18. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLOS Computational Biology* **11**, e1004333 (2015).

19. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods* **15**, 255–261 (2018).

20. Dal Molin, A., Baruzzo, G. & Di Camillo, B. Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods. *Frontiers in Genetics* **8**, (2017).

21. Jaakkola, M. K., Seyednasrollah, F., Mehmood, A. & Elo, L. L. Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in Bioinformatics* bbw057 (2016) doi:10.1093/bib/bbw057.

22. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, (2015).

23. Brooks, M. E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* **9**, 378–400 (2017).

24. Højsgaard, S., Halekoh, U. & Yan, J. The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software* **15**, 1–11 (2005).

25. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

26. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 21 (2014).

27. Suomi, T., Seyednasrollah, F., Jaakkola, M. K., Faux, T. & Elo, L. L. ROTS: An R package for reproducibility-optimized statistical testing. 10.

28. Trapnell, C. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *nature biotechnology* **32**, 11 (2014).

29. Ziegler, A. & Vens, M. Generalized Estimating Equations. *Methods Inf Med* **49**, 421–425 (2010).

30. Draper, N. R. Analysis of Messy Data Volume 1: Designed Experiments, Second Edition by George A. Milliken, Dallas E. Johnson. *International Statistical Review* **77**, 321–322 (2009).

Acknowledgments

We thank T. D. Howard, L. D. Miller, and M. Olivier (All WFU School of Medicine) for critical review of this content. This work was supported by The Center for Public Health Genomics and grant U01 NS036695 (Co-PI Langefeld) from NIH.

Author contributions

C.D.L. and K.D.Z conceived the study together. K.D.Z implemented simulations and analyses with guidance from C.D.L. K.D.Z wrote the original draft and reviewed and edited it with M.A.E. and C.D.L.

Competing interests

Authors declare no competing interests.

Data and materials availability: All data are publically available under the accession numbers GSE81861, GSE72056, and E-MTAB-5061. All base code is available in the supplementary materials.

Supplementary Materials:

Materials and Methods

Figures S1-S31 (tSNE and power curves)

Tables S1-S4

R code – Correlation

R code – Simulation