Comparison and evaluation of statistical error models for scRNA-seq

Saket Choudhary[1] **and Rahul Satija**[1,2,✉]

[1]New York Genome Center, 101 6th Ave, New York, NY, 10013 USA
[2]Center for Genomics and Systems Biology, New York University, 12 Waverly Pl, New York, NY, 10003 USA
rsatija@nygenome.org

Heterogeneity in single-cell RNA-seq (scRNA-seq) data is driven by multiple sources, including biological variation in cellular state as well as technical variation introduced during experimental processing. Deconvolving these effects is a key challenge for preprocessing workflows. Recent work has demonstrated the importance and utility of count models for scRNA-seq analysis, but there is a lack of consensus on which statistical distributions and parameter settings are appropriate. Here, we analyze 58 scRNA-seq datasets that span a wide range of technologies, systems, and sequencing depths in order to evaluate the performance of different error models. We find that while a Poisson error model appears appropriate for sparse datasets, we observe clear evidence of overdispersion for genes with sufficient sequencing depth in all biological systems, necessitating the use of a negative binomial model. Moreover, we find that the degree of overdispersion varies widely across datasets, systems, and gene abundances, and argues for a data-driven approach for parameter estimation. Based on these analyses, we provide a set of recommendations for modeling variation in scRNA-seq data, particularly when using generalized linear models or likelihood-based approaches for preprocessing and downstream analysis.

Introduction

Single-cell RNA-sequencing (scRNA-seq) represents a powerful approach for the unsupervised characterization of molecular variation in heterogeneous biological systems (1, 2). However, separating biological heterogeneity across cells that corresponds to differences in cell type and state from alternative sources of variation represents a key analytical challenge in the normalization and preprocessing of single-cell RNA-seq data (3, 4). Upstream analytical workflows typically aim to achieve two separate but related tasks. First, data normalization aims to correct for differences in cellular sequencing depth, which collectively arise from fluctuations in cellular RNA content, efficiency in lysis and reverse transcription, and stochastic sampling during next-generation sequencing (5). Second, variance stabilization aims to address the confounding relationship between gene abundance and gene variance, and to ensure that both lowly and highly expressed genes can contribute to the downstream definition of cellular state. Although the use of unique molecular identifiers (UMIs), random sequences that label individual molecules, has been a promising approach to limit amplification bias (6, 7), variation due to sequencing depth still arises in such datasets and can be a major source of technical variance. These challenges are not unique to single-cell sequencing (8), but the sparsity of scRNA-seq data, coupled with substantial diversity in profiling technologies, necessitates the development and assessment of new methods.

While initial work focused on the development of cell 'size-factors' for normalization, recent methods have been focused on the development and application of statistical models for scRNA-seq analysis. In particular, two recent studies proposed to use generalized linear models (GLMs), where cellular sequencing depth was included as a covariate, as part of scRNA-seq preprocessing workflows. Our sctransform (9) approach utilizes the Pearson residuals from negative binomial regression as input to standard dimensional reduction techniques, while GLM-PCA (10) focuses on a generalized version of principal component analysis (PCA) for data with Poisson-distributed errors. More broadly, multiple techniques aim to learn a latent state that captures biologically relevant cellular heterogeneity using either matrix factorization or neural networks (11–13), alongside a defined error model that describes the variation that is not captured by the latent space.

Together, these studies demonstrate the importance and potential of statistical models to assist in the normalization, variance stabilization, and downstream analysis of scRNA-seq data. However, each of these approaches requires an explicit definition of a statistical error model for scRNA-seq, and there is little consensus on how to define or parameterize this model. While multiple groups have utilized a Poisson error model (10, 14–18), others argue that the data exhibit evidence of overdispersion, requiring the use of a negative-binomial (NB) distribution (5, 19–21). Even for methods that assume a NB distribution, different groups propose different methods to parameterize their model. For example, a recent study (22) argued that fixing the NB inverse overdispersion parameter θ to a single value is an appropriate estimate of technical overdispersion for all genes in all scRNA-seq datasets, while others (23) propose learning unique parameter values for each gene in each dataset. This lack of consensus is further exemplified by the scvi-tools (11, 24) suite, which supports nine different methods for parameterizing error models. The purpose of this error model is to describe and quantify heterogeneity that is not captured by biologically relevant differences in cell state, and highlights a specific question: How can we model the observed variation in gene expression for an scRNA-seq experiment conducted on a biologically 'homogeneous' population?

Results

Shallow sequencing masks overdispersion in scRNA-seq data

We first explored whether a Poisson distribution was capable of fully encapsulating heterogeneity in scRNA-seq data that was independent of biological variation in the cellular state (i.e., 'independent of the latent space' (25)). The rationale behind a Poisson model assumes that homogeneous cells express mRNA molecules for a given gene at a fixed underlying rate, and the variation in scRNA-seq results specifically from a stochastic sampling of mRNA molecules, for example due to inefficiencies in reverse transcription and PCR, combined with incomplete molecular sampling during DNA sequencing (5, 25). The Poisson distribution constrains the variance of a random variable to be equal to its mean, and has been utilized for modeling UMI counts in multiple previous studies (15, 16). While the Poisson distribution is well suited to capture variation driven by stochastic technical loss and sampling noise, it cannot capture other sources of biological heterogeneity between cells that are not driven by changes in cell state, for example, intrinsic variation caused by stochastic transcriptional bursts (26–28). These fluctuations would cause scRNA-seq data to deviate from Poisson statistics, exhibiting overdispersion that can be modeled using a negative binomial distribution.

We therefore asked whether scRNA-seq data exhibited evidence of overdispersion by exploring the mean-variance relationship using technical controls (endogenous RNA and spike-ins), cell line (HEK293 and NIH3T3), and heterogeneous (PBMC; mouse cortex; fibroblasts) datasets profiled using multiple technologies (Supplementary Table S1). These datasets have varying sequencing depths with median UMIs per cell spanning from approximately 375 to more than $195,000$ (Supplementary Figure S1). In each dataset, we performed a goodness-of-fit test, independently modeling the observed counts for each gene to be Poisson distributed, while accounting for differences in sequencing depth between individual cells (Supplementary Methods). For the technical control datasets (8, 14), where the input to each 'cell' represented a uniform source of RNA, observed variation was largely consistent with the Poisson model (Figure 1B). In contrast, when analyzing a human PBMC dataset profiled using Smart-seq3 (29), thousands of genes were poorly fit by a Poisson distribution (Figure 1A and B), even after accounting for cell-to-cell variation in sequencing depth (Supplementary Table S2). While we expected to observe overdispersion for a subset of genes, particularly for those whose expression varies across multiple cell types, we were surprised to see that 97.6% of genes with average expression > 1 UMI/cell failed the Poisson goodness-of-fit test. We observed a similar phenomenon when analyzing data from homogeneous HEK293 cells profiled with the 10X Chromium v2 system (HEK-r2; Figure 1A and B), with 93% of genes exhibiting average abundance of > 1 UMI/cell demonstrating evidence of overdispersion. In each of the 58 datasets we analyzed, genes exhibiting Poisson variation were overwhelmingly lowly expressed compared to genes that were overdispersed (Supplementary Figure S2). Moreover, when comparing results for cell-line datasets where we expect low levels of variation in cell state, we found that the global fraction of genes deviating from a Poisson distribution was correlated with the average sequencing depth of the dataset (Figure 1C).

Our results suggest that scRNA-seq datasets commonly exhibit biological variation that exceeds Poisson sampling, but that the statistical power to detect these fluctuations requires sufficient sequencing depth. For example, when observing molecular counts in the deeply sequenced PBMC dataset (median $8,288$ UMI/cell), highly expressed genes such as TPT1, RPS19 exhibited particularly strong deviations from Poisson variability (Figure 1D). However, we found that when artificially downsampling the same dataset to $1,000$ UMI/cell, a depth that is common to shallowly sequenced scRNA-seq datasets, deviations from a Poisson distribution were strongly reduced (Figure 1E). After downsampling, only 0.5% genes failed the Poisson goodness-of-fit test, demonstrating that reducing cellular sequencing depth can artificially create the appearance of Poisson variation. We conclude that the Poisson error model may represent an acceptable approximation for scRNA-seq datasets with shallow sequencing, but as the sensitivity of molecular profiling continues to increase, error models that allow for overdispersion are required for scRNA-seq analysis. Furthermore, we reiterate that the use of a Poisson error model does not account for the possibility of intrinsic stochastic noise in single-cell datasets, though this type of noise has been extensively described and does not correlate with changes in cell type or state.

The level of overdispersion varies substantially across datasets

We next focused on the application of negative binomial error models, and considered different strategies for parameterizing the level of overdispersion associated with each gene. Recent work (22) suggested that a negative binomial model with a fixed parameterization (for example, inverse overdispersion parameter $\theta = 100$) could be applied to all scRNA-seq datasets to achieve effective variance stabilization. To explore whether a single value of θ could be applied to diverse scRNA-seq datasets, we first independently fit θ estimates for each gene in each dataset using a GLM with negative binomial errors (NB GLM), using library size as an offset to account for variation in cellular sequencing depth. We observed substantial differences in the magnitude of the estimated θ across different datasets, though replicate datasets from the same study yielded concordant results (Figure 2A and B). Consistent with our previous results (Figure 1B), θ values for each dataset varied across different biological systems, technologies, and sequencing depths.

We next tested the ability for a single value of θ to perform effective variance stabilization across a range of datasets. We processed each of our 58 datasets using an NB GLM after fixing θ to a single value for all genes in the dataset (for example, θ=100). We found that no sin-

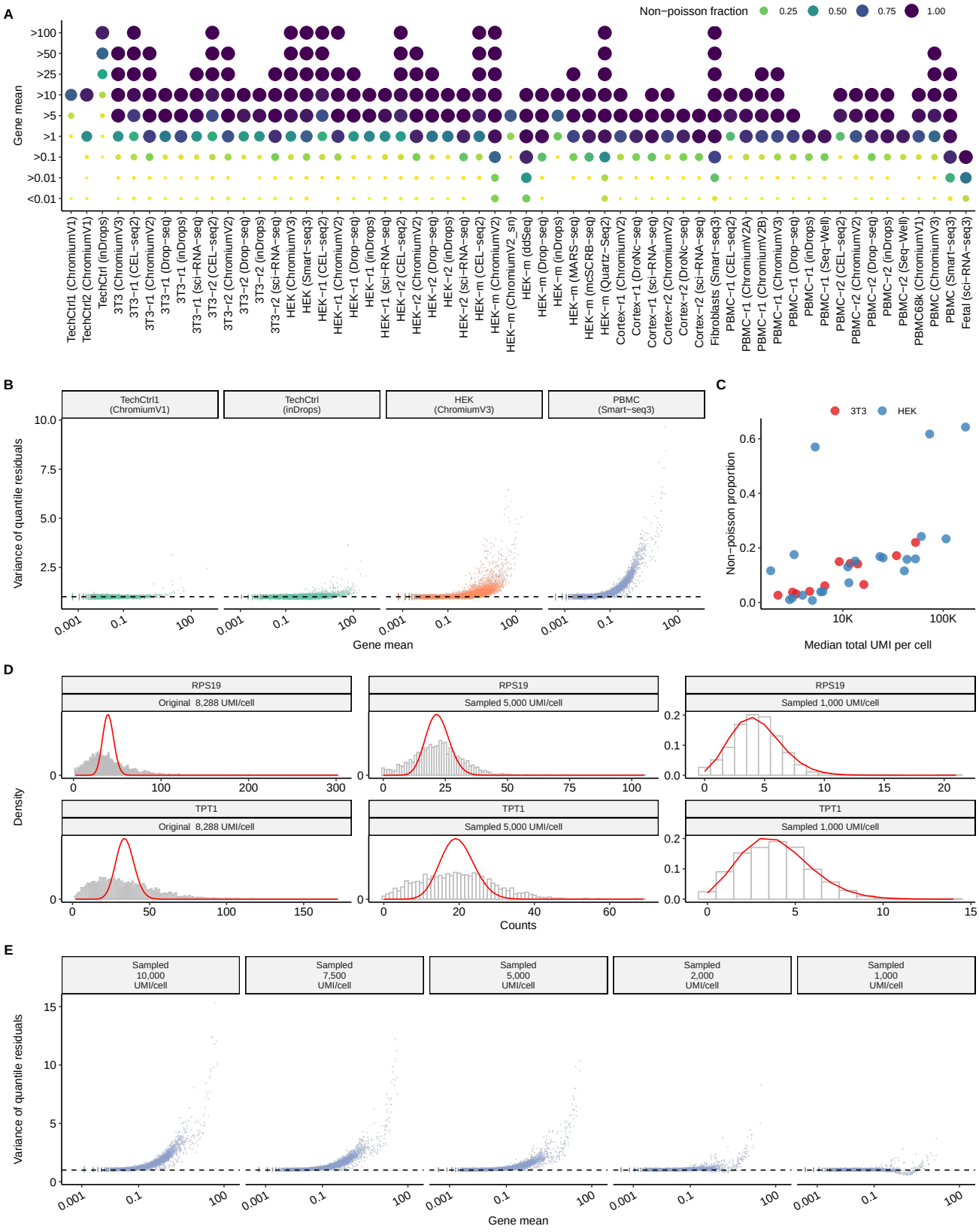

Figure 1. Shallow sequencing masks overdispersion in scRNA-seq data. A) Proportion of genes that fail a goodness-of-fit test for a Poisson GLM (Supplementary Methods), as a function of gene abundance, for 58 scRNA-seq datasets. For visual clarity, both the color and diameter of each dot correspond to the fraction of genes that exhibit overdispersion. Y-axis represents non-cumulative gene abundance bins between two consecutive labels (for example, > 1 refers to all genes with average abundance > 1 UMI and ≤ 5 UMI). Values are listed in Supplementary Table S2. **B)** Relationship between average gene abundance and quantile residual variance, after applying a Poisson GLM (Supplementary Methods). Results are shown for datasets profiling endogenous RNA ('technical controls'), a HEK293 cell line ('biological controls'), and human PBMC ('heterogeneous'). **C)** In datasets profiling cell lines, the fraction of genes that exhibit overdispersion is correlated with average sequencing depth. **D)** Distribution of molecular counts for highly expressed genes in the PBMC Smart-seq3 dataset after downsampling to two different sequencing depths. The expected density assuming a Poisson distribution is shown in red. **E)** Same as (A) but after downsampling the PBMC Smart-seq3 dataset to five different sequencing depths.

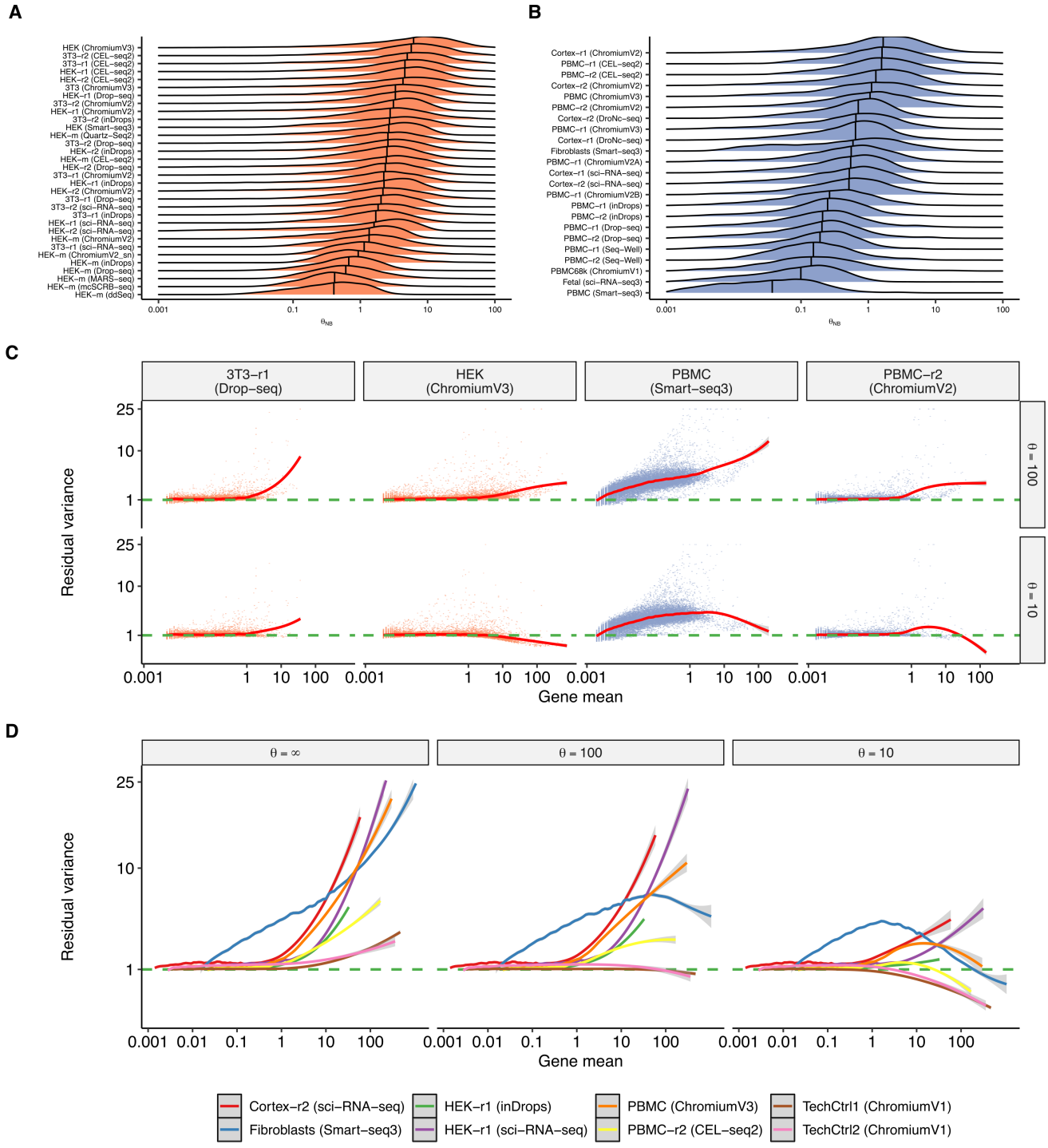

Figure 2. Overdispersion varies across datasets. A), B) Distribution of per-gene values for the estimated inverse overdispersion θ_{NB} of a NB GLM across a range of cell lines (A) and heterogeneous datasets (B). We estimated parameters only for genes where the variance of counts exceeds the mean. Vertical bar indicates the median of the distribution, which varies substantially across datasets, but is concordant across replicate experiments within the same study. **C)** Relationship between gene mean and the variance of Pearson residuals resulting from an NB GLM with $\theta = 10$ or $\theta = 100$. Each dot is a gene and the trendline (LOESS) is shown in red. **D)** Same as (C), but shown for additional datasets and for $\theta = \infty$ (Poisson). Only trendlines are shown for visual clarity.

gle value of θ could achieve effective variance stabilization across all datasets. For example, a negative binomial error model with $\theta = 100$ resulted in clear heteroskedasticity in multiple datasets (Figure 2C), as we observed a strong relationship between the mean expression of a gene, and its residual variance. This will artificially boost the weight of all highly expressed genes in downstream analysis such as dimensional reduction and clustering. We repeated the analysis with two alternative models, setting $\theta = \infty$ and $\theta = 10$, both of which revealed similar shortcomings in multiple datasets (Figure 2D and Supplementary Figures S3 - S10). We conclude that fixing a single value of θ may achieve effective performance in certain cases, but is unlikely to generalize across the diversity of systems and technologies represented by scRNA-seq data.

Gene overdispersion varies as a function of abundance

An alternative strategy for parameterizing θ leverages a well-characterized strategy for modeling counts in bulk RNA-seq data, where per-gene dispersion estimates have repeatedly been found to vary as a function of expression abundance (30–36). In sctransform (9), we aim to estimate a global relationship between gene abundance and θ by employing a regularization procedure where parameters are first fit for each gene individually, but information from genes with similar average abundances is subsequently pooled together in order to improve the robustness of parameter estimates. The underlying rationale for this choice is the non-decreasing relationship between gene abundance and θ that has been repeatedly observed in bulk RNA-seq studies (30–36). When analyzing each of the technologies and biological systems explored in this manuscript, we identified the same global patterns relating gene abundance and overdispersion levels (Supplementary Figures S11 - S14).

We also considered the findings from (22), which proposed that θ values should not vary as a function of gene abundance, and suggested that the relationship between these two variables was driven entirely by biases in the parameter estimation procedure, especially when analyzing lowly expressed genes. We first confirmed that lowly expressed genes, particularly those with average abundance < 0.1 UMI/cell, posed difficulties for parameter estimation. This is because the vast majority of count values for these genes are 0, creating inherent challenges in maximum likelihood estimation. When estimating parameters on simulated data drawn from a negative binomial with fixed θ, we confirmed a bias for these genes that resulted in decreased parameter estimates for θ (Supplementary Figure S15). However, using two complementary analyses, we found that this bias was not sufficient to explain the true relationships we observed in biological data. First, we observed a non-decreasing relationship between gene abundance (μ) and dispersion (θ) even when moving beyond the threshold for lowly expressed genes, which we did not observe when analyzing simulated data (Supplementary Figure S16). Additionally, we attempted to increase ('upsample') the depth of single cell datasets by pooling together molecular counts from cells with similar molecular profiles (Supplementary Methods) as inspired by the MetaCell framework (37). We repeated the parameter estimation procedure on metacells generated either from single-cell data, or using our simulation framework (Methods). Increasing the depth of sampling removed the effects of bias when analyzing simulated data, but preserved the observed relationship between μ and θ on real biological datasets (Supplementary Figure S16). We conclude that when modeling scRNA-seq data using a negative binomial distribution, the inverse overdispersion parameter θ does vary as a function of gene abundance, but the true nature of this relationship can be masked for genes with low molecular counts.

Recommendations for modeling heterogeneity in scRNA-seq datasets

Our findings highlight how the extensive diversity of scRNA-seq datasets poses challenges in identifying uniform procedures for the preprocessing and normalization of scRNA-seq data. Sparsely sequenced datasets may appear to be compatible with the use of Poisson error models, but datasets with additional sequencing depth reveal clear evidence of overdispersion. The level of overdispersion, exemplified by the NB parameter θ, also can vary substantially across datasets, technologies and systems, and even varies within a dataset as a function of gene abundance. However, the estimation of robust parameter estimates for θ can be challenging for lowly expressed genes, especially when analyzing datasets with sparse sequencing. We therefore considered recommendations for balancing these considerations, providing the flexibility to learn error models that can be robustly applied to our full spectrum of scRNA-seq datasets.

We first recommend the use of negative binomial observation model as an alternative to the Poisson distribution. Our analyses show that the Poisson distribution is a reasonable approximation for technical-control datasets consisting of endogenous or spike-in RNA, as well as for some scRNA-seq experiments with shallow sequencing. However, scRNA-seq datasets from cell lines exhibit clear evidence of overdispersion at higher sequencing depths, even for genes that do not correlate with changes in cell type or state. At least some of this overdispersion likely originates from 'intrinsic' noise, stochastic cellular variation that is inherent to the processes of mRNA transcription and degradation, and will affect the expression heterogeneity of all genes. While this variation is not a result of measurement error, it is not the primary focus of downstream scRNA-seq analyses, including the identification of cell types and states, and the inference of developmental trajectories. We therefore recommend that this variation be modeled independently of the latent space, which requires the use of a negative binomial error model. We note that the Poisson distribution is a special case of the negative binomial, and therefore the NB model can be successfully applied for datasets with very shallow sequencing, with appropriate parameter settings.

Second, we recommend learning negative binomial pa-

rameters separately for each dataset, rather than fixing them to a single value across all analyses. Moreover, we recommend allowing θ to vary across genes within a dataset, as a function of average gene abundance. The relationship between μ and θ has been repeatedly demonstrated in bulk RNA-seq, and is apparent across diverse scRNA-seq datasets as well, particularly for genes with sufficient sequencing depth. Using a fixed θ to parameterize all genes in a scRNA-seq dataset leads to ineffective variance stabilization and results in a global relationship between expression level and expression variance (Figure 2 and Supplementary Figures S3 and S4). We note that the recommendations described above relate not only to GLM-based preprocessing workflows, but also probabilistic or likelihood-based models (11, 24, 38).

Our analyses highlighted that lowly expressed genes with particularly sparse molecular counts often lacked sufficient information content to robustly detect overdispersion. We therefore designed a modified regularization procedure for learning GLM parameter estimates (Supplementary Methods). First, following the recommendations from (22), we fix the slope of the NB GLM to its analytically derived solution of $\ln(10)$, so that only the overdispersion and intercept parameters are estimated for each gene. Second, we reasoned that for genes with very low expression ($\mu < 0.001$), or where the variance of their molecular counts does not exceed the mean (i.e. $\sigma^2 \leq \mu$), we do not have sufficient evidence for overdispersion to fit negative binomial parameters. We therefore removed these genes from the regularization process and fixed their θ parameter to ∞, exemplifying a Poisson distribution. For example, in the scRNA-seq dataset of HEK cells profiled with SMART-Seq3, we removed $1,577$ genes (8.5%) at this stage, the majority of which were lowly expressed ($66.64\% < 0.1$ UMI/cell). We found that our modified regularization enables us to reproducibly learn gene-specific parameters even when using a subset of cells in the estimation procedure. This indicates increased robustness (Figure 3A), and allows us to learn a regularized relationship between μ and θ using only a subset of cells that achieves nearly identical results (Figure 3B) with increased computational efficiency (Figure 3C).

To test the broad applicability of this procedure, we applied it to each of the 58 datasets examined in this manuscript. In each case, we achieved effective variance stabilization as we observed no global relationship between gene expression levels and the variance of the resulting Pearson residuals (Supplementary Figures S17 - S20). Moreover, in each case, genes with the highest residual variance were distributed across a range of expression levels and - when analyzing heterogeneous samples - represented markers that have been strongly associated with individual cell types. As a result, application of this preprocessing pipeline will give the greatest weight to these markers, while downweighting fluctuations in the most highly expressed genes, which often appear to exhibit extensive heterogeneity in the absence of variance stabilization. These results indicate that our preprocessing workflow has sufficient flexibility to accurately model a wide variety of scRNA-seq datasets and serves as a basis for our recommendations in this manuscript.

Discussion

The application of statistical count models for preprocessing of scRNA-seq data overcomes important challenges that cannot be addressed by using linear size or scaling factor-based normalization. However, these techniques require the selection of an appropriate error distribution and accompanying parameter settings. Here, we explore these questions through the analysis of a wide diversity of scRNA-seq datasets varying across technologies, biological systems, and sequencing depths.

Our analyses revealed three key insights. First, we found that all scRNA-seq datasets exhibited clear evidence of overdispersion (i.e. deviation from a Poisson distribution), even after accounting for differences in sequencing depth, once exceeding a minimum expression level. This threshold varied across datasets as a function of average sequencing depth. This result strongly supports the use of negative binomial error models when analyzing UMI datasets. Second, we found that the negative binomial overdispersion parameter θ varied substantially across datasets, arguing against the use of a fixed θ estimate. Finally, we found that all datasets exhibited a dependence between gene abundance and overdispersion estimates. This result is robust even when considering potential biases in the overdispersion parameter estimation, and supports an empirical approach to learn regularized parameter estimates, as is commonly performed in bulk RNA-seq analysis.

Taken together, these results are compatible with the idea that cell-to-cell variation in scRNA-seq count data can be decomposed into multiple broad categories. The first represents variation in cell type and state which is biologically driven and encoded in cellular transcriptomes. This heterogeneity is the primary interest and focus of downstream analysis, and is typically represented in a latent space that can be learned via linear or non-linear dimensional reduction techniques. A second source represents technical measurement error arising from the stochastic loss of molecules during library preparation and sequencing. This sampling error can be modeled using a Poisson distribution and, particularly for shallowly sequenced datasets, represents a substantial source of remaining heterogeneity.

Our analyses suggest a third level of variation that should be accounted for: fluctuations in gene expression which are driven by the noise that is inherent to the processes of mRNA transcription and degradation (i.e. ‘intrinsic noise’) and manifests as overdispersion in scRNA-seq data. The presence of intrinsic noise has been extensively characterized and is an inevitable consequence of the gene regulatory process. Therefore, no two cells can generate mRNA molecules at exactly the same rate (an assumption of a Poisson process), even if they originate from the same ‘homogeneous’ population. Our analyses demonstrate that intrinsic noise is readily detectable for genes with sufficient sequencing depth, but can be masked in shallow datasets (Figure 1E). While intrinsic

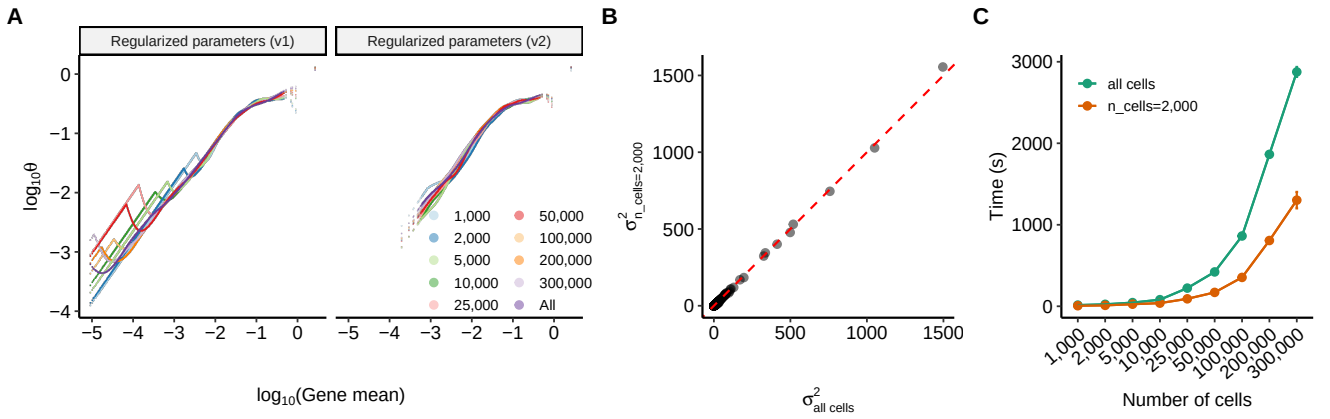

Figure 3. A modified regularization procedure improves the robustness of parameter estimates. A) Left: Estimated parameter estimates for θ on the Fetal sci-RNA-seq3 dataset (39), using the original regularization procedure from (9) (v1 regularization). Regularized estimates were learned using all cells (purple line), or downsampled cell subsets. Right: Same as (A), but using a modified procedure where the GLM slope was fixed, and genes where $\sigma^2 \leq \mu$ and $\mu < 0.001$ were excluded from regularization (v2 regularization) which improves robustness, and enables us to learn parameter estimates from a subsample of $2,000$ cells. **B)** Correlation of Pearson residual variance after applying a NB GLM with v2 regularization where parameters were estimated from all $377,456$ cells (x-axis), and a subsample of $2,000$ cells (y-axis). **C)** Green curve: total sctransform run time as a function of dataset size, using all cells to estimate parameters. Orange curve: total runtime when using a subsample of $2,000$ cells, which increases computational efficiency for large datasets.

noise it is not driven by measurement error, it should also be modeled independently of the latent space. Therefore, as the sensitivity and depth of scRNA-seq datasets continue to increase, the use of negative binomial error models will become increasingly important. Moreover, the level of intrinsic noise can vary across biological systems and gene abundance levels, motivating the use of a data-driven regularization procedure to learn gene-level overdispersion parameters.

Our analyses highlight the importance of considering a diversity of datasets when evaluating the statistical properties of new data types. While our results are therefore applicable to scRNA-seq measurements, they cannot be directly applied to new single-cell modalities, including protein measurements (i.e. CITE-seq (40)), chromatin accessibility profiles (i.e. scATAC-seq (41)), and DNA interaction maps (i.e. scCUT&TAG (42, 43)). As with cellular transcriptomes, these modalities can be profiled using multiple technologies that vary in their sensitivity and sparsity. We anticipate exciting future work that will characterize and parameterize heterogeneity in these data types, to achieve effective preprocessing and normalization.

Availability of materials and data

Raw datasets used in the main text are available from public URLs listed in Supplementary Table S1. Scripts to reproduce the analyses are available at https://github.com/saketkc/scRNA_NB_comparison.

Source code for sctransform along with the modifications described in this manuscript is available in the forked repository at https://github.com/saketkc/sctransform. A Python implementation that interfaces with the Scanpy (44) package is available at https://github.com/saketkc/pysctransform.

To invoke 'v2' regularization in `SCTransform` using Seurat (45):

```
library(Seurat)
object <- CreateSeuratObject(counts = counts)
object <- SCTransform(object, vst.flavor='v2')
```

Analogously, to use SCTransform in Python (using Scanpy (44)):

```
from pysctransform import SCTransform
adata = sc.read_h5ad("anndata_object.h5ad")
residuals = SCTransform(adata, vst_flavor='v2')
```

Acknowledgements

The authors would like to thank Christoph Hafemeister, and members of the Satija Lab for thoughtful discussions related to this work. This work was supported by the Chan Zuckerberg Initiative (EOSS-0000000082, HCA-A-1704-01895), and the NIH (RM1HG011014-02, 1OT2OD026673-01, DP2HG009623-01) to R.S.

Competing interests

In the past three years, R.S. has worked as a consultant for Bristol-Myers Squibb, Regeneron, and Kallyope and served as an SAB member for ImmunAI, Resolve Biosciences, Nanostring, and the NYC Pandemic Response Lab.

References

1. R. D. Hodge, J. A. Miller, M. Novotny, B. E. Kalmbach, J. T. Ting, T. E. Bakken, B. D. Aevermann, E. R. Barkan, M. L. Berkowitz-Cerasano, C. Cobbs, F. Diez-Fuertes, S.-L. Ding, J. McCorrison, N. J. Schork, S. I. Shehata, K. A. Smith, S. M. Sunkin, D. N. Tran, P. Venepally, A. M. Yanny, F. J. Steemers, J. W. Phillips, A. Bernard, C. Koch, R. S. Lasken, R. H. Scheuermann, and E. S. Lein, "Transcriptomic evidence that von economo neurons are regionally specialized extratelencephalic-projecting excitatory neurons," *Nat. Commun.*, vol. 11, p. 1172, Mar. 2020.
2. B. M. Colquitt, D. P. Merullo, G. Konopka, T. F. Roberts, and M. S. Brainard, "Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits," *Science*, vol. 371, Feb. 2021.
3. C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni, "Normalizing single-cell RNA sequencing data: challenges and opportunities," *Nat. Methods*, vol. 14, pp. 565–571, June 2017.
4. O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nat. Rev. Genet.*, vol. 16, pp. 133–145, Mar. 2015.

5. D. Grün, L. Kester, and A. van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nat. Methods*, vol. 11, pp. 637–640, June 2014.
6. S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson, "Quantitative single-cell RNA-seq with unique molecular identifiers," *Nat. Methods*, vol. 11, pp. 163–166, Feb. 2014.
7. D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit, "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types," *Science*, vol. 343, pp. 776–779, Feb. 2014.
8. V. Svensson, K. N. Natarajan, L. H. Ly, R. J. Miragaia, and others, "Power analysis of single-cell RNA-sequencing experiments," *Nature*, 2017.
9. C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression," *Genome Biol.*, vol. 20, p. 296, Dec. 2019.
10. F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, "Author correction: Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model," *Genome Biol.*, vol. 21, p. 179, July 2020.
11. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nat. Methods*, vol. 15, pp. 1053–1058, Dec. 2018.
12. G. La Manno, K. Siletti, A. Furlan, D. Gyllborg, E. Vinsland, and others, "Molecular architecture of the developing mouse brain," *BioRxiv*, 2020.
13. H. M. Levitin, J. Yuan, Y. L. Cheng, F. J. Ruiz, E. C. Bush, J. N. Bruce, P. Canoll, A. Iavarone, A. Lasorella, D. M. Blei, and P. A. Sims, "De novo gene signature identification from single-cell RNA-seq with hierarchical poisson factorization," *Mol. Syst. Biol.*, vol. 15, p. e8557, Feb. 2019.
14. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, pp. 1187–1201, May 2015.
15. F. Wagner, Y. Yan, and I. Yanai, "K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data," *BioRxiv*, 2017.
16. C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard, "Comparative analysis of Single-Cell RNA sequencing methods," *Mol. Cell*, vol. 65, pp. 631–643.e4, Feb. 2017.
17. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nat. Commun.*, vol. 8, p. 14049, Jan. 2017.
18. T. H. Kim, X. Zhou, and M. Chen, "Demystifying "drop-outs" in single-cell UMI data," *Genome Biol.*, vol. 21, p. 196, Aug. 2020.
19. L. Amrhein, K. Harsha, and C. Fuchs, "A mechanistic model for the negative binomial distribution of single-cell mRNA counts," *bioRxiv*, 2019.
20. B. Vieth, C. Ziegenhain, S. Parekh, W. Enard, and I. Hellmann, "powsimr: power analysis for bulk and single cell RNA-seq experiments," *Bioinformatics*, vol. 33, pp. 3486–3488, Nov. 2017.
21. L. He, J. Davila-Velderrain, T. S. Sumida, D. A. Hafler, M. Kellis, and A. M. Kulminski, "Nebula is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data," *Communications biology*, vol. 4, no. 1, pp. 1–17, 2021.
22. J. Lause, P. Berens, and D. Kobak, "Analytic pearson residuals for normalization of single-cell RNA-seq UMI data," *bioRxiv*, 2020.
23. D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J. P. Vert, "ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data," *BioRxiv*, 2017.
24. A. Gayoso, R. Lopez, G. Xing, P. Boyeau, K. Wu, and others, "scvi-tools: a library for deep probabilistic analysis of single-cell omics data," *bioRxiv*, 2021.
25. A. Sarkar and M. Stephens, "Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis," *Nature Genetics*, vol. 53, pp. 770–777, May 2021.
26. P. S. Swain, M. B. Elowitz, and E. D. Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, pp. 12795–12800, Oct. 2002.
27. M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, pp. 1183–1186, Aug. 2002.
28. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mRNA synthesis in mammalian cells," *PLoS Biol.*, vol. 4, p. e309, Oct. 2006.
29. M. Hagemann-Jensen, C. Ziegenhain, P. Chen, D. Ramsköld, G.-J. Hendriks, A. J. M. Larsson, O. R. Faridani, and R. Sandberg, "Single-cell RNA counting at allele and isoform resolution using smart-seq3," *Nat. Biotechnol.*, vol. 38, pp. 708–714, June 2020.
30. M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139–140, Jan. 2010.
31. Y.-H. Zhou, K. Xia, and F. A. Wright, "A powerful and flexible approach to the analysis of RNA sequence count data," *Bioinformatics*, vol. 27, pp. 2672–2678, Oct. 2011.
32. Y. Di, D. W. Schafer, J. S. Cumbie, and J. H. Chang, "The NBP negative binomial model for assessing differential gene expression from RNA-Seq," *Stat. Appl. Genet. Mol. Biol.*, vol. 10, Jan. 2011.
33. S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," *Genome Res.*, vol. 22, pp. 2008–2017, Oct. 2012.
34. D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Res.*, vol. 40, pp. 4288–4297, May 2012.
35. H. Wu, C. Wang, and Z. Wu, "A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data," *Biostatistics*, vol. 14, pp. 232–243, Apr. 2013.
36. M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.
37. Y. Baran, A. Bercovich, A. Sebe-Pedros, Y. Lubling, A. Giladi, E. Chomsky, Z. Meir, M. Hoichman, A. Lifshitz, and A. Tanay, "MetaCell: analysis of single-cell RNA-seq data using k-nn graph partitions," *Genome Biol.*, vol. 20, p. 206, Oct. 2019.
38. G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.
39. J. Cao, D. R. O'Day, H. A. Pliner, P. D. Kingsley, M. Deng, R. M. Daza, M. A. Zager, K. A. Aldinger, R. Blecher-Gonen, F. Zhang, M. Spielmann, J. Palis, D. Doherty, F. J. Steemers, I. A. Glass, C. Trapnell, and J. Shendure, "A human cell atlas of fetal gene expression," *Science*, vol. 370, Nov. 2020.
40. M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert, "Simultaneous epitope and transcriptome measurement in single cells," *Nat. Methods*, vol. 14, pp. 865–868, Sept. 2017.
41. J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf, "Single-cell chromatin accessibility reveals principles of regulatory variation," *Nature*, vol. 523, pp. 486–490, July 2015.
42. M. Bartosovic, M. Kabbe, and G. Castelo-Branco, "Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues," *Nat. Biotechnol.*, Apr. 2021.
43. S. J. Wu, S. N. Furlan, A. B. Mihalas, H. S. Kaya-Okur, A. H. Feroze, S. N. Emerson, Y. Zheng, K. Carson, P. J. Cimino, C. D. Keene, J. F. Sarthy, R. Gottardo, K. Ahmad, S. Henikoff, and A. P. Patel, "Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression," *Nat. Biotechnol.*, Apr. 2021.
44. F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biol.*, vol. 19, p. 15, Feb. 2018.
45. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, 2021.
46. P. K. Dunn and G. K. Smyth, "Randomized quantile residuals," *J. Comput. Graph. Stat.*, vol. 5, pp. 236–244, Sept. 1996.
47. C. Feng, L. Li, and A. Sadeghpour, "A comparison of residual diagnosis tools for diagnosing regression models for count data," *BMC Med. Res. Methodol.*, vol. 20, p. 175, July 2020.
48. J. D. Storey, "A direct approach to false discovery rates," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 64, pp. 479–498, Aug. 2002.
49. C. Ahlmann-Eltze and W. Huber, "glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data," *Bioinformatics*, vol. 36, pp. 5701–5702, Apr. 2021.
50. C. Hafemeister and R. Satija, "Analyzing scRNA-seq data with the sctransform and offset models," *https://satijalab.org/pdf/sctransform_offset.pdf*, 2020.
51. V. Svensson, "Droplet scRNA-seq is not zero-inflated," *Nat. Biotechnol.*, vol. 38, pp. 147–150, Feb. 2020.
52. J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, L. T. Nguyen, J. Y. H. Kwon, B. Barak, W. Ge, A. J. Kedaigle, S. Carroll, S. Li, N. Hacohen, O. Rozenblatt-Rosen, A. K. Shalek, A.-C. Villani, A. Regev, and J. Z. Levin, "Systematic comparison of single-cell and single-nucleus RNA-sequencing methods," *Nat. Biotechnol.*, vol. 38, pp. 737–746, June 2020.
53. E. Mereu, A. Lafzi, C. Moutinho, C. Ziegenhain, D. J. McCarthy, A. Álvarez-Varela, E. Batlle, Sagar, D. Grün, J. K. Lau, S. C. Boutet, C. Sanada, A. Ooi, R. C. Jones, K. Kaihara, C. Brampton, Y. Talaga, Y. Sasagawa, K. Tanaka, T. Hayashi, C. Braeuning, C. Fischer, S. Sauer, T. Trefzer, C. Conrad, X. Adiconis, L. T. Nguyen, A. Regev, J. Z. Levin, S. Parekh, A. Janjic, L. E. Wange, J. W. Bagnoli, W. Enard, M. Gut, R. Sandberg, I. Nikaido, I. Gut, O. Stegle, and H. Heyn, "Benchmarking single-cell RNA-sequencing protocols for cell atlas projects," *Nat. Biotechnol.*, vol. 38, pp. 747–755, June 2020.

Supplementary Methods

Data sources and preprocessing

All datasets were obtained as preprocessed count matrices from Gene expression omnibus (GEO), EBI ArrayExpress, or author's website. In each case, we utilized cells that had passed the QC thresholds set by the original study authors. However, to minimize the effect of potential cell outliers in our data, we filtered out cells that fell outside of the 5% to 95% UMI quantiles in each dataset. Additionally, we removed all cells where more than 15% of reads mapped to mitochondrial transcripts. We did not perform any filtering for the Fetal sci-RNA-seq3 dataset as it had already been filtered and annotated by the authors. The dataset source and associated publication are available in Supplementary Table S1.

Goodness of Fit test using a Poisson GLM

To explore whether a Poisson distribution represents an appropriate error model for UMI-based scRNA-seq data, we fit a Poisson GLM adjusting for differences in library size modeled as an offset. To reduce the computational complexity, we subsampled $1,000$ cells in a density dependent manner from the count matrices of each dataset: the probabilty of selecting a cell c is $\frac{1}{d(\log_{10} N_c)}$, where d is the density estimate of all $\log_{10}$-transformed total cell UMIs and N_c is the total UMI counts in cell c. These subsampled count matrices were then used to fit a Poisson GLM for each gene UMI vector with total UMI content of each cell modeled as an offset vector (`glm.fit(gene_umi ~ 1, offset=log(total_umi), family=Poisson(link="log"))` in `R`. We then performed a goodness of fit test on the randomized quantile residuals (46) of this GLM model fit calculated using `statmod::qresid(model)`. If the data is well-described by the model, the sum of squares of the quantile residuals is expected to follow a Chi-squared distribution with degrees of freedom = $N_{\text{cells}} - 1$ where N_{cells} represents the total number of cells in the dataset. We chose quantile residuals to measure the goodness of fit, as they have lower type-I error and higher power in comparison to other residuals for identifying misspecification (47). To calculate p-values, we used the `pchisq` function in R (`pchisq(qresid, df=model$df.residual, lower.tail=FALSE)`). To control for multiple testing, we adjusted p-values using the `qvalue` method available through the qvalue package (48). We used a q-value threshold of 0.01 to accept or reject the fit to the Poisson model. Library sizes reflected in Figure 1E were calculated based on the subset count matrices.

Assessing overdispersion after downsampling sequence depth

In Figure 1D-E we assess the level of dispersion in the PBMC Smart-seq3 dataset, after downsampling the dataset to different sequencing depths. The full dataset contains $2,629$ cells with a median UMI/cell of $8,288$ with a maximum coverage of $20,463$ UMI/cell. When downsampling to $10,000$ UMI/cell, we first excluded $1,837$ cells where $< 9,900$ UMIs were detected in the dataset. For the remaining cells, we randomly sampled molecules at a proportion expected to yield $10,000$ UMI/cell on average and retained only cells that contained UMIs in the range $10,000 \pm 100$ to minimize the effect of differences in sequencing depth. We repeated this process for multiple sequencing depths shown in Figure 1D-E.

Comparing levels of overdispersion across datasets

In Figure 2A-B, we fit NB GLM to each gene in each dataset, in order to estimate the inverse overdispersion parameter θ. We model the observed counts for each gene using the following model `gene_umi` ~ 1, and estimate parameters using `glmGamPoi::glm_gp(gene_umi, model, offset=log(total_umi), size_factors=FALSE)` using the glmGamPoi package (49). We perform this procedure for all genes where the variance of the observed counts exceeds the mean.

Modeling scRNA-seq datasets with sctransform

For clarity, we restate the modeling framework used in sctransform (9). In sctransform, UMI counts across cells in a dataset are modeled using a generalized linear model (GLM). The total UMI count per cell is used as an offset in the GLM. Thus, for a given gene g in cell c, we have

$$\begin{aligned} x_{gc} &\sim \text{NB}(\mu_{gc}, \theta_g) \\ \ln \mu_{gc} &= \beta_{g0} + \ln n_c, \end{aligned}$$

where θ_g is the gene-specific dispersion parameter, $n_c = \sum_g x_{gc}$ is the total sequencing depth and the variance of the negative binomial (NB) is given by $\mu_{gc} + \mu_{gc}^2/\theta_g$.

We perform three steps to remove technical noise and perform variance stabilization. In the first step, the inverse overdispersion parameter θ is individually estimated using a subset of genes (2000 by default), which are sampled in a density-dependent manner according to their average abundance. In the next step, we calculate a smoothed curve that characterizes the global relationship between μ and θ, thereby regularizing θ estimates as a function of gene mean. We perform the same regularization for the intercept parameter. We use the geometric mean to estimate gene abundance, which is more robust to outlier values in scRNA-seq. As outlier values can originate from multiple sources including the presence of cell doublets, errors in UMI collapsing, or ambient RNA, we have empirically improved performance when using the geometric mean instead of the arithmetic mean. Although sctransform supports multiple estimators for θ, we recommend the use of glmGamPoi (49), an alternate estimator that is more robust and faster.

In the final step, we use the regularized parameter estimates to calculate Pearson residuals Z_{gc}. For each gene-cell combination, the Pearson residuals Z_{gc} are given by

$$
\begin{aligned}
Z_{gc} &- \frac{x_{gc} - \mu_{gc}}{\sigma_{gc}} \\
\mu_{gc} &= \exp \beta_{g0} + \ln n_c \\
\sigma_{gc} &= \sqrt{\mu_{gc} + \frac{\mu_{gc}^2}{\theta_{gc}}},
\end{aligned}
$$

The 'residual variance' for a gene represents the remaining variation in gene expression that is not explained by the sctransform model, and is defined as:

$$
\begin{aligned}
\text{residual variance}_g &= \frac{1}{C-1} \sum_{c=1}^{C} (Z_{gc} - \bar{Z}_g)^2 \\
\bar{Z}_g &= \sum_{c=1}^{C} Z_{gc},
\end{aligned}
$$

where C represents the number of total cells in the dataset.

Evaluating the performance of a GLM with fixed θ

In Figure 2C-D, as well as associated supplementary figures (S3 - S10) we model each of the scRNA-seq datasets using a NB GLM with a fixed value of θ for each gene in each dataset. To test this, we utilize the 'offset' model as described by Lause et al. in (22). We repeated the analysis with three different values for the fixed overdispersion parameter, $\theta = \infty$, $\theta = 100$, and $\theta = 10$.

Improving the robustness of parameter regularization

In Figure 3 we explore a modified regularization procedure to improve the robustness of NB parameter estimates, particularly for lowly expressed genes, and to increase computational efficiency. We make two changes to the estimation procedure described in (9). First, we fix the slope parameter of the GLM to $\ln(10)$ with $\log_{10}(\text{total UMI})$ used as the predictor. As described in (22), this value represents the analytically derived solution for this parameter, and closely mirrors the regularized values we had obtained for the slope parameter in the original sctransform procedure. While (22) also recommends fixing the intercept parameter for the GLM, an approximate solution to the maximum likelihood estimate of this parameter can only be obtained for large values of θ. As our data-driven estimates for θ demonstrate that this parameter can vary substantially across datasets, we continue to set the intercept parameter for the GLM through regularization.

As a second modification, we remove a subset of genes prior to performing regularization. In particular, we reasoned that for genes with either very low abundance ($\mu < 0.001$), or where the variance of count values did not exceed the average abundance (i.e. $\sigma^2 \leq \mu$), we lacked sufficient information to learn robust NB parameter estimates. We therefore exclude these genes from the regularization procedure, and set their θ parameter estimates to ∞ for all downstream analyses. We note that this filtration step occurs rapidly, as the per-gene mean and variance can be efficiently calculated. For this filtration step, we use the arithmetic mean to set abundance, as this value should be compared with gene variance to determine evidence of overdispersion. For these genes, the regularized intercept ($\hat{\beta}_{g0}^{\text{poisson}}$) is set to the analytically derived solution from (22), with a fixed slope of $\ln(10)$:

$$
\hat{\beta}_{g0}^{\text{poisson}} = \ln\left(\sum_c x_{gc}\right) - \ln\left(\sum_c n_c\right)
$$

Simulation of UMI counts with fixed overdispersion

To explore the potential bias of maximum-likelihood (ML) estimators, we simulated an scRNA-seq dataset with fixed levels of overdispersion. We fixed θ to different values $\{0.001, 0.01, 0.1, 1, 10, 100\}$, and simulated scRNA-seq counts from an NB distribution, using gene means that were taken from the PBMC Smart-Seq3 dataset. We next estimated parameter values for θ using both the v1 and v2 versions of our sctransform regularization procedure using glmGamPoi (49) as an estimation engine. We also estimated a maximum likelihood of θ using glmGamPoi without explicitly accounting for library size (MLE). To create an 'upsampled' dataset where the sequencing depth is higher, we multiplied the estimated means x_{gc} by a factor of 500, and repeated the sampling procedure.

Increasing sequencing depth by creating metacells

In order to 'upsample' the PBMC Smart-seq3 dataset, we ran MetaCells v0.3.5 (37) for three different values of 'K' parameter (200, and 300, and 400) with all other parameters as defaults. UMI counts of cells belonging to one metacell were consolidated to create a metacell count, resulting in a higher sequencing depth. These metacells were then used as input to sctransform to estimate per gene θ.

Supplementary Figures & Tables

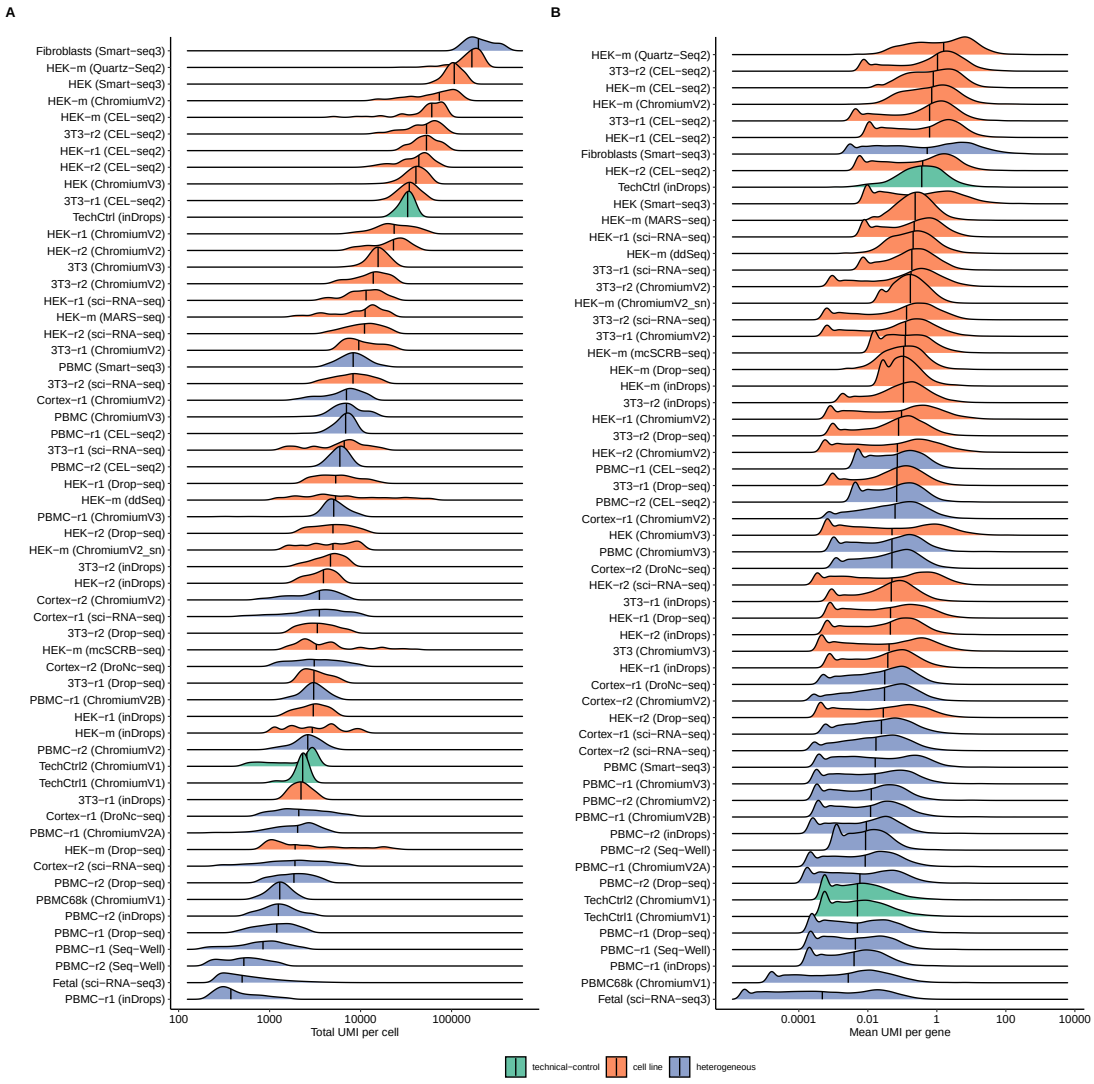

Figure S1. UMI statistics for the 58 datasets analyzed in this manuscript. A) Distribution of total UMI per cell across datasets **B)** Distribution of mean UMI per gene across datasets (technical control = endogeneous or spike-in RNA; cell line = HEK293 and 3T3 cell lines; heterogeneous = samples consisting of multiple cell types).

Figure S2. Genes that exhibit Poisson heterogeneity are more lowly expressed. In each dataset, we performed a per-gene goodness-of-fit test based on a GLM with a Poisson error model (Supplementary Methods). Shown are the distribution of gene abundances (average UMI/gene) for genes that passed (blue) and failed (red) the goodness-of-fit test.

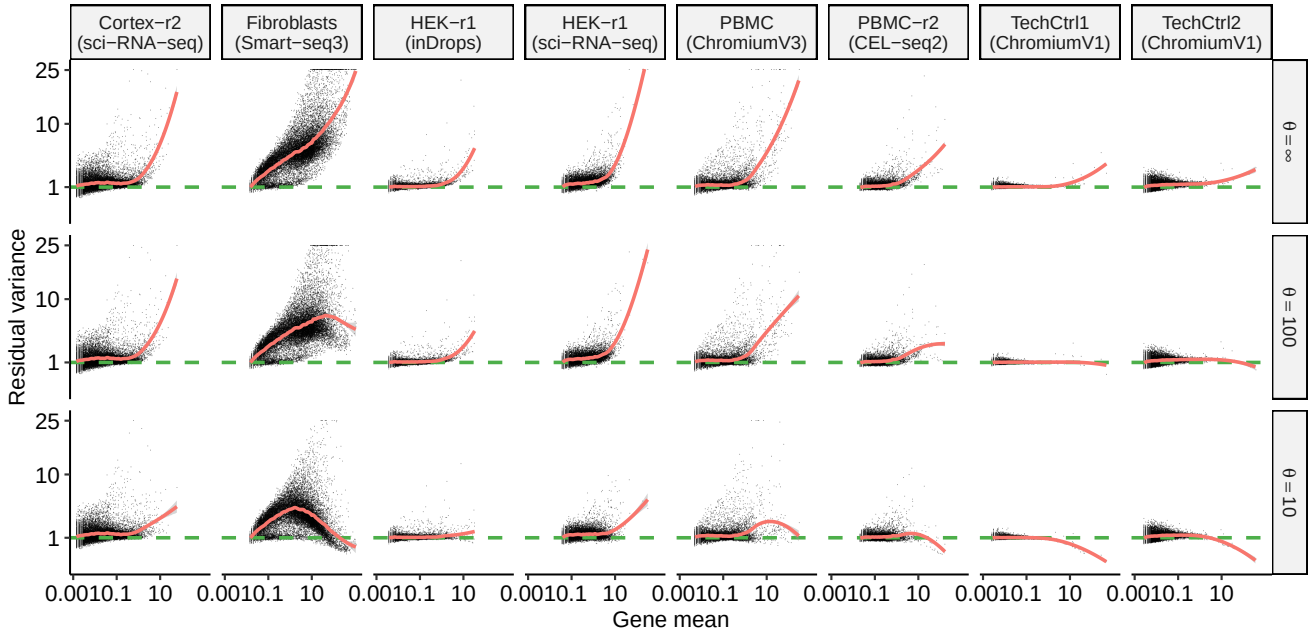

Figure S3. Relationship between gene abundance and the variance of Pearson residuals. Values shown are resulting from an NB GLM with three different values of θ. Same as Figure 2D but with per-gene estimates highlighted instead of smoothed curves.

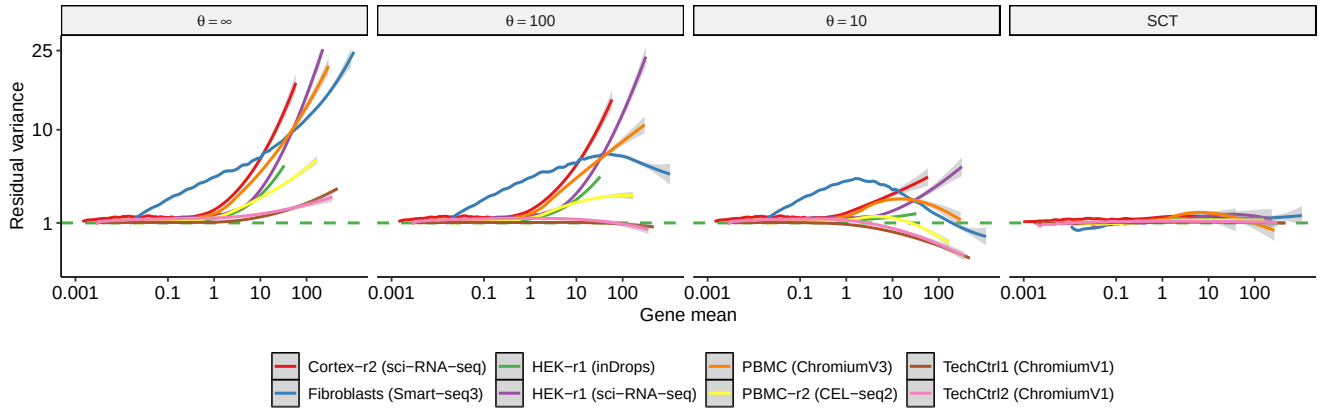

Figure S4. Relationship between gene abundance and the variance of Pearson residuals. Same as Figure 2D but additionally showing results for sctransform (v2 regularization).

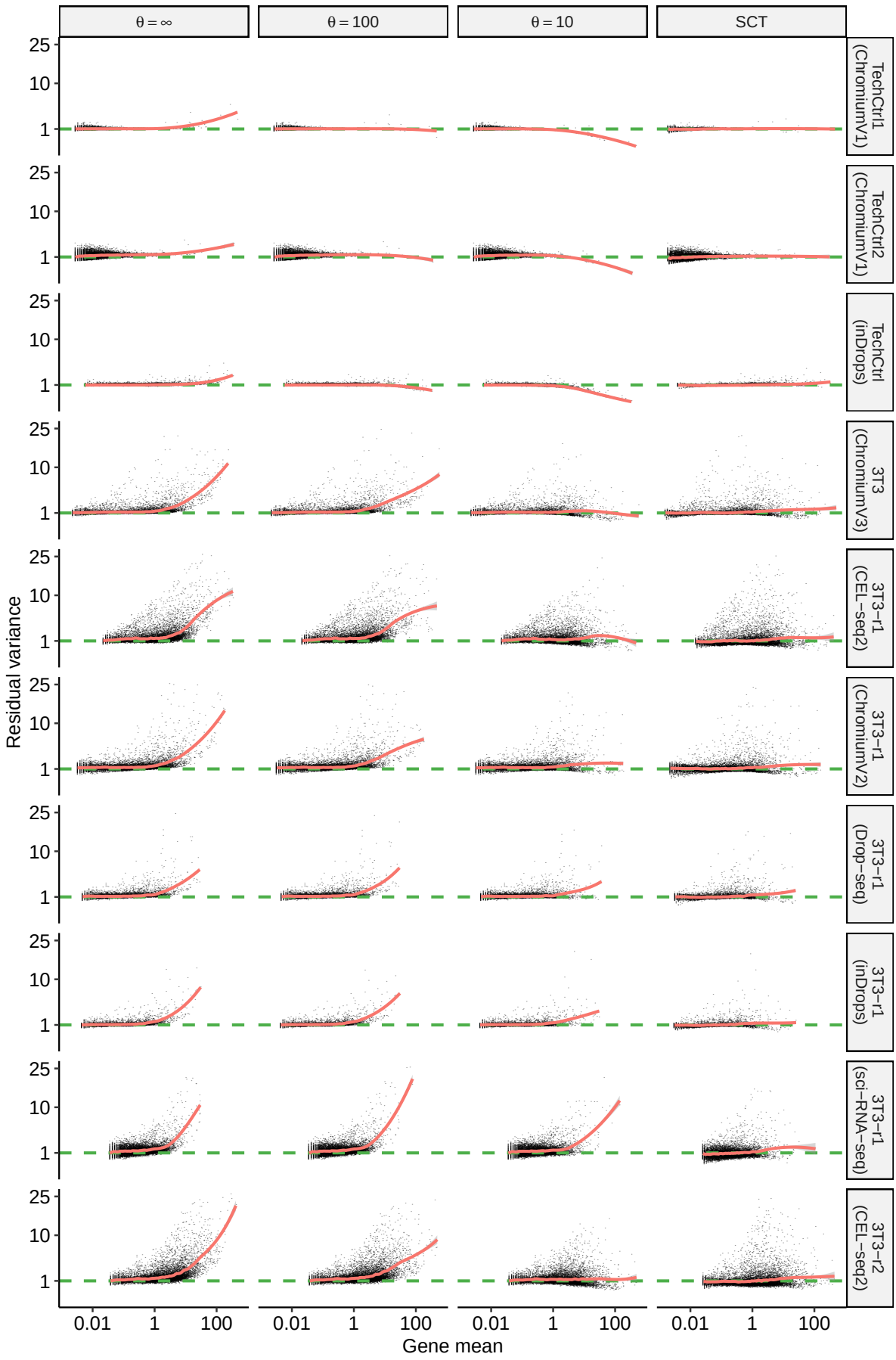

Figure S5. Evaluating variance stabilization for different error models. Same as in Figure 2C, but with additional datasets 1-10. Also shown are results from sctransform (v2 regularization).

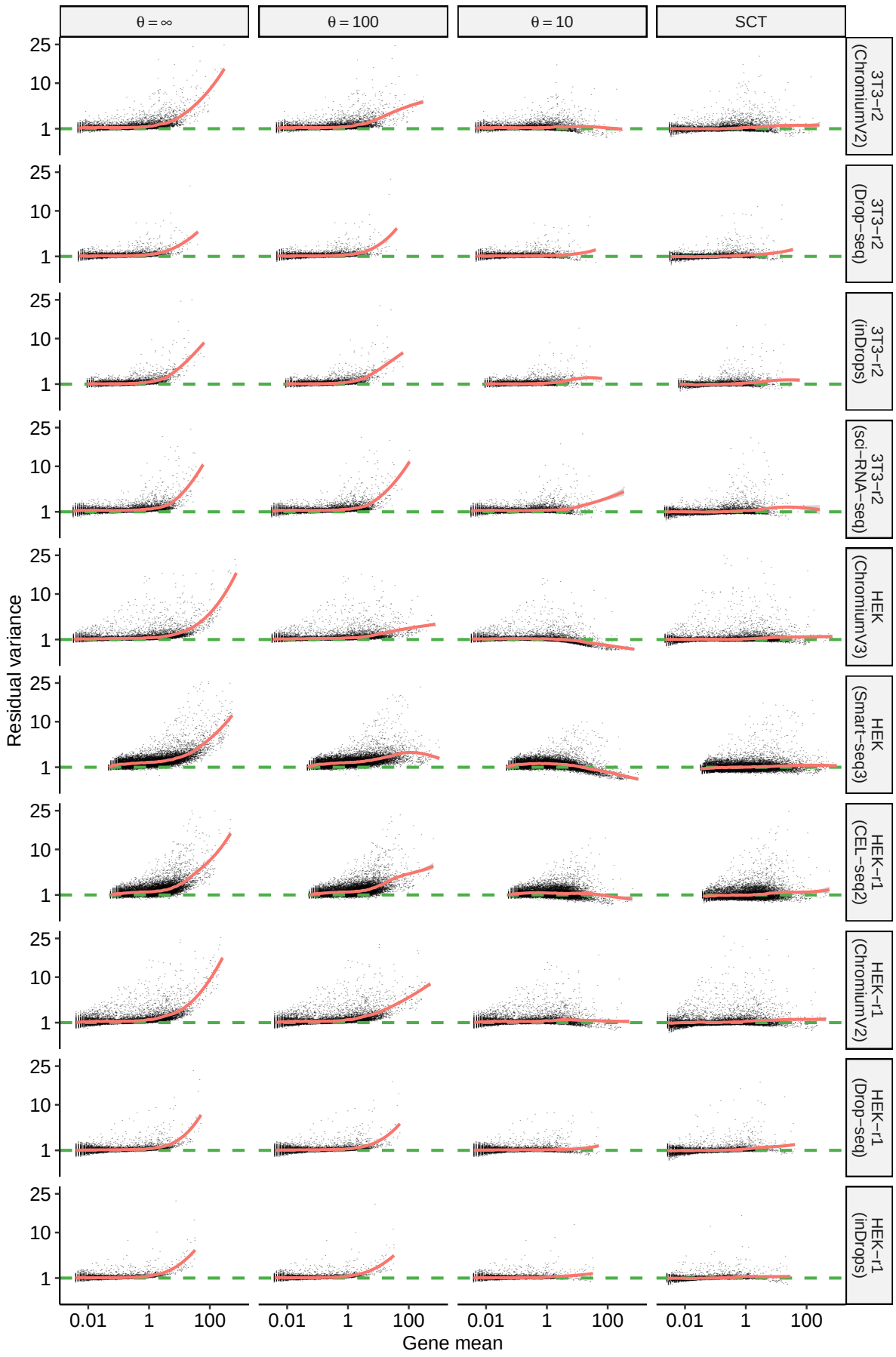

Figure S6. Evaluating variance stabilization for different error models. Same as in Figure 2C, but with additional datasets 11-20. Also shown are results from sctransform (v2 regularization).

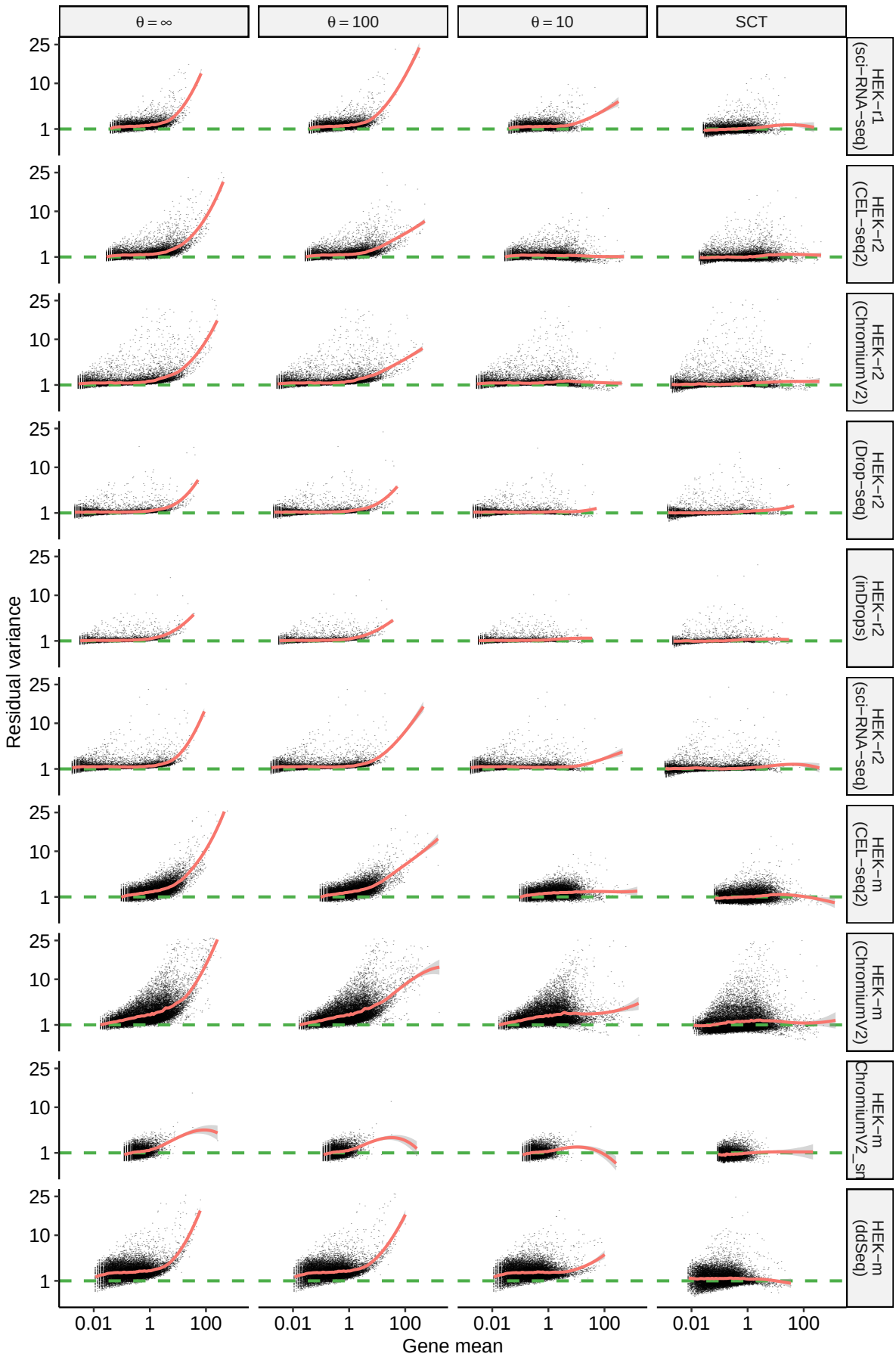

Figure S7. Evaluating variance stabilization for different error models. Same as in Figure 2C, but with additional datasets 21-30. Also shown are results from sctransform (v2 regularization).

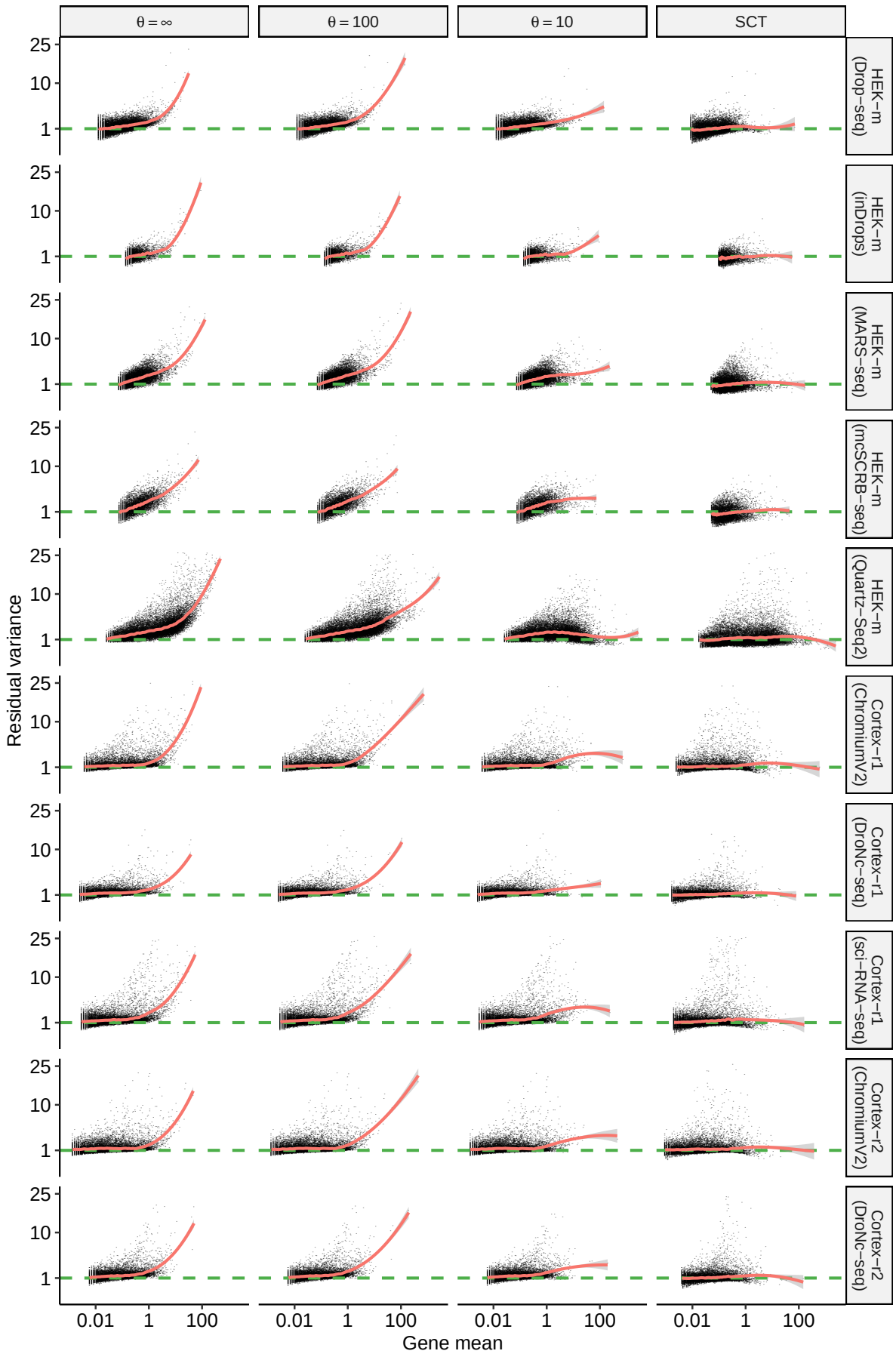

Figure S8. Evaluating variance stabilization for different error models. Same as in Figure 2C, but with additional datasets 31-40. Also shown are results from sctransform (v2 regularization).

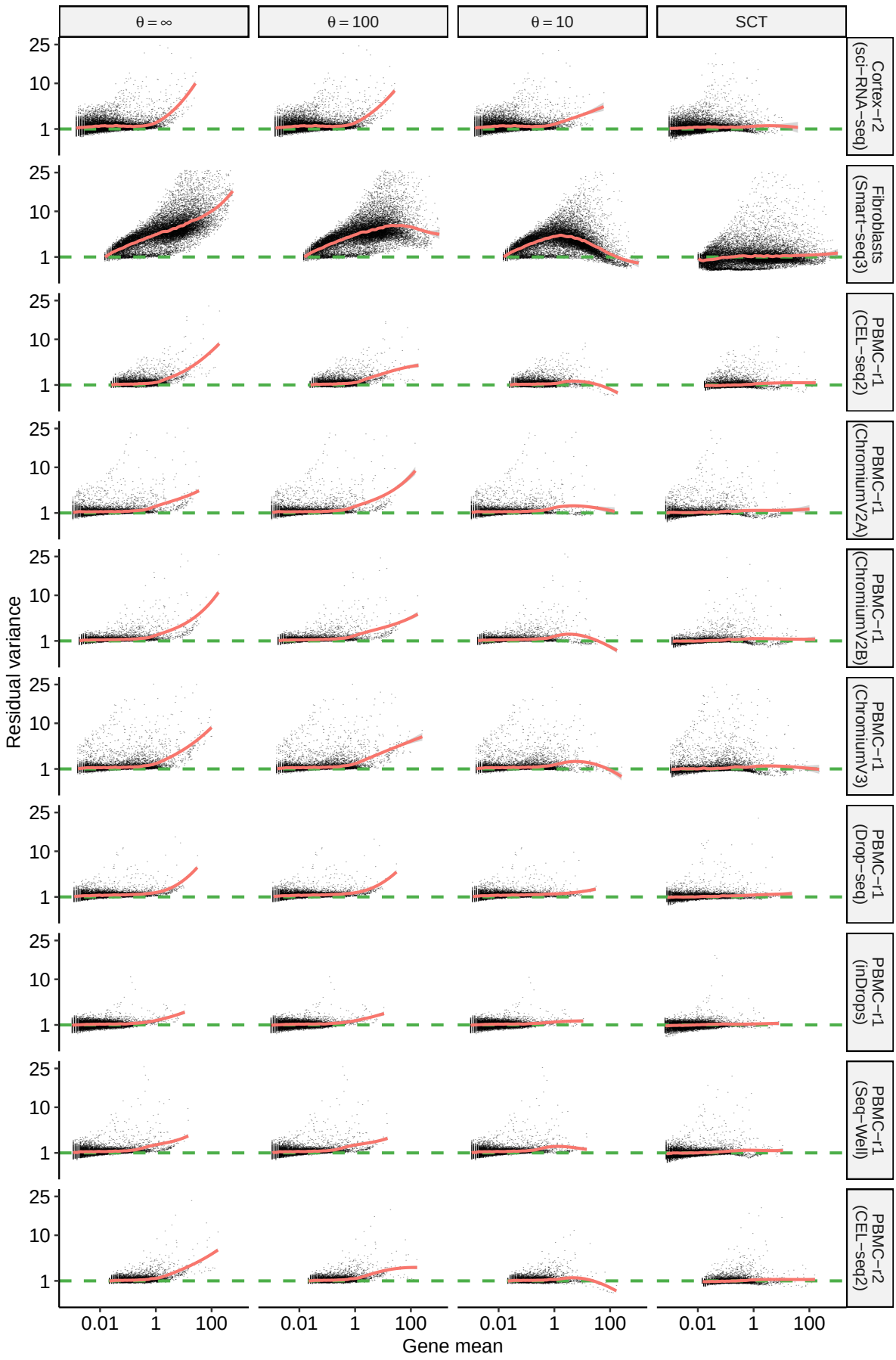

Figure S9. Evaluating variance stabilization for different error models. Same as in Figure 2C, but with additional datasets 41-50. Also shown are results from sctransform (v2 regularization).

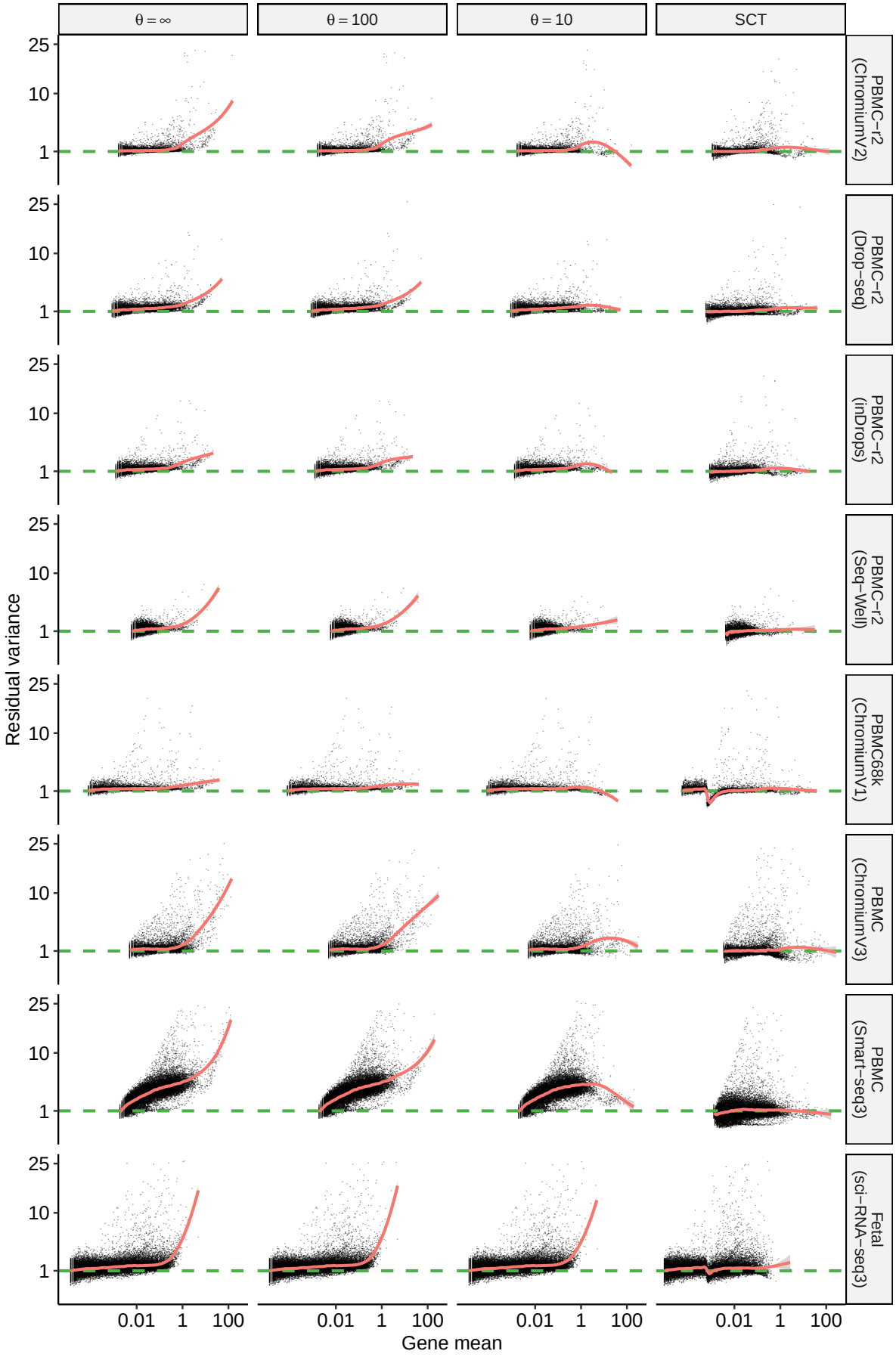

Figure S10. Evaluating variance stabilization for different error models. Same as in Figure 2C, but with additional datasets 51-58. Also shown are results from sctransform (v2 regularization).

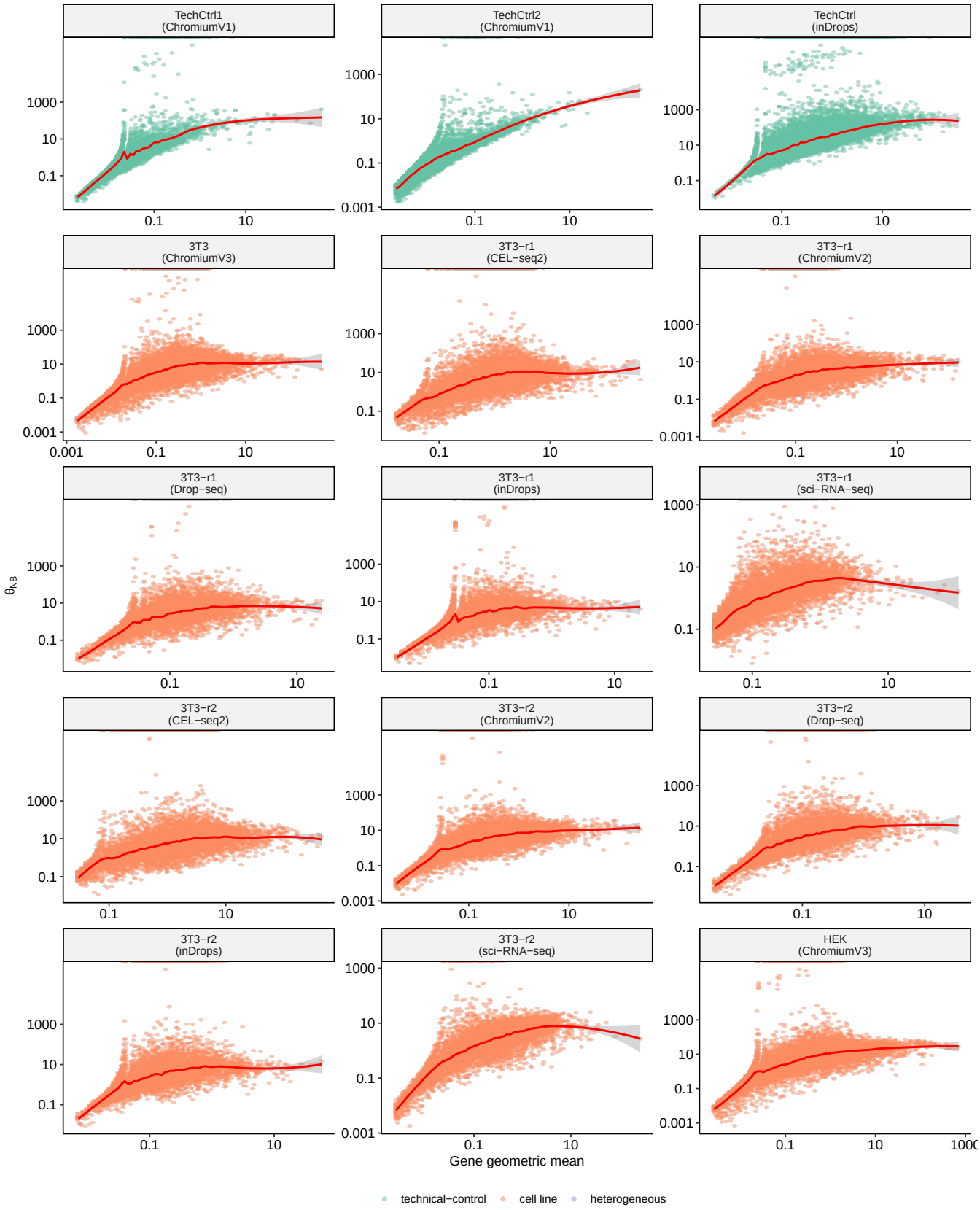

Figure S11. Relationship between inverse overdispersion parameter θ and gene abundance μ. Overdispersion was estimated using all cells after accounting for library size using a negative-binomial GLM. The red curve indicates a LOESS fit. All datasets exhibit a relationship between gene mean and θ [Datasets 1-15].

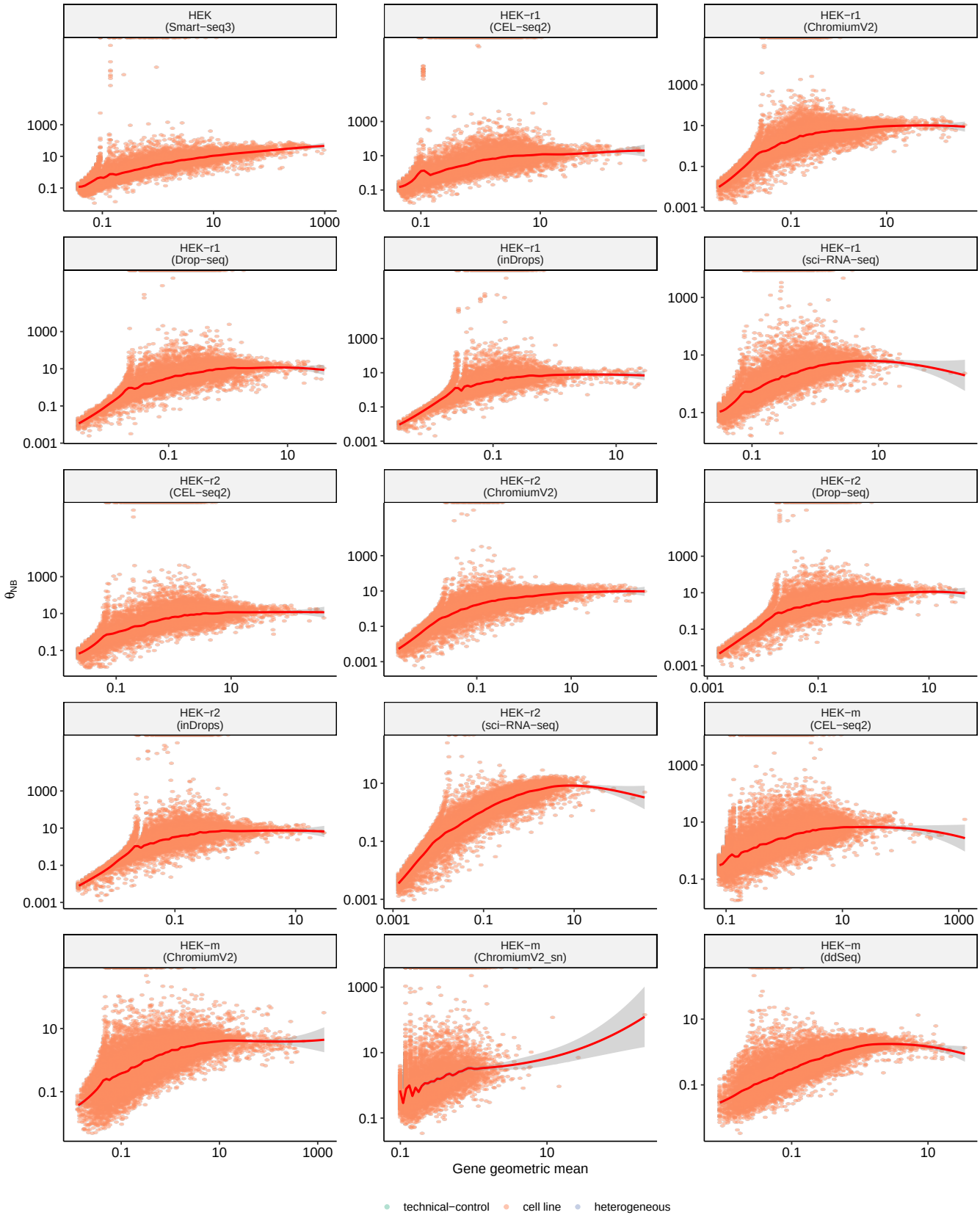

Figure S12. Relationship between inverse overdispersion parameter θ and gene abundance μ. Overdispersion was estimated using all cells after accounting for library size using a negative-binomial GLM. The red curve indicates a LOESS fit. All datasets exhibit a relationship between gene mean and θ [Datasets 16-30].

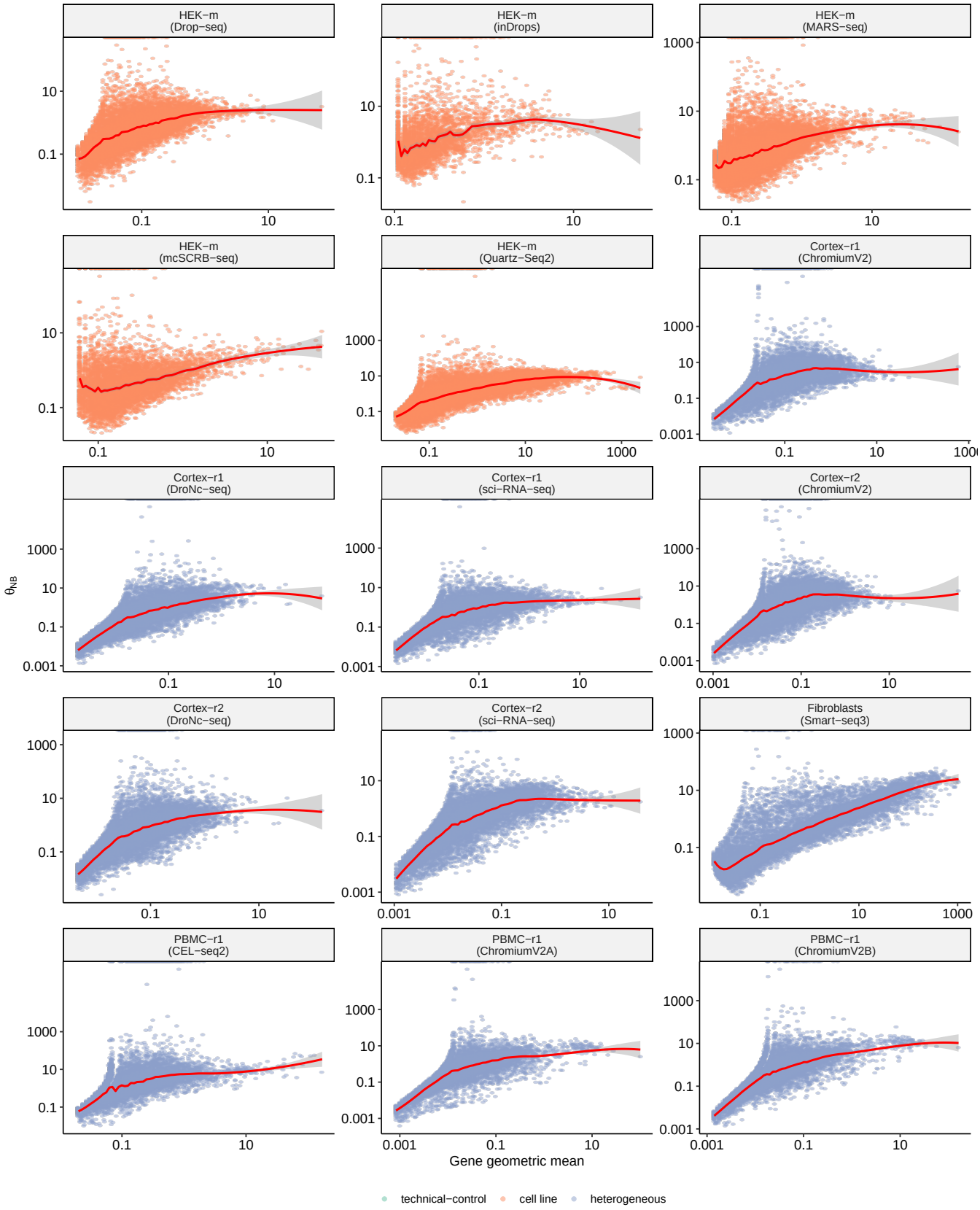

Figure S13. Relationship between inverse overdispersion parameter θ and gene abundance μ. Overdispersion was estimated using all cells after accounting for library size using a negative-binomial GLM. The red curve indicates a LOESS fit. All datasets exhibit a relationship between gene mean and θ [Datasets 31-45].

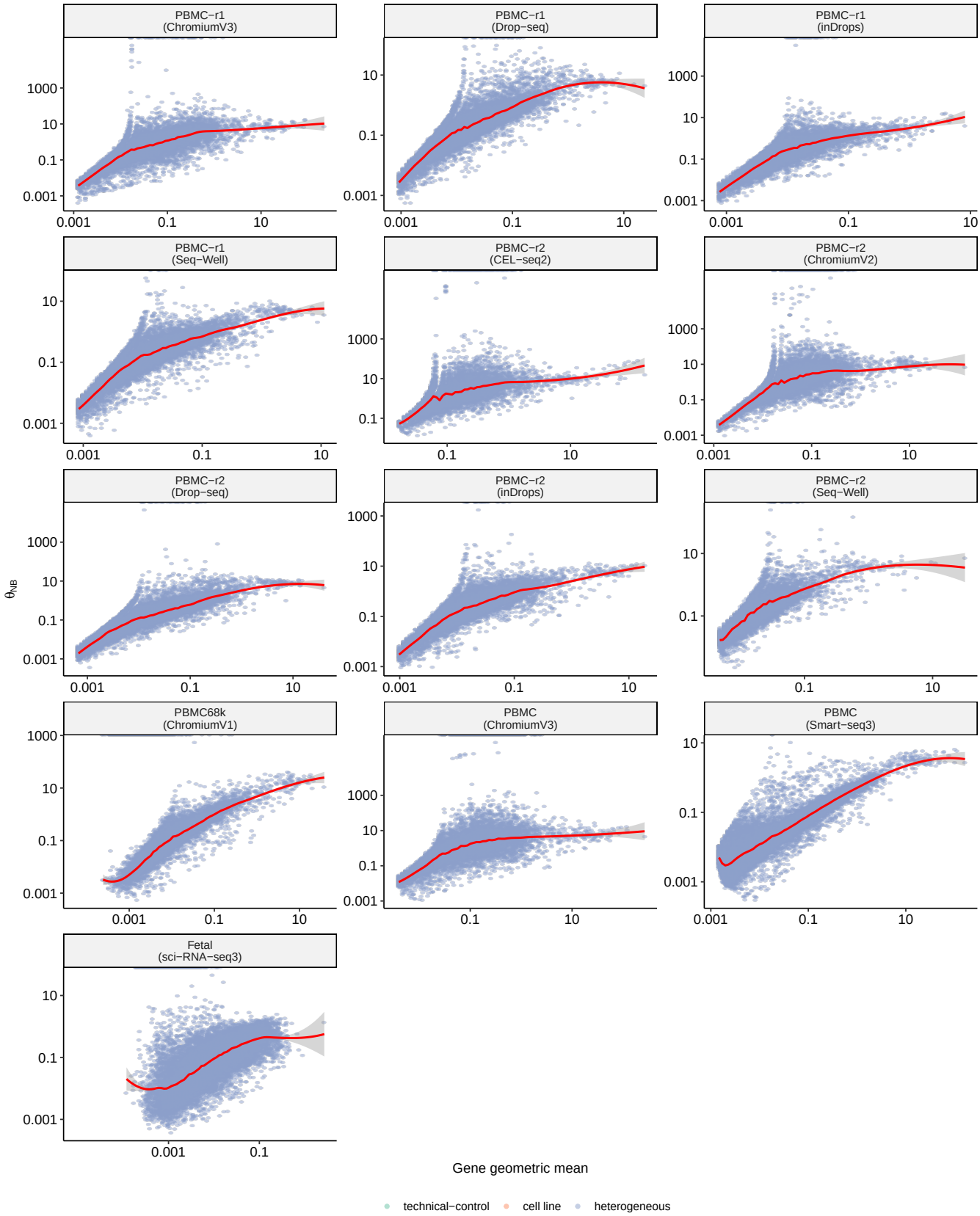

Figure S14. Relationship between inverse overdispersion parameter θ and gene abundance μ. Overdispersion was estimated using all cells after accounting for library size using a negative-binomial GLM. The red curve indicates a LOESS fit. All datasets exhibit a relationship between gene mean and θ [Datasets 46-58].

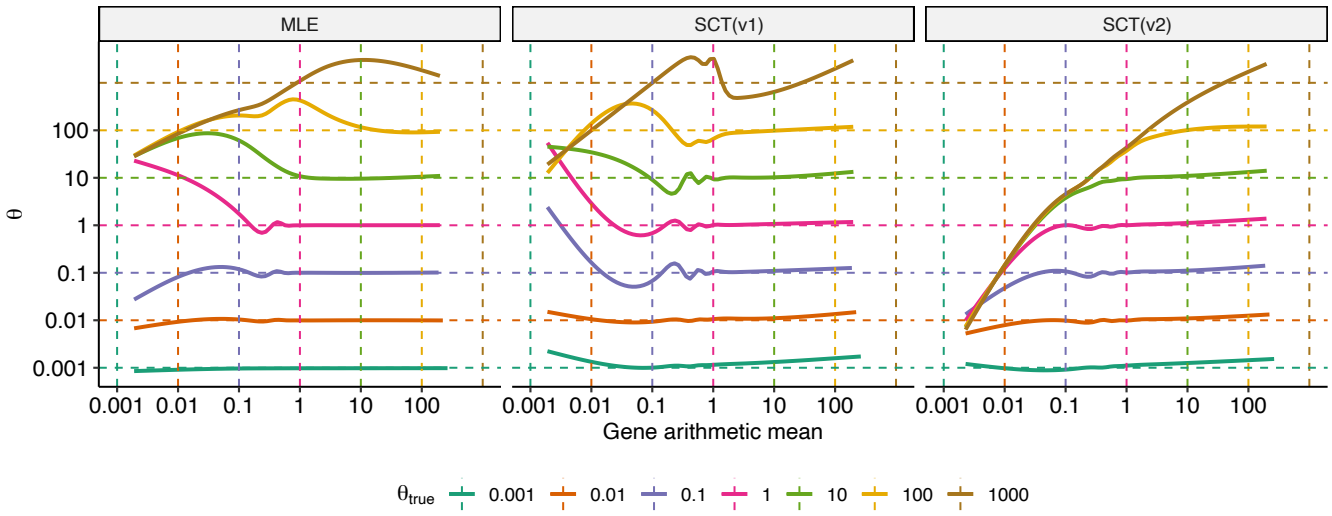

Figure S15. Estimation of dispersion in simulated datasets. Using mean counts distribution from PBMCs profiled using Smart-seq3, synthetic counts matrices were generated using a fixed $\theta = \{0.001, 0.01, 0.1, 1, 10, 100\}$. There is a bias in estimated θ from all the three methods: MLE (glmGamPoi (49)), SCT (sctransform) and SCT2 (sctransform, v2 regularization). The bias arises from difficulty in estimating the true θ when $\mu < 1$ and $\mu < \theta$. We note that the variance of the NB model is given by $\mu_{gc} + \mu_{gc}^2/\theta_g$. The second term approaches 0 for small values of μ, which is where we observe this bias. Therefore, the bias in parameter estimation has minimal impact on both the expected NB variance, and the final Pearson residuals (50).

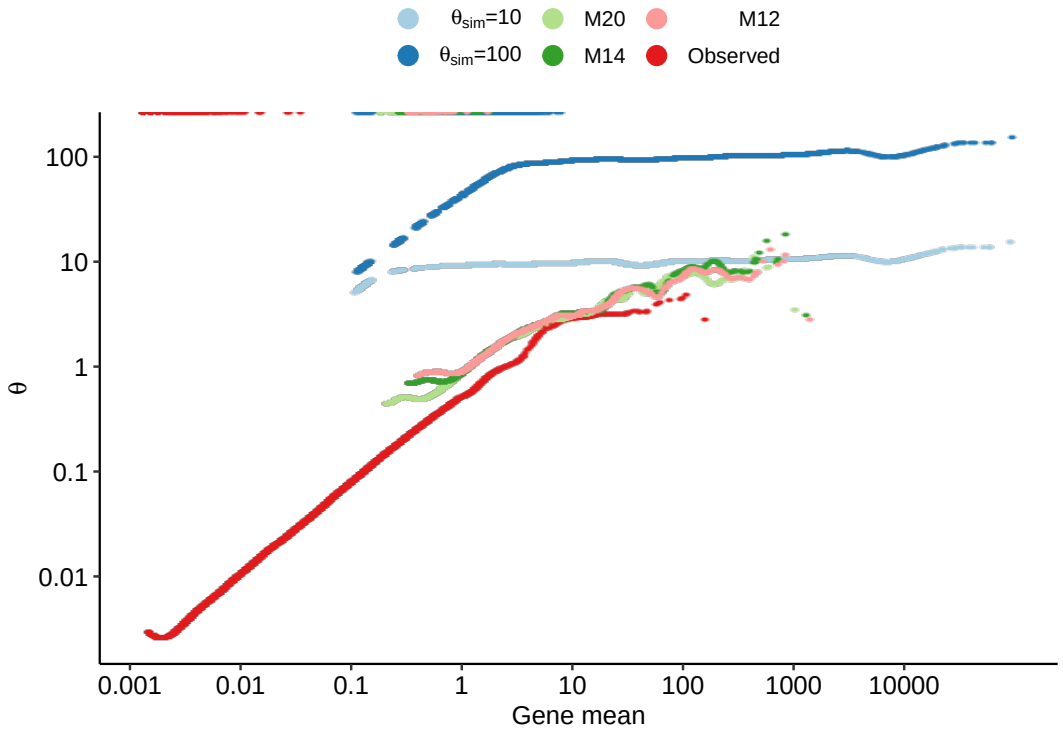

Figure S16. Effect of 'upsampling' on $\mu - \theta$ relationship Relationship between gene mean and dispersion observed in PBMC Smart-seq3 dataset, simulated dataset with different true dispersions ($\theta_{\text{sim}} = 10$ and 100) and Metacells (M20, M14, M12). The simulated datasets were generated using mean counts from the observed PBMC Smart-seq3 dataset, but by 'upsampling' the means to be 500 times larger. Metacells were generated using MetaCell (37) using different parameters of K for the KNN graph. M20, M14, and M12 represents 20, 14, and 12 metacells constructed using K = 200, 300, and 400 respectively.

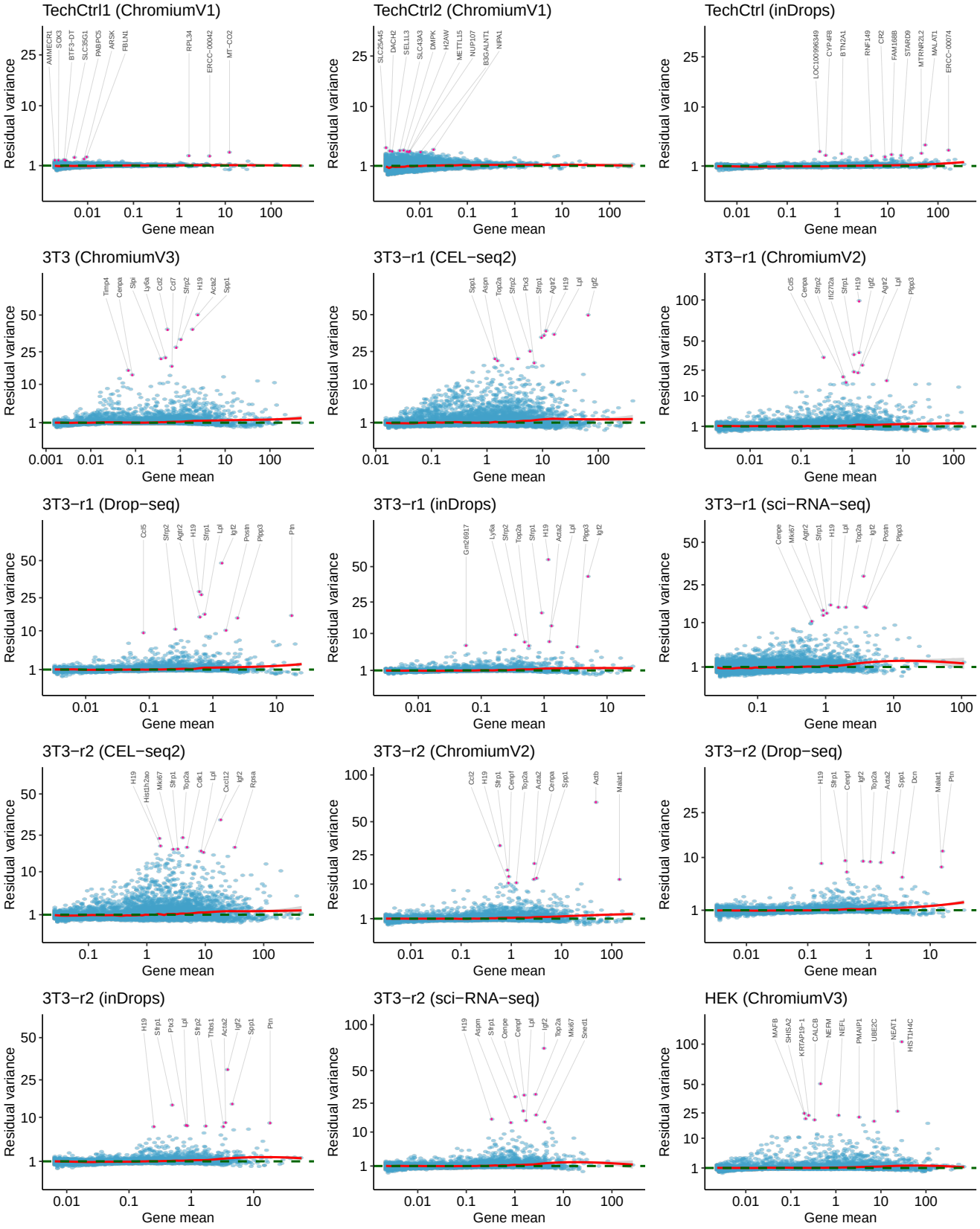

Figure S17. Variance stabilization achieved by sctransform2 across datasets. Y-axis shows variation of pearson residuals as estimated by sctransform (v2 regularization) for datasets 1-15. Top 10 genes with highest residual variances are highlighted.

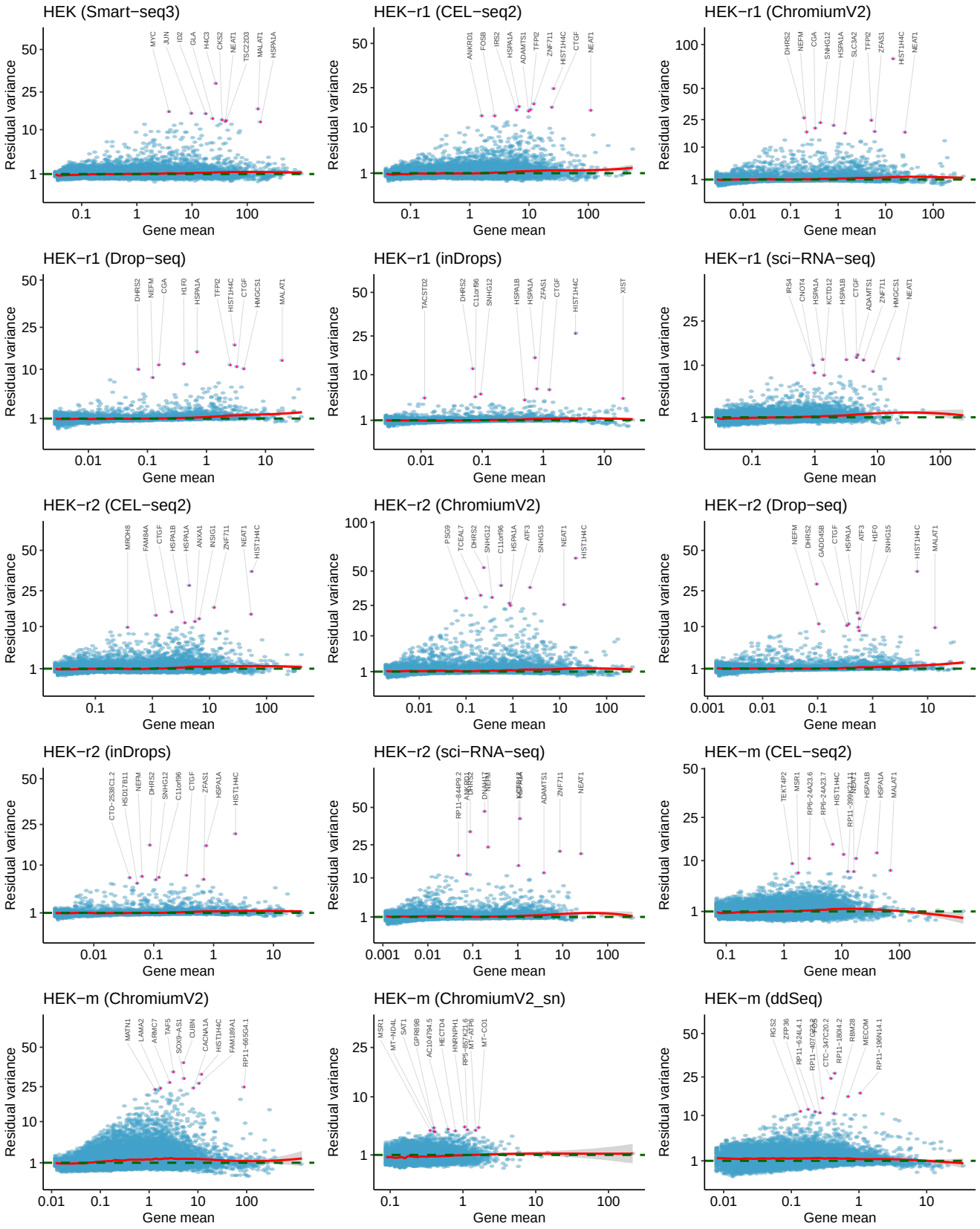

Figure S18. Variance stabilization achieved by sctransform2 across datasets. Y-axis shows variation of pearson residuals as estimated by sctransform (v2 regularization) for datasets 16-30. Top 10 genes with highest residual variances are highlighted.

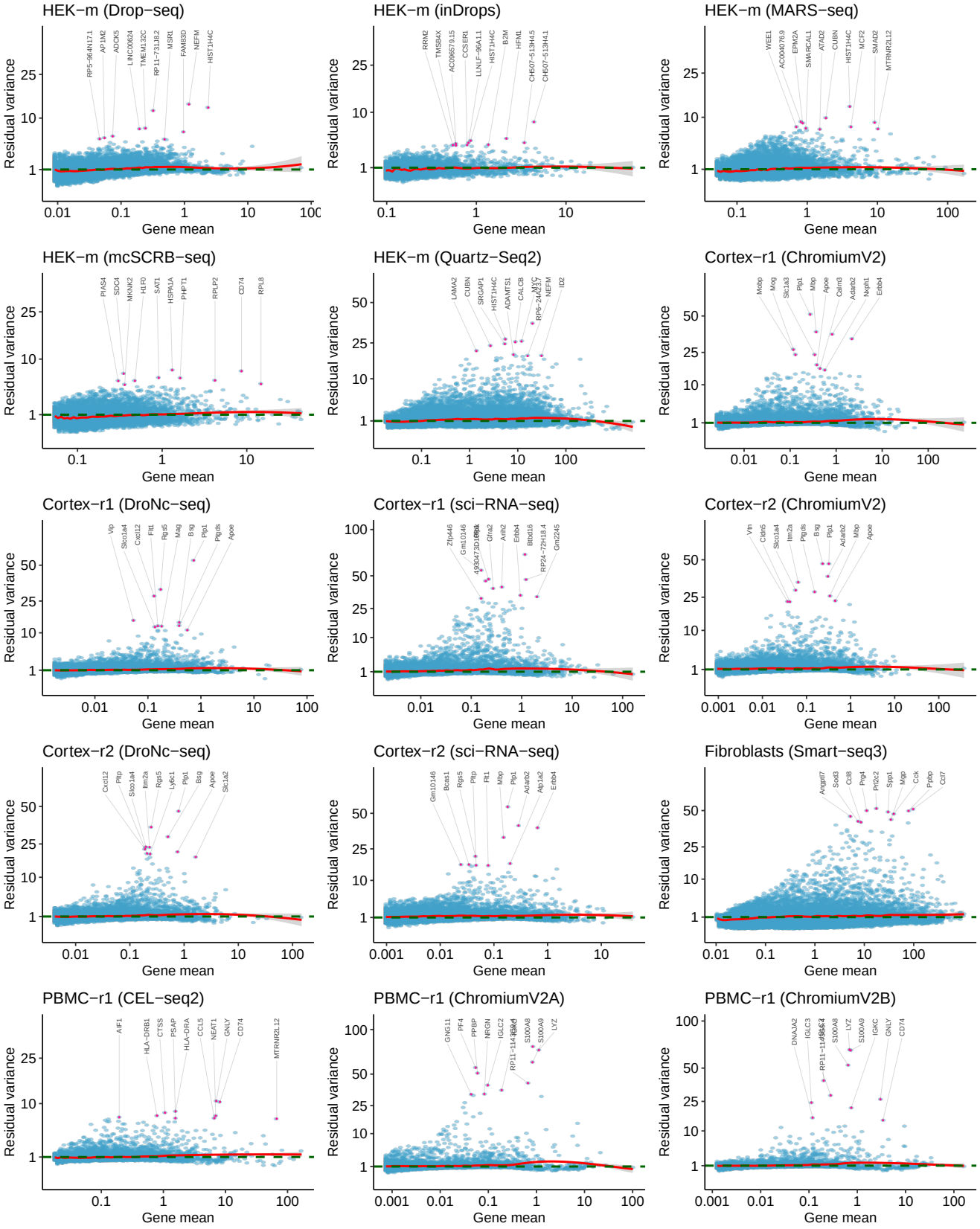

Figure S19. Variance stabilization achieved by sctransform2 across datasets. Y-axis shows variation of pearson residuals as estimated by sctransform (v2 regularization) for datasets 31-45. Top 10 genes with highest residual variances are highlighted.

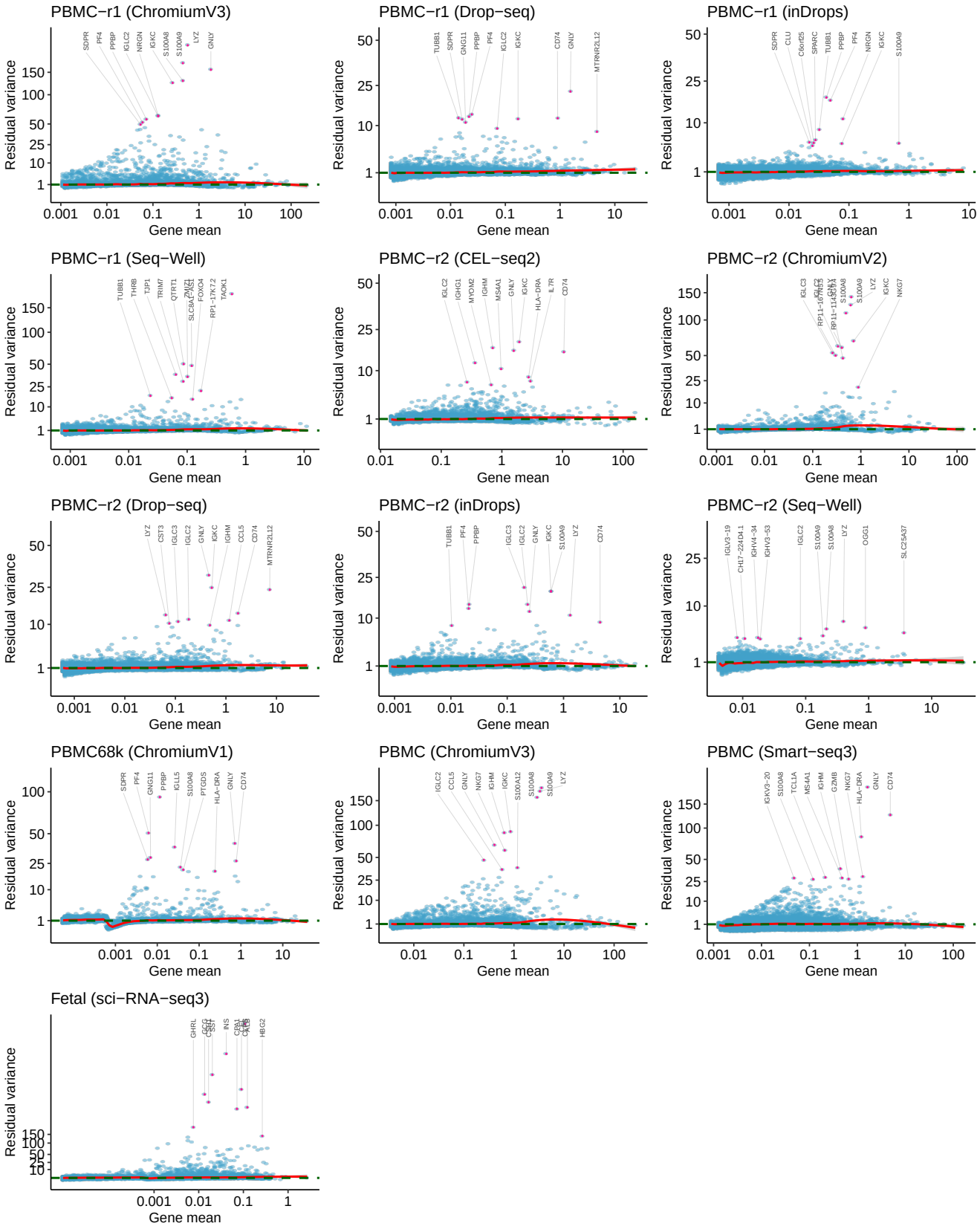

Figure S20. Variance stabilization achieved by sctransform2 across datasets. Y-axis shows variation of pearson residuals as estimated by sctransform (v2 regularization) for datasets 46-58. Top 10 genes with highest residual variances are highlighted.

Sample name	Technology	Tissue	Datatype	Raw data	Reference
TechnicalControl1 (ChromiumV1)	ChromiumV1	TechnicalControl	technical-control	link	(8, 51)
TechnicalControl2 (ChromiumV1)	ChromiumV1	TechnicalControl	technical-control	link	(8, 51)
TechnicalControl (inDrops)	inDrops	TechnicalControl	technical-control	link	(14, 51)
3T3 (ChromiumV3)	ChromiumV3	3T3	cell line	link	(51)
3T3-r1 (ChromiumV2)	ChromiumV2	3T3	cell line	link	(52)
3T3-r1 (Drop-seq)	Drop-seq	3T3	cell line	link	(52)
3T3-r1 (inDrops)	inDrops	3T3	cell line	link	(52)
3T3-r1 (sci-RNA-seq)	sci-RNA-seq	3T3	cell line	link	(52)
3T3-r2 (CEL-seq2)	CEL-seq2	3T3	cell line	link	(52)
3T3-r2 (ChromiumV2)	ChromiumV2	3T3	cell line	link	(52)
3T3-r2 (Drop-seq)	Drop-seq	3T3	cell line	link	(52)
3T3-r2 (inDrops)	inDrops	3T3	cell line	link	(52)
3T3-r2 (sci-RNA-seq)	sci-RNA-seq	3T3	cell line	link	(52)
HEK (ChromiumV3)	ChromiumV3	HEK	cell line	link	(51)
HEK (Smart-seq3)	Smart-seq3	HEK	cell line	link	(29)
HEK-r1 (CEL-seq2)	CEL-seq2	HEK	cell line	link	(52)
HEK-r1 (ChromiumV2)	ChromiumV2	HEK	cell line	link	(52)
HEK-r1 (Drop-seq)	Drop-seq	HEK	cell line	link	(52)
HEK-r1 (inDrops)	inDrops	HEK	cell line	link	(52)
HEK-r1 (sci-RNA-seq)	sci-RNA-seq	HEK	cell line	link	(52)
HEK-r2 (CEL-seq2)	CEL-seq2	HEK	cell line	link	(52)
HEK-r2 (ChromiumV2)	ChromiumV2	HEK	cell line	link	(52)
HEK-r2 (Drop-seq)	Drop-seq	HEK	cell line	link	(52)
HEK-r2 (inDrops)	inDrops	HEK	cell line	link	(52)
HEK-r2 (sci-RNA-seq)	sci-RNA-seq	HEK	cell line	link	(52)
HEK-m (CEL-seq2)	CEL-seq2	HEK	cell line	link	(53)
HEK-m (ChromiumV2)	ChromiumV2	HEK	cell line	link	(53)
HEK-m (ChromiumV2_sn)	ChromiumV2	HEK	cell line	link	(53)
HEK-m (ddSeq)	ddSeq	HEK	cell line	link	(53)
HEK-m (Drop-seq)	Drop-seq	HEK	cell line	link	(53)
HEK-m (inDrops)	inDrops	HEK	cell line	link	(53)
HEK-m (MARS-seq)	MARS-seq	HEK	cell line	link	(53)
HEK-m (mcSCRB-seq)	mcSCRB-seq	HEK	cell line	link	(53)
HEK-m (Quartz-Seq2)	Quartz-Seq2	HEK	cell line	link	(53)
Cortex-r1 (ChromiumV2)	ChromiumV2	Cortex	heterogeneous	link	(52)
Cortex-r1 (DroNc-seq)	DroNc-seq	Cortex	heterogeneous	link	(52)
Cortex-r1 (sci-RNA-seq)	sci-RNA-seq	Cortex	heterogeneous	link	(52)
Cortex-r2 (ChromiumV2)	ChromiumV2	Cortex	heterogeneous	link	(52)
Cortex-r2 (DroNc-seq)	DroNc-seq	Cortex	heterogeneous	link	(52)
Cortex-r2 (sci-RNA-seq)	sci-RNA-seq	Cortex	heterogeneous	link	(52)
Fibroblasts (Smart-seq3)	Smart-seq3	Fibroblasts	heterogeneous	link	(29)
PBMC-r1 (CEL-seq2)	CEL-seq2	PBMC	heterogeneous	link	(52)
PBMC-r1 (ChromiumV2A)	ChromiumV2	PBMC	heterogeneous	link	(52)
PBMC-r1 (ChromiumV2B)	ChromiumV2	PBMC	heterogeneous	link	(52)
PBMC-r1 (ChromiumV3)	ChromiumV3	PBMC	heterogeneous	link	(52)
PBMC-r1 (Drop-seq)	Drop-seq	PBMC	heterogeneous	link	(52)
PBMC-r1 (inDrops)	inDrops	PBMC	heterogeneous	link	(52)
PBMC-r1 (Seq-Well)	Seq-Well	PBMC	heterogeneous	link	(52)
PBMC-r2 (CEL-seq2)	CEL-seq2	PBMC	heterogeneous	link	(52)
PBMC-r2 (ChromiumV2)	ChromiumV2	PBMC	heterogeneous	link	(52)
PBMC-r2 (Drop-seq)	Drop-seq	PBMC	heterogeneous	link	(52)
PBMC-r2 (inDrops)	inDrops	PBMC	heterogeneous	link	(52)
PBMC-r2 (Seq-Well)	Seq-Well	PBMC	heterogeneous	link	(52)
PBMC68k (ChromiumV1)	ChromiumV1	PBMC	heterogeneous	link	
PBMC (ChromiumV3)	ChromiumV3	PBMC	heterogeneous	link	(51)
PBMC (Smart-seq3)	Smart-seq3	PBMC	heterogeneous	link	(29)
Fetal (sci-RNA-seq3)	sci-RNA-seq3	Fetus	heterogeneous	link	(39)

Table S1. List of datasets used in this study. Raw data can be downloaded from the hyperlinks under ‘raw data‘ column. Similar sample names with ‘-r1’ and ‘-r2’ denote replicates from Ding et al.(52) study.

Sample name	<0.01	>0.01	>0.1	>1	>5	>10	>25	>50	>100
TechCtrl1 (ChromiumV1)	0.000			0.01	0.10	0.70			
TechCtrl2 (ChromiumV1)	0.000	0.000	0.018	0.51		0.93			
TechCtrl (inDrops)			0.000	0.01	0.04	0.13	0.41	0.67	0.93
3T3 (ChromiumV3)	0.002	0.007	0.066	0.49	0.99	1.00	1.00	1.00	
3T3-r1 (CEL-seq2)	0.007	0.007	0.118	0.37	0.84	0.98	1.00	1.00	1.00
3T3-r1 (ChromiumV2)	0.004	0.011	0.196	0.82	1.00	1.00	1.00	1.00	
3T3-r1 (Drop-seq)	0.002	0.001	0.054	0.60	1.00	1.00			
3T3-r1 (inDrops)	0.001	0.001	0.046	0.78	1.00	1.00			
3T3-r1 (sci-RNA-seq)	0.001	0.003	0.056	0.50	0.99	1.00	1.00		
3T3-r2 (CEL-seq2)	0.008	0.009	0.104	0.39	0.77	0.95	1.00	1.00	1.00
3T3-r2 (ChromiumV2)	0.008	0.006	0.143	0.74	1.00	1.00	1.00	1.00	
3T3-r2 (Drop-seq)	0.001		0.039	0.47	0.95	1.00			
3T3-r2 (inDrops)	0.001	0.001	0.029	0.48	0.97	1.00			
3T3-r2 (sci-RNA-seq)	0.005	0.008	0.209	0.87	1.00	1.00	1.00		
HEK (ChromiumV3)	0.003	0.006	0.076	0.51	0.97	1.00	1.00	1.00	1.00
HEK (Smart-seq3)	0.012	0.012	0.172	0.70	0.94	0.99	1.00	1.00	1.00
HEK-r1 (CEL-seq2)		0.008	0.092	0.38	0.76	0.94	1.00	1.00	1.00
HEK-r1 (ChromiumV2)	0.005	0.018	0.158	0.85	1.00	1.00	1.00		1.00
HEK-r1 (Drop-seq)	0.000	0.003	0.038	0.47	1.00	1.00	1.00		
HEK-r1 (inDrops)	0.001	0.001	0.020	0.54	1.00	1.00			
HEK-r1 (sci-RNA-seq)	0.005	0.007	0.076	0.52	0.97	1.00			
HEK-r2 (CEL-seq2)	0.005	0.008	0.085	0.45	0.90	0.98	1.00	1.00	1.00
HEK-r2 (ChromiumV2)	0.006	0.019	0.203	0.91	1.00	1.00	1.00	1.00	
HEK-r2 (Drop-seq)	0.001	0.003	0.059	0.62	1.00	1.00	1.00		
HEK-r2 (inDrops)	0.001	0.003	0.028	0.60	1.00	1.00			
HEK-r2 (sci-RNA-seq)	0.006	0.018	0.265	0.91	1.00	1.00			
HEK-m (CEL-seq2)		0.019	0.136	0.61	0.93	0.99	1.00	1.00	1.00
HEK-m (ChromiumV2)	0.188	0.204	0.675	0.97	1.00	1.00	1.00	1.00	1.00
HEK-m (ChromiumV2_sn)			0.005	0.17	0.73				
HEK-m (ddSeq)	0.222	0.479	0.912	0.99	1.00	1.00			
HEK-m (Drop-seq)	0.006	0.014	0.291	0.98	1.00	1.00			
HEK-m (inDrops)			0.008	0.25	0.69	1.00			
HEK-m (MARS-seq)		0.010	0.209	0.80	0.99	1.00	1.00		
HEK-m (mcSCRB-seq)		0.007	0.334	0.94	1.00	1.00			
HEK-m (Quartz-Seq2)	0.103	0.133	0.525	0.96	1.00	1.00	1.00	1.00	1.00
Cortex-r1 (ChromiumV2)	0.001	0.009	0.124	0.77	1.00	1.00			
Cortex-r1 (DroNc-seq)	0.001	0.007	0.200	0.91	1.00				
Cortex-r1 (sci-RNA-seq)	0.001	0.011	0.258	0.97	1.00	1.00			
Cortex-r2 (ChromiumV2)	0.001	0.006	0.104	0.80	1.00	1.00			
Cortex-r2 (DroNc-seq)	0.002	0.007	0.204	0.92	1.00				
Cortex-r2 (sci-RNA-seq)	0.001	0.010	0.218	0.99	1.00				
Fibroblasts (Smart-seq3)	0.048	0.274	0.799	0.99	1.00	1.00	1.00	1.00	1.00
PBMC-r1 (CEL-seq2)	0.001	0.001	0.017	0.31	0.95	1.00			
PBMC-r1 (ChromiumV2A)	0.000	0.004	0.109	0.87	0.98	1.00			
PBMC-r1 (ChromiumV2B)	0.001	0.002	0.079	0.80	0.98	1.00	1.00		
PBMC-r1 (ChromiumV3)	0.002	0.011	0.145	0.78	1.00	1.00	1.00		
PBMC-r1 (Drop-seq)	0.000	0.002	0.128	0.83	1.00				
PBMC-r1 (inDrops)	0.000	0.003	0.189	0.97					
PBMC-r1 (Seq-Well)	0.001	0.005	0.214	0.96					
PBMC-r2 (CEL-seq2)		0.001	0.013	0.30	0.95	1.00			
PBMC-r2 (ChromiumV2)	0.000	0.001	0.061	0.74	0.98	1.00			
PBMC-r2 (Drop-seq)	0.001	0.004	0.206	0.92	1.00	1.00			
PBMC-r2 (inDrops)	0.001	0.005	0.145	0.96	1.00	1.00			
PBMC-r2 (Seq-Well)	0.000		0.095	0.93					
PBMC68k (ChromiumV1)	0.000	0.003	0.069	0.76	0.98	1.00			
PBMC (ChromiumV3)	0.002	0.011	0.111	0.66	0.95	1.00	1.00	1.00	
PBMC (Smart-seq3)	0.031	0.359	0.894	0.97	1.00	1.00	1.00		
Fetal (sci-RNA-seq3)	0.116	0.617	0.990						

Table S2. Proportion of non-poisson genes across different gene-mean bins. Columns indicate non-cumulative gene abundance bins between two consecutive labels (for example, > 1 refers to all genes with mean > 1 and ≤ 5). Each cell entry summarizes the total proportion of genes belonging to a mean abundance bin that were detected to be non-poisson for a dataset.