Analyzing scRNA-seq data with the sctransform and offset models

Christoph Hafemeister and Rahul Satija

We enjoyed reading the recent preprint from Lause, Berens, and Kobak (2020), and appreciate the authors' careful consideration of both our sctransform manuscript (Hafemeister and Satija 2019), and the co-published GLM-PCA paper (Townes et al. 2019). We were pleased to see that many of the conclusions of Lause, Berens, and Kobak (2020) are fully consistent with both of these manuscripts. In particular, the authors explore the use of generalized linear models (GLM) for scRNA-seq normalization. They find that a negative binomial distribution (without zero inflation) is an appropriate statistical error model for analyzing these data. Moreover, they find that the Pearson residuals from this model can be effectively applied for the purposes of identifying highly variable genes, and performing dimensional reduction and clustering. Each of these findings support and affirm the original conclusions of both the Hafemeister and Satija (2019) and Townes et al. (2019) manuscripts. In addition, the authors suggest two modifications to the sctransform model, which we address below.

Comparison of the offset model and sctransform

Our sctransform model learns regularized parameters for negative binomial regression directly from each scRNA-seq dataset. However, Lause, Berens, and Kobak (2020) argue that instead of learning these parameters from the data, it is possible to fix their values to a constant level across datasets, resulting in a less flexible model. They propose an offset model, which is more parsimonious than the sctransform model, and is well-justified under a set of technical assumptions that describe data without biological variability.

We respectfully disagree with the authors' claim that the regularized sctransform model is overspecified, particularly when analyzing data with biological heterogeneity. We agree that their proposed offset model is more parsimonious. However, our goal in sctransform is not to necessarily use the simplest model, but to perform a broadly applicable normalization procedure that focuses downstream analyses on relevant biological variation. Our decision to allow θ to vary flexibly, as a learned function of gene mean, is inspired by methods for differential expression such as DESeq (Anders and Huber 2010) and DESeq2 (Love, Huber, and Anders 2014), which learn regularized models for gene variance directly from each dataset. Moreover, we have found that empirical estimation of the slope and intercept of the GLM model returns parameter estimates that are very similar to those proposed by the offset model, but flexibly allows sctransform to adapt to artifacts and biases which commonly affect scRNA-seq experiments but are extremely challenging to analytically model. These include contamination from ambient RNA and variable levels of mitochondrial contamination (Luecken and Theis 2019), and may violate the technical assumptions used to justify the offset model.

We compared the performance of the sctransform (sct) and offset models on a publicly available scRNA-seq human peripheral blood mononuclear cells (PBMC) generated by 10x Genomics. For each model, we considered the standard deviation of the resulting Pearson residuals for each gene g (denoted here as σ_g), which represents a quantification of the biological variation associated with each gene and determines its weight in downstream analyses. When applying an appropriate statistical model, genes with the highest σ_g should represent markers of heterogeneous cell states, while housekeeping genes, ribosomal proteins, and mitochondrial genes should have reduced σ_g.

Overall, we observed largely concordant transcriptome-wide results (Figure 1A; R=0.880) between $\sigma_{g,sct}$ and $\sigma_{g,\text{offset}}$. However, genes with increased residual variation in the offset model were heavily enriched for ribosomal proteins and other highly expressed housekeeping genes (e.g. RPS27, RPS12, RPL30; ten genes shown in Table 1). Each of these genes is ranked within the top 100 transcriptome-wide based on $\sigma_{g,\text{offset}}$. Therefore, when applying the offset model, these genes will have greater weight in downstream tasks such as

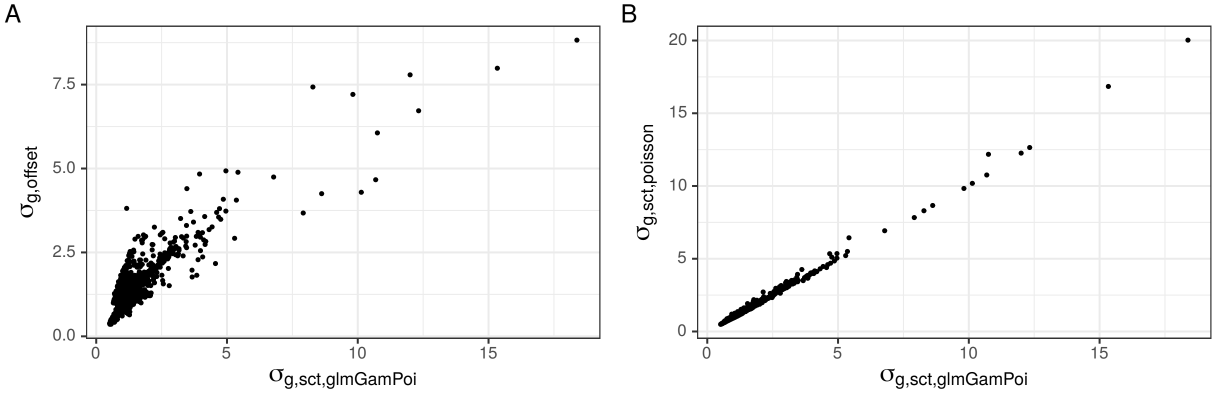

Figure 1: (A) Each point represents the standard deviation of the Pearson residuals for one gene, after applying either the sctransform or offset models. While we generally observe concordance between the two methods, genes with the most substantial differences are highlighted in Table 1. (B) same as in (A), but comparing the standard deviation of Pearson residuals calculated using the sctransform model and two estimation procedures for θ. The results are extremely similar under both approaches.

clustering and visualization. By contrast, genes with increased residual variation in the sctransform model represented canonical markers of distinct human immune cell types (e.g. S100A9, GNLY, LYZ; Table 1). These genes will therefore have an increased role in defining cellular state when using the sctransform model, and we argue that this is consistent with effective normalization and variance stabilization. We observed very similar results when analyzing two additional publicly available 10x datasets from embryonic mouse heart tissue and a Hodgkin's lymphoma tumor (Table 1). We conclude that the offset model does not improve in the normalization and variance stabilization of heterogeneous scRNA-seq datasets when compared to sctransform.

Improved methods for estimating θ

In addition, Lause, Berens, and Kobak (2020) argue that the estimation procedure for θ originally proposed in the sctransform manuscript (denoted as poisson) can result in a bias, particularly for lowly expressed genes. They demonstrate this bias by analyzing negative control datasets of spike-in mRNA that lack biological variation. We agree with these findings, which arise from the inherent challenges in fitting negative binomial GLMs to data from lowly expressed genes, where most cell expression values are 0. We are grateful to the authors for identifying this issue. However, we emphasize that this does not result in a noticeable effect on the outputs of the sctransform procedure. We have previously implemented an alternative estimation procedure (glmGamPoi, as proposed by Ahlmann-Eltze and Huber (2020)) , which alleviates this bias, and also substantially improves the speed of the learning procedure. When we compare $\sigma_{g,sct,poisson}$ with $\sigma_{g,sct,glmGamPoi}$ we observe essentially identical results (Figure 1B; R=0.996). This is because, when considering lowly expressed genes, small fluctuations in θ may not materially influence gene variance, which is defined as $\mu + \mu^2/\theta$. We released the ability to apply glmGamPoi estimation as part of sctransform v0.3 (released on September 9, 2020), and invite users to apply the updated method, and to compare their results with previous versions.

We again thank the authors for a thoughtful and careful discussion of these important issues, and for the opportunity to discuss their findings with them. We note that a Bayesian approach, where the prior distributions for GLM parameter values are set by the offset model, but whose posterior estimates may vary if there is strong evidence in the data, may represent an attractive possibility for future work. All data used here is available for public download, along with code to reproduce the analyses in Figure 1 and Table 1.

PBMC		Heart		Lymphoma	
$\sigma_{g,sct} > \sigma_{g,\text{offset}}$	$\sigma_{g,sct} < \sigma_{g,\text{offset}}$	$\sigma_{g,sct} > \sigma_{g,\text{offset}}$	$\sigma_{g,sct} < \sigma_{g,\text{offset}}$	$\sigma_{g,sct} > \sigma_{g,\text{offset}}$	$\sigma_{g,sct} < \sigma_{g,\text{offset}}$
S100A9	MALAT1	Tnnt2	mt-Co3	IGKC	MALAT1
S100A8	RPS27	Actc1	mt-Atp6	LYZ	FTH1
IGLC2	RPS12	Col1a1	mt-Co2	HLA-DRA	HSP90AA1
IGHM	RPL30	Dcn	mt-Co1	CXCL13	JUND
GNLY	MT-ATP6	Col13a1	Malat1	G0S2	FTL
LYZ	RPS27A	Tnnc1	mt-Cytb	CST3	BTG1
IGLC3	RPL32	Col1a2	Tmsb4x	TIMP1	UBC
MZB1	RPL34	Hbb-bt	mt-Nd2	ALOX15	B2M
IGKC	RPL10	Tnni3	mt-Nd1	EREG	RPS12
JCHAIN	RPL13	Myl3	Hmgb2	IGHA1	RPS27

Table 1: Ranked list of genes whose Pearson residual differs most between the sctransform and offset models. Genes are ordered by the magnitude of difference between $\sigma_{g,sct}$ and $\sigma_{g,\text{offset}}$. After normalization of all three datasets, genes with increased residual variation under the sctransform model are cell type markers, while genes with increased variation under the offset model are heavily enriched for mitochondrial, ribosomal, and highly expressed housekeeping genes.

References

Anders, Simon and Wolfgang Huber (2010). “Differential expression analysis for sequence count data”. In: *Genome biology*, pp. 1–1.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12, p. 550.

Hafemeister, Christoph and Rahul Satija (2019). “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. In: *Genome biology* 20.1, pp. 1–15.

Luecken, Malte D and Fabian J Theis (2019). “Current best practices in single-cell RNA-seq analysis: a tutorial”. In: *Molecular systems biology* 15.6, e8746.

Townes, F William et al. (2019). “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model”. In: *Genome biology* 20.1, pp. 1–16.

Ahlmann-Eltze, Constantin and Wolfgang Huber (2020). “glmGamPoi: Fitting Gamma-Poisson Generalized Linear Models on Single Cell Count Data”. In: *bioRxiv*.

Lause, Jan, Philipp Berens, and Dmitry Kobak (2020). “Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data”. In: *bioRxiv*. DOI: 10.1101/2020.12.01.405886.