Statistical Science
2020, Vol. 35, No. 1, 112–128
https://doi.org/10.1214/19-STS7560

Data Denoising and Post-Denoising Corrections in Single Cell RNA Sequencing

Divyansh Agarwal, Jingshu Wang and Nancy R. Zhang

Abstract. Single cell sequencing technologies are transforming biomedical research. However, due to the inherent nature of the data, single cell RNA sequencing analysis poses new computational and statistical challenges. We begin with a survey of a selection of topics in this field, with a gentle introduction to the biology and a more detailed exploration of the technical noise. We consider in detail the problem of single cell data denoising, sometimes referred to as "imputation" in the relevant literature. We discuss why this is not a typical statistical imputation problem, and review current approaches to this problem. We then explore why the use of denoised values in downstream analyses invites novel statistical insights, and how denoising uncertainty should be accounted for to yield valid statistical inference. The utilization of denoised or imputed matrices in statistical inference is not unique to single cell genomics, and arises in many other fields. We describe the challenges in this type of analysis, discuss some preliminary solutions, and highlight unresolved issues.

Key words and phrases: Single cell biology, RNA sequencing, imputation, post-denoising inference, empirical Bayes, deep learning.

1. INTRODUCTION

1.1 Statistics: The Lens That [Single] Cell Biology Needs

Karl Pearson called statistics "the grammar of science" in the context of how statistical models can provide structure and meaning to physical or biological data (Pearson, 1982). Today's biologists increasingly recognize that statistics is indispensable to their work, and similarly, biology-driven problems continue to inspire statistical thinking. In this paper, we will survey a relatively recent domain in biology, one where instead of analyzing the average signals from many cells through the sequencing of whole tissue, scientists can examine the properties of individual cells. Single cell technologies have become immensely popular over the past five years, making this an exciting time for statistical and computational developments in the field. Through this review, we will first give the reader a general background on single cell RNA sequencing data, and then focus on the problem of data denoising. Data denoising is one way of enhancing the biological signals in single cell sequencing data, but it is also a topic imbued with statistical challenges which we hope to clarify. Beyond single cell sequencing, data denoising and imputation is a common data preprocessing step in many other fields, such as genetics, proteomics, metabolomics and neuroimaging (Hsu et al., 2005, Chiron et al., 2014). Often, denoised/imputed data matrices, rather than the original raw matrices, are used for downstream visualization, parameter estimation and hypothesis testing. We will examine the validity of using denoised matrices for such types of analyses and discuss future directions for statistical research.

1.2 Why Single Cell?

As multicellular organisms, each patch of tissue in our bodies is a heterogeneous community of dissimilar cells, where the cell is the basic unit of structure and function. Comprehensive study of a population requires observation of its multifaceted members. Analogously, comprehensive study of a tissue necessitates nuanced information about its member cells. Until about a decade ago,

Divyansh Agarwal is Ph.D. and M.D. candidate, Graduate Program in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: divyansh.agarwal@pennmedicine.upenn.edu). Jingshu Wang is Assistant Professor, Department of Statistics, The University of Chicago, Chicago, Illinois 60637, USA (e-mail: jingshuw@uchicago.edu). Nancy R. Zhang is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: nzh@wharton.upenn.edu).

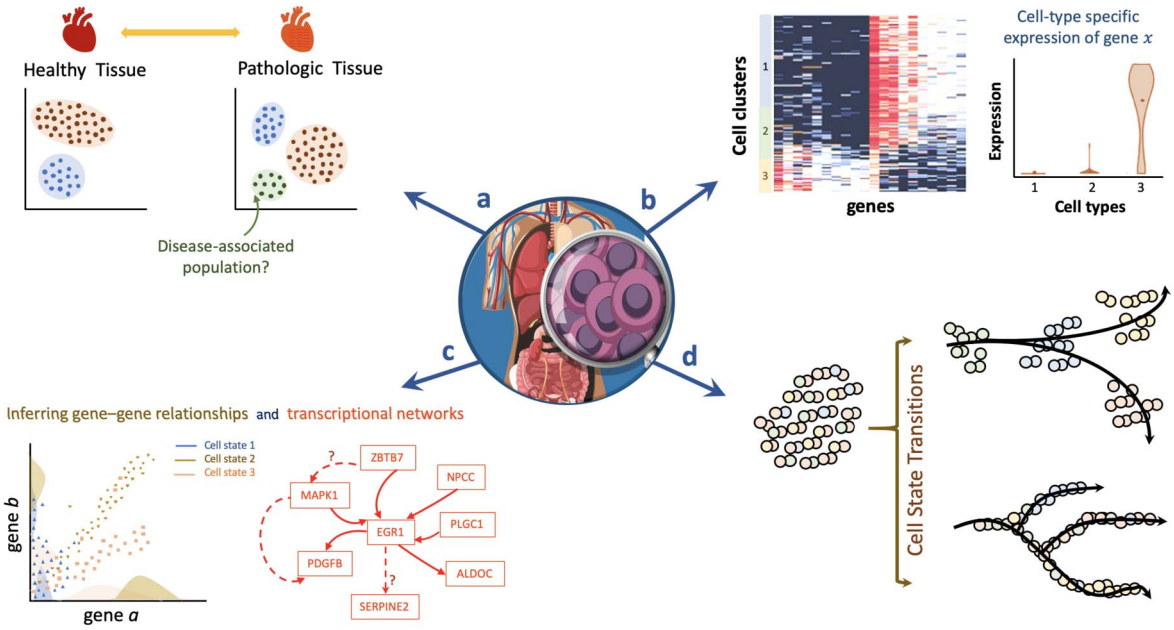

FIG. 1. *Common types of analyses with scRNA-seq data*: (*a*) *cell type identification and detecting disease-associated cell populations*, (*b*) *differential expression analysis and comparison of expression distributions across cell types*, (*c*) *analyses involving gene-gene correlations, network reconstruction, and* (*d*) *trajectory inference.*

biologists relied on high throughput assays, such as microarrays or bulk sequencing, that required the pooling of materials (e.g., RNA, DNA) over a large number of cells. Thus, it was very difficult to study the heterogeneity of cells within a tissue. Especially, if the dynamics of component populations are offset in time, or if a cell type of interest is rare, much information is lost in bulk tissue experiments (Raj and van Oudenaarden, 2008, Gossett et al., 2012).

The instructions needed to build and maintain cells are encoded in its DNA, and cells carry out these instructions by reading and transcribing DNA into RNA. The RNA readouts for a given gene are called transcripts, and the collection of all gene readouts, the transcriptome. Although recent technological advances have made possible the high throughput single-cell profiling of many types of features, such as DNA copy number, chromatin accessibility, methylation, RNA and surface level proteins (see the review by Stuart and Satija (2019)), this paper will focus on single cell RNA sequencing (scRNA-seq). In scRNA-seq, the entire transcriptome is profiled for each cell, across a large number of cells simultaneously. This technology has allowed the comprehensive cataloging of the different cell types that constitute organs, leading to the discovery of new subpopulations of cells that were previously hidden by bulk sequencing analyses. Through scRNA-seq, we are gaining a deeper understanding of what constitutes specific cell types, how cell types function, and how they may change during disease.

Single cell sequencing has spurred rapid methods development to address its manifold challenges, reviewed recently in Zappia, Phipson and Oshlack (2018) and Hwang, Lee and Bang (2018). In this review, we focus on an inherent limitation of scRNA-seq: only a small fraction of the transcripts present in each cell are sequenced, leading to unreliable gene quantification that hinders downstream analysis. Denoising scRNA-seq data and imputing the missing transcripts can be an effective pre-processing step. However, while accurate denoising can enhance downstream visualization and analysis, it can also introduce biases and spurious correlations between genes and cells. We will take an in-depth look at the problem of denoising scRNA-seq data, and, in particular, examine the validity of downstream statistical estimation and testing procedures based on the denoised values. We envision that the lessons learned here may also apply to denoising/imputation efforts in other fields.

1.3 Outline

We start, in Section 2, with a gentle introduction to single cell biology and a brief survey of some of the typical questions that drive single cell studies (Figure 1). Since there has been rapid progress in this area, we will point the reader to other review papers when available. In Section 3, we will discuss the technical noise in single cell experiments, focusing on experiments that utilize a barcoding strategy called unique molecular identifier (UMI, more on this in Section 3.2). In Section 4, we will turn to the specific topic of scRNA-seq denoising. We discuss why this is not a typical statistical imputation problem, canvass current strategies, and highlight common pitfalls. Data denoising uses correlations detected in the data matrix to reduce the noise in each matrix entry, and there

is a concern that this introduces bias and spurious correlations. In Section 5, we examine the issue of bias and explore strategies for bias-adjustment.

2. A BRIEF SKETCH OF SINGLE CELL BIOLOGY

2.1 What's in a Cell Type, Anyway?

The multimillion dollar global initiative, Human Cell Atlas, will soon "create comprehensive reference maps of all human cells to give us a unique ID card for each cell type" (Regev et al., 2017). The Human Cell Atlas begins in the rather philosophical quest of defining a cell type. Canonical cell types—think of muscle (myocyte), nerve (neuron) or fat (adipocyte)—were originally defined by both the functions of the tissues in which they reside *and* histological cell classifications based on morphology, often based on imaging of stained cells mounted on a microscope slide. To reconsider the definition of a cell type, let us revisit what a cell is. On the one hand, it is a collection of "stuff": mainly DNA, RNA, proteins and metabolites enclosed by a cell wall. On the other, a cell is a dynamic entity that operates in an ecological niche: in response to its environment, it can adapt and modify. With technologies such as scRNA-seq, one can begin to quantify both the fixed and dynamic states of single cells, and broaden the existing catalogue of cell types. Are existing classifications appropriate, or should cell types be defined to encompass the dynamic landscapes on which cells reside? This question, which marks the beginning of a paradigm shift on how cell types should be defined by the scientific community, is an open invitation for statistical ideas.

Although the transcriptome of single cells can be used to ascribe cell type identity, there is considerable cell-to-cell variation in expression within a given cell type. This variation can reflect the inherent stochasticity of RNA transcription and degradation at the single cell level, as well as differences in micro-environment, in cell cycle stage, and in other *latent* factors. Furthermore cells often don't fall into discrete classes, and can transition dynamically between similar types (Clevers et al., 2017) or have multiple ID cards! Just as there is no singular, all-encompassing way of classifying people, maybe there is none for cells.

And as with people, history and context matters for cells. How is a cell related to another cell and from where did it arise? What molecular events or external stimuli influenced its transcriptome? What identity might a cell assume next, given its current environment and its history? The concept of a cell type becomes particularly important if we think of them not just as static molecular snapshots, but as histories unfolding in time (Trapnell, 2015). Each cell in the human body has its origins in a single fertilized oocyte (female egg cell). Cells make decisions as they divide along their developmental journey; some decisions are definitive, and others are more flexible with potential for reversal. For understanding both normal development and disease, it is imperative to have a fundamental grasp of the cell-fate transitions that occur in complex cellular ecosystems. Although scRNA-seq captures only a static snapshot of each cell's transcriptome, ingenious computational methods have been developed to infer the dynamic context. For example, there are many methods that use graphical representations in low dimensional projections to reconstruct developmental trajectories. Others have adapted optimal transport methods to longitudinal cell sampling designs to infer population dynamics. The work of La Manno et al. (2018) is particularly noteworthy, which exploits the fact that freshly transcribed mRNA is unspliced and thus uses the ratio of unspliced mRNA to spliced mRNA to deduce the future transcriptomic state of cells. Recently, (Saelens et al., 2019) reviewed more than 70 software tools for constructing cellular trajectories from scRNA-seq data, highlighting the interest on this type of analysis.

2.2 Biological "Noise," Does It Matter?

Cell types and cell trajectories connote macroscopic changes in a cell's phenotype, often realized through broad, sweeping changes in its transcriptome. Zooming into cells of the same type, or cells at the same point in a differentiation trajectory, how do we expect gene expression to vary at this level? The stochastic variation in RNA count between cells of the same overt type is often referred to as "gene expression noise." Studies of gene expression noise and its functional ramifications date back to the early 1950s, when Novick and Weiner (Novick and Weiner, 1957) showed that the production of beta-galactosidase in bacteria grown in a homogeneous environment is random and highly variable across genetically identical cells. Since then, many studies, mostly conducted on yeast and bacteria, have demonstrated the generality and relevance of biological noise. The impact of noise on biological pathways has also been explored by computational models (McAdams and Arkin, 1997, Arkin, Ross and McAdams, 1998). For unicellular organisms, gene expression noise is now understood as a mechanism through which populations hedge bets in an unpredictable environment (Raj and van Oudenaarden, 2008).

What role does transcriptional noise play in multicellular organisms, where cells are expected to act in concert to maintain the fitness of the whole? In fact, stochastic variations in gene expression play key roles in tissue development and maintenance (Losick and Desplan, 2008). In development, the inherent stochasticity of early cellular events allows for the diversification needed for cell type differentiation (Eldar and Elowitz, 2010). Such stochasticity may also affect disease progression and cellular response to treatment. Preliminary evidence has emerged

from scRNA-seq studies that show an increase in biological noise with aging (Enge et al., 2017, Song, Sarnoski and Acar, 2018). Interestingly, in some situations, variation between cells can drive inter-cellular competition within a seemingly homogeneous tissue niche, and such competition can be fundamental to the health of the tissue (Di Gregorio, Bowling and Rodriguez, 2016).

Single cell technologies provide an unprecedented opportunity to study gene expression noise and characterize its functional roles. However, what seems missing is a clear formulation of biological noise that can be estimated from sequencing data. Recall that "noise" refers to variation between cells of the same type. However, in multicellular organisms, "cell type" is now appreciated as a somewhat fluid concept, where detailed analysis can often partition a previously presumed homogeneous cell type into finer subtypes. If gene expression noise is estimated based on a clustering of the data, how should the resolution be chosen? What types of inter-cellular expression variation can be attributed to noise, rather than to latent macroscopic variables? We found an insightful description by Elowitz et al. (2002), who highlighted two mechanisms that underlie noise. The expression of a gene can vary between similar cells due to differences in factors such as cell size, cell location within tissue, and fluctuations in the expression of upstream genes. Variation due to such global, environmental factors are referred to as "extrinsic noise." In contrast, Elowitz et al. (2002) defined "intrinsic noise" to be noise due to "stochasticity inherent in the biochemical process of gene expression," that is, noise due to the stochasticity in *cis*-regulatory binding, transcription, and RNA degradation. Borrowing from these definitions, it seems natural to think of extrinsic noise as latent factors that induce correlations between genes and cells, and intrinsic noise as fluctuations that are independent across genes and cells. In Section 4.2, we will adopt this framework to formulating the SAVER-X denoising model.

3. TECHNICAL NOISE IN SINGLE CELL RNA SEQUENCING

Although protocols for RNA sequencing in individual cells were first described in the 1990s (Van Gelder et al., 1990, Eberwine et al., 1992), it was not until 2009 that the entire transcriptome was quantified for six individual cells (Tang et al., 2009). Technology has evolved during the past ten years to allow current studies to easily scale up to thousands of cells per run (Svensson, Vento-Tormo and Teichmann, 2018). In this parallelization across cells, usually only a small fraction of the RNA molecules in each cell are sequenced and thus counted. Due to this low per-cell coverage and other technical issues, scRNA-seq data is much noisier than bulk RNA sequencing data. The technical noise of scRNA-seq has been extensively studied (see reviews in Kolodziejczyk et al. (2015), Ziegenhain et al., 2017), here we will provide a brief overview and describe a Poisson model for UMI-based data.

3.1 Sources of the Technical Noise

There are many scRNA-seq protocols (Papalexi and Satija, 2018, Hedlund and Deng, 2018), which differ in how the cells are dissociated and isolated into individual compartments for processing, as well as how the RNA is barcoded and amplified. For simplicity, we focus on error propagation in the steps that follow RNA extraction, shown in Figure 2.

First, the reverse transcription (RT) step, which converts RNA (X_{gc} for gene g and cell c) into cDNA (W_{gc}), has a limited efficiency where only a fraction of the RNA molecules in the cell are successfully reverse transcribed. This is called a "dropout" event in scRNA-seq. The efficiency of this step can be both gene- and cell- specific. The cDNA molecules are given a cell-specific barcode and pooled across cells. Then, each cDNA molecule is amplified into multiple copies ($\tilde{W}_{gc}$), accumulating amplification noise which are exponential in magnitude and difficult to model (Degrelle et al., 2008, Parekh et al., 2016). Most protocols use a barcoding strategy, unique molecular identifiers (UMI) (Islam et al., 2014), where each cDNA molecule is tagged with a unique barcode before amplification, and thus amplified copies from the same cDNA molecule all share the same barcode. Then, reads for the same barcode can be collapsed and we can simply count the number of unique barcodes, instead of the number of reads, mapping to each gene.

In the final sequencing step, a random sample (Y_{gc}) of the amplified (and barcoded) cDNA library is read by the sequencer, and the reads are mapped to an annotated genome template. This step is the same as for bulk sequencing, and a simple binomial sampling model suffices. The average per cell coverage, that is, the average number of reads sequenced per cell, is a parameter that can be approximately controlled. The higher the coverage, the lower the sampling noise.

As in bulk RNA sequencing, external RNA controls, such as ERCC spike-in mixtures (Brennecke et al., 2013), have been used to characterize technical noise. ERCC spike-ins are comprised of distinct synthetic RNAs, at varying concentrations, which are added into each cell at known dilution ratios. As the expected concentration of these spike-ins is known and does not vary across cells, we can use their measured reads/UMI counts to characterize measurement error. However, researchers have shown that the technical bias for spike-ins can be quite different from that for real genes (Tung et al., 2017), which has limited their practical utility. Additionally, it is not always possible to add spike-ins to each cell, especially in microfluidic and droplet-based protocols.

3.2 Models for the Technical Noise

Compared to bulk RNA-seq data, scRNA-seq data is much more sparse and highly dispersed (Svensson et al.,

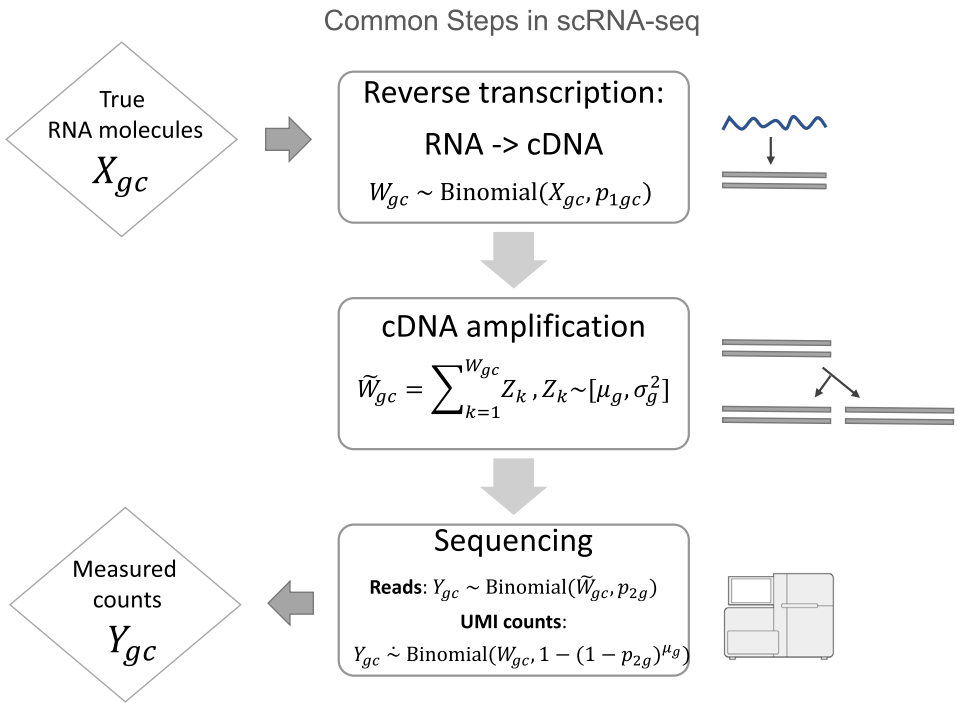

FIG. 2. *Propagation of measurement error during scRNA-seq a experiment. In the notation, subscript c indexes cells and g indexes genes. See the Supplementary Material* (Agarwal, Wang and Zhang, 2020) *for details.*

2017). It is difficult to quantify technical noise in scRNA-seq because it is impossible to have true technical replicates. We cannot sequence a cell twice, and comparisons between cells are confounded by true biological differences. Early studies performed variance decomposition to calculate how much of the between cell variance can be attributed to technical noise (Klein et al., 2015, Kim et al., 2015). Such calculations, based only on moment equations, avoid explicit assumptions about the noise distribution.

Yet, assumptions about the technical noise is needed to denoise the data and estimate the underlying gene expression values. We focus here on scRNA-seq that utilize UMIs, and refer the reader to the Supplementary Material (Agarwal, Wang and Zhang, 2020) for modeling the measurement error in general. Assume that the true number of RNA copies of gene g in cell c is X_{gc} and the observed UMI count is Y_{gc}. UMI allows us to ignore the PCR amplification variation ($\sigma_g^2 = o(\mu_g)$ in Figure 2), and Kim et al. (2015) derived a Poisson model,

$$Y_{gc}|X_{gc} \sim \text{Binomial}(X_{gc}, \alpha_{gc}),$$

where α_{gc} is the overall efficiency.

Since α_{gc} is typically small (less than 10%), we further get

$$Y_{gc}|X_{gc} \sim \text{Poisson}(\alpha_{gc} X_{gc}). \tag{3.1}$$

Through a deconvolution approach, Wang et al. (2018) found extensive empirical evidence for this Poisson model. Compared with noise models for read counts in experiments without UMI, an important advantage is that we do not need an extra zero inflation term for dropouts, and all technical zeroes are modeled by Poisson-based downsampling (Wang et al., 2018). Evidence of overdispersion for UMI-based data has motivated more sophisticated models (Hafemeister and Satija, 2019), yet we found that for most genes, the Poisson-alpha model gives an adequate approximation, and its computational attractiveness motivates its use.

As with other high-throughput genomics experiments, scRNA-seq data is plagued by batch effects (Hicks et al., 2018). Batch adjustment is difficult in scRNA-seq as batch can be confounded with important features such as cell type. Recently, many methods (Haghverdi et al., 2018, Butler et al., 2018, Stuart and Satija, 2019) were proposed to remove batch effects (also called "data alignment" in the literature). However, given the complexity of the issue, more statistical treatment is needed in this direction. In simple scenarios, the efficiency α_{gc} in (3.1) can also be modeled as being linearly depending on known covariates, such as batches. We refer the reader to Wang et al. (2018) for a more detailed discussion.

4. DENOISING SINGLE CELL DATA

4.1 A Review of Current Approaches

Single cell sequencing gives a patchy picture of gene expression wherein most (usually $> 90\%$ and sometimes

up to 99%) of the entries in the cell-by-gene count matrix are zeros. This motivates the question of whether we could "denoise," that is, recover the original RNA counts for each gene in each cell. This has sometimes also been referred to as "imputation," although it differs from the classical imputation setup in important ways. Many of the zeros in scRNA-seq data are true zeros due to lack of expression. For genes that are expressed, the likelihood of a technical zero, either due to dropout during reverse transcription or due to sampling during sequencing, decreases with its expression level. Thus, not all zeros are due to dropout, and those that are missing are not missing at random.

In data denoising, we not only want to recover the genes that are missing (a.k.a. the technical zeros), but also improve the estimates for all genes. All methods for scRNA-seq denoising work by borrowing information across related genes or similar cells. A large category of methods, for example, estimate the expression of each gene in each cell by smoothing across "neighboring" cells that are proximal in some lower-dimensional geometrical representation, such as a manifold. Examples of such methods include kNN-smoothing (Wagner, Yan and Yanai, 2017), scImpute (Li and Li, 2018), VIPER (Chen and Zhou, 2018), DrImpute (Gong et al., 2018) and Markov Affinity-based Graph Imputation of Cells (MAGIC) (Van Dijk et al., 2018).

Alternatively, one can also exploit gene-gene correlations. Such methods, which include SAVER (Huang et al., 2018), DCA (Eraslan et al., 2019), scVI (Lopez et al., 2018), ALRA (Linderman, Zhao and Kluger, 2018), attempt to improve the estimates for each gene using the observed counts for related genes in the same cell. In such gene-level models, it is natural to leverage public data sets to more accurately estimate gene-gene relationships. Transfer learning from public datasets is particularly enticing in light of recent initiatives to build detailed cellular atlases for each anatomic organ in mouse (Han et al., 2018, The Tabula Muris Consortium, 2018) and human (Rozenblatt-Rosen et al., 2017). Recent methods that look to transfer information across scRNA-seq datasets include SAVER-X (Wang et al., 2019) and TRANSLATE (Badsha et al., 2018); both use an autoencoder, a neural network that compresses the input into a latent-space representation, and then reconstructs the output from this representation.

In general, denoising can be a double-edged sword. It can be tempting, during denoising, to tune algorithms to introduce structures into the data that the eye wishes to see. We believe that denoising methods should be evaluated under the null—does it introduce correlations that are not real? In particular, methods that learn from external data should not introduce bias, or force the new data to conform to patterns that only exist in the external data. Ultimately, denoising should increase the reproducibility of discoveries across replicate experiments and varying protocols.

Since the truth is unknown in almost all scRNAseq datasets, benchmarking methods is difficult. Several studies have performed evaluations that assess methods in their ability to: (i) recover gene-gene correlations, (ii) enhance the visualization of distinct cell clusters, and (iii) improve other downstream analyses such as differential expression. For example Tian et al. (2019) created "gold standard" benchmarking data sets by mixing distinct cell populations at known ratios. Comparison of twelve normalization and denoising methods on these data showed substantial differences between methods in their ability to minimize bias (introducing false signals in the data). Another comparison of denoising methods on negative binomial simulations and data from the Mouse Cell Atlas stressed that most methods, except SAVER (Huang et al., 2018), tend to introduce spurious correlations between genes (Andrews and Hemberg, 2018). Moreover, detection of differentially expressed genes in denoised data can have type-1 error inflation, depending on which denoising approach is employed (Zhang and Zhang, 2018). These studies motivate our inquiry in Section 5, where we investigate the effects of existing denoising methods on downstream analyses, and explore remedies to bias and type-1 error inflation.

4.2 The Single-Cell Analysis via Expression Recovery (SAVER) Model

In Section 5, we will investigate the issue of bias across several denoising methods and explore how to obtain unbiased estimation under the framework of the SAVER model we proposed in (Huang et al., 2018). Here we will review the SAVER model in more depth, which should further clarify the concepts of technical versus biological variation, and of intrinsic versus extrinsic noise.

In denoising scRNA-seq data, the quantity we would like to recover is well defined: within each cell c, each gene g had a true realized expression level X_{gc} when the cell's RNA molecules were extracted. Then, each step in the experiment introduces technical noise, eventually resulting in the observed expression count Y_{gc}. We seek to recover the true, unobserved value X_{gc}. Focusing on UMI-based data, let Y_{gc} be the count of the number of unique UMI barcodes for gene g in cell c. To recover the true expression levels we need to distinguish biological variation from technical noise in the matrix Y, and thus require careful choice of a technical noise model. In SAVER and SAVER-X, we assume the technical noise to follow (3.1). To quantify the biological variation between cells, consider a general framework where the true gene expression X_{gc} is derived by adding independent stochastic noise to an underlying correlated component Λ_{gc}:

$$X_{gc}|\Lambda_{gc} \overset{\text{indep}}{\sim} F(\Lambda_{gc}, \varphi_g \Lambda_{gc}), \tag{4.1}$$

where F is an arbitrary distribution with mean Λ_{gc} and variance $\varphi_g \Lambda_{gc}$. The independence in (4.1) is across both genes and cells. Λ_{gc} can be interpreted as the portion of gene g's expression that is predictable given the expression of other genes. As discussed in Section 2.2, at the level of single cells, gene expression can be idiosyncratic and unpredictable due to intrinsic noise. This motivates the independent deviations from the "predictable" component Λ_{gc}. Thus, one can interpret F to be the distribution for the intrinsic noise. An alternative argument for the conditional independence of X_{gc} given Λ_{gc} is that this is a natural consequence of the model's construction, that is, we assume that all of the dependence between genes has already been absorbed into Λ. So far the only assumption on F is that the expected magnitude of the deviation of X_{gc} from Λ_{gc} is controlled by a single gene-specific dispersion parameter φ_g. This is, of course, a simplification, as it would not be surprising for the magnitude of intrinsic noise to, say, vary across cell states. One could envision a more complex model for ϕ_g, but for now we abide by this simple assumption.

We have yet to specify the model for the correlated component Λ, on which we need restrictions to allow identifiability. In the first version of this model (Huang et al., 2018), we assumed that each gene can be predicted with only a small set of other genes (sparsity). Then, in Wang et al. (2019), SAVER-X assumes that Λ is low-rank and smooth, that is, the points $\mathbf{\Lambda}_c = (\Lambda_{gc} : g = 1, \ldots, G)$ lie on a manifold.

As we described in Section 2, currently the most common types of scRNA-seq analyses are the detection cell types and continuous cell trajectories, followed by the identification of genes that show differential expression patterns across types or along trajectories. Since cell types and continuous cell trajectories are characterized by the concerted up- and down-regulation of *groups* of genes, we expect these features to be fully captured by the correlated component Λ. In contrast, if one were interested in characterizing intrinsic transcriptional noise within a cell type (Enge et al., 2017, Martinez-Jimenez et al., 2017, Barroso, Puzovic and Dutheil, 2018), or in estimating the entropy of cells (Teschendorff and Enver, 2017), one would need to look at X and its deviation from Λ. Although denoising methods may have a propensity to over-smooth the data to yield an output that is close to Λ, the SAVER and SAVER-X models are unique in that they try to differentiate between intrinsic biological noise and technical noise.

5. POST-DENOISING ANALYSIS AND INFERENCE

A naive plug-in approach after denoising is to treat the denoised/imputed matrix $\widehat{X}$ as the unknown true X and use it for downstream analyses. However, this ignores estimation uncertainty. For which types of downstream analyses would this be problematic? As mentioned in Section 4.1, some benchmark studies have found data denoising to introduce bias and spurious correlations. In this section, we discuss denoising bias, bias adjustment, and post-denoising inference based on the SAVER model, with a focus on two types of downstream analyses: (1) estimation of functions of X (measures of gene dispersion, gene-gene correlations, cell-cell distances, and cell clustering labels) and (2) hypothesis testing (differential expression analysis).

5.1 Estimating Functions of X

A wide range of applications require reliable estimation of functions of X. For instance, we may want to quantify the biological variation of a gene g across cells. Or, we may want to quantify functions of X involving two or more genes, for example to compute pairwise gene-gene correlations or cell-to-cell distances for clustering or trajectory analysis.

Let $f(X)$ be the function of interest. Directly plugging in $\widehat{X}$ to estimate $f(X)$, which is relatively straightforward and widely used in practice, ignores any uncertainty in the estimation of $\widehat{X}$. Since the SAVER framework also gives the posterior distribution of X, we may get improved estimates of $f(X)$ by its posterior mean:

$$E\big[f(X) \mid Y, \Lambda\big]. \tag{5.1}$$

Since Λ is not known, we would ideally also like to incorporate the uncertainty in the estimation of Λ as well. However, we have found this to be very difficult and, as yet, have not found strategies that are computationally attractive. Thus, we propose to use the plug-in estimate

$$E\big[f(X)|Y, \Lambda = \hat{\Lambda}\big]. \tag{5.2}$$

Below, we will examine how well this strategy works.

5.1.1 *Variance, correlation, and other simple functions.* There are two strategies one could employ to estimate (5.2). For simple functions of X, such as the variance of a gene, or the Pearson correlation between two genes, we can directly derive an analytical formula of $E[f(X) \mid Y, \Lambda]$.

For example, let $f(X) = V_g(X) = \frac{1}{C}\sum_{c=1}^{C}(X_{gc} - \bar{X}_{g\cdot})^2$ be the true biological variance of gene g. The denoised matrix $\hat{X}$, given by SAVER, is an estimate of $E[X|Y, \hat{\Lambda}]$. For gene g cell c let $v_{gc} = \mathrm{Var}[X|Y, \Lambda]$ be the posterior variance, and let $\hat{v}_{gc}$ be the estimated posterior variance computed by SAVER, the latter assuming that Λ is known and equal to $\hat{\Lambda}$. Then, it is easy to show (see the Supplementary Material (Agarwal, Wang and Zhang, 2020)) that

$$\begin{aligned} &E\big[V_g(X) \mid Y, \Lambda\big] \\ &\quad\approx \frac{1}{C}\left[\sum_{c=1}^{C}(\widehat{X}_{gc} - \bar{\widehat{X}}_{g\cdot})^2 + \sum_{c=1}^{C}\widehat{v}_{gc}\right]. \end{aligned} \tag{5.3}$$

If we were to simply use the variance of the gene computed from the denoised values, we would underestimate by approximately the amount that is the sum of the variances v_{gc} across cells, ignoring the uncertainty in the estimate of Λ. However, with SAVER's estimates of v_{gc}, we can attempt to correct for this bias by using (5.3) instead.

Next, consider Pearson's correlation between two genes g_1 and g_2, denote

$$C_{g_1g_2}(X) = \frac{\frac{1}{C}\sum_{c=1}^{C}(X_{g_1c} - \bar{X}_{g_1\cdot})(X_{g_2c} - \bar{X}_{g_2\cdot})}{\sqrt{V_{g_1}(X)}\sqrt{V_{g_2}(X)}}.$$

When the number of cells C is sufficiently large,

$$\begin{aligned} &E\big[C_{g_1g_2}(X) \mid Y, \Lambda\big] \\ &\quad\approx \frac{\frac{1}{C}\sum_{c=1}^{C}(\hat{X}_{g_1c} - \bar{\hat{X}}_{g_1\cdot})(\hat{X}_{g_2c} - \bar{\hat{X}}_{g_2\cdot})}{E[\sqrt{V_{g_1}(X)} \mid Y, \Lambda]E[\sqrt{V_{g_2}(X)} \mid Y, \Lambda]}. \end{aligned} \tag{5.4}$$

To estimate the denominator, we will simply plug-in the square-root of (5.3) for genes g_1 and g_2.

For more complicated functions $f(X)$ where an analytic formula is harder to derive, one could sample from the estimated posterior distribution of X to get at (5.1). We will show in simulations the performance of the sampling strategy when $f(X)$ is the Spearman correlation or a cell-to-cell distance function.

Following the strategy of Huang et al. (2018), we perform down-sampling simulations on four datasets—(Baron et al., 2016, Chen et al., 2017, La Manno et al., 2016, Zeisel et al., 2015). For each dataset, we select a subset of cells and highly expressed genes, and treat them as the underlying true X. Then we down-sample the counts in each entry and obtain a simulated Y following (3.1), where the average α is 10% for the first three datasets and 5% for the last dataset. We then denoise Y and compare the denoised $\hat{X}$ with X.

We compare four recent scRNA-seq denoising methods: SAVER-X (without transfer learning), DCA (Eraslan et al., 2019), scVI (Lopez et al., 2018) and ALRA (Linderman, Zhao and Kluger, 2018). DCA and scVI use an autoencoder while ALRA uses a thresholded SVD to approximate the data matrix.

First, we evaluate the estimation of three gene-level functions: (1) coefficient of variation (CV) measuring gene expression dispersion, (2) gene-gene Pearson correlation, and (3) the gene-gene Spearman correlation. For SAVER-X, We compare the estimates obtained by directly using the denoised values with those obtained using sampling-based bias correction. For CV and Pearson correlation, since analytic formulas (5.3), (5.4) are available, we also include these analytic corrections in the comparisons.

As shown in Figure 3, the down-sampled observed counts Y always overestimate the dispersion and underestimate gene-gene correlations due to the introduction of technical noise. In contrast, directly using the denoised values tends to underestimate the dispersion and overestimate gene-gene correlations. For SAVER/SAVER-X, we found that the estimation bias can be effectively removed by incorporating the uncertainty in $\hat{X}$ through the proposed analytical and sampling strategies that use $\hat{\Lambda}$ as a plug-in for Λ in the approximation of the posterior. Since other denoising methods face the issue of over-smoothing as well, we surmise that taking the posterior randomness of X into consideration can provide a general strategy for ameliorating such biases.

Note that this down-sampling simulation is not perfect. Our "true" X is a filtered, high-quality matrix of observed UMI counts. Here, the gene-gene sample correlations are close to 0, and as a consequence, the underestimation of gene-gene correlations using Y is not apparent in some simulation datasets. In other datasets, however, the actual underlying true gene-gene correlation would be higher, and thus the denoising methods would easily facilitate the recovery of such correlation patterns.

5.1.2 *Dimension reduction and visualization.* Visualization is an integral part of exploratory data analysis, and single cell data is often first scrutinized by eye. Dimension reduction tools, such as PCA, tSNE and UMAP, are part of every scRNA-seq analysis pipeline. Such methods start with a cell-to-cell similarity or distance matrix, and find lower dimension projections that preserve the distances. Treating this distance matrix as a function on X, does denoising improve its estimation and enhance visualization? What role does denoising uncertainty play? When computing the cell-to-cell distance (d), Euclidean distance is a commonly used metric, although other measures have also been proposed. For a detailed overview and comparison of the similarity measures, we encourage the reader to see Skinnider, Squair and Foster (2019), Kim et al. (2019).

For simplicity, we will consider how the use of denoised expression estimates affects d, and in turn, the subsequent visualization of cell clusters. For two cells, c and c', the Euclidean distance between their true transcriptome profiles is

$$d^2(c, c') = \sum_g (X_{gc} - X_{gc'})^2.$$

Using the notations from the last section, we condition on the observed data Y and plug in $\hat{\Lambda}$ for Λ to get

$$\begin{aligned} &E\big[d^2(c, c')|Y, \Lambda\big] \\ &\quad\approx \sum_g (\hat{X}_{gc} - \hat{X}_{gc'})^2 + \sum_g \widehat{v}_{gc} + \sum_g \widehat{v}_{gc'}. \end{aligned} \tag{5.5}$$

Please also see the Supplementary Material (Agarwal, Wang and Zhang, 2020) for a more lengthy discussion. (5.5) is fast to compute and simple to interpret: The second and third terms are the sums of the posterior variances across genes for cells c and c', respectively, and can be

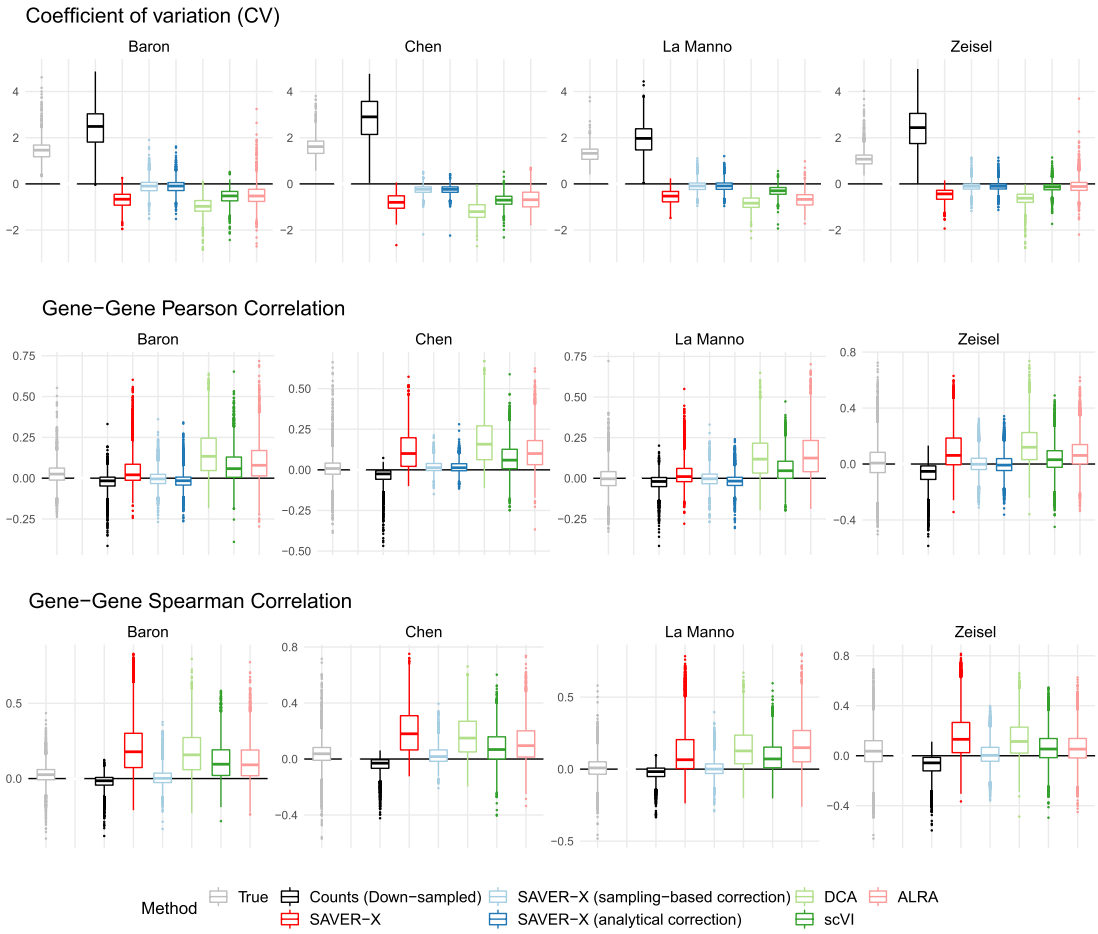

FIG. 3. *Effectiveness of various estimates* (5.2) *for estimates of gene coefficient of variation, Pearson correlatoin, and Spearman correlation. The boxplots are grouped by data set* (*Baron*, *Chen*, *La Manno*, *and Zeisel*). *For each data set, the left-most boxplot in gray shows the distribution of* $f(X)$ *on the original* X, *followed by six boxplots in differing colors showing the distribution of* $\widehat{f(X)} - f(X)$ *with* $\widehat{f(X)}$ *obtained by the methods listed in the legend at bottom.*

interpreted as quantifying the uncertainty in the cell positions. It is natural that the distances computed from the denoised matrix must be adjusted accounting for cell position uncertainty. In this formula, distances between cells with higher denoising uncertainty would be expanded more than distances between cells with lower denoising uncertainty. Note that (5.5) could also be computed through sampling from the posterior distribution using the posterior parameters given by SAVER and SAVER-X. This would be more computationally intensive as one would need to repeatedly sample $X|Y, \Lambda = \hat{\Lambda}$ and then compute d^2 from the sampled X's. We will refer to this as the posterior sampled distances, as opposed to the analytically corrected distances given by (5.5).

Should these corrections help visualization? To explore whether and how these corrections might affect downstream visualization results, we test their effects on several scRNA-seq data sets. First, consider data generated from healthy mouse kidneys from Park et al. (2018). Since the kidney is a multifaceted organ containing diverse cell types, we chose a subset of cell types that have a protective and immunomodulatory role (viz. NK cells, macrophages and podocytes) to visualize. We considered the cell type labels assigned by the authors to be the "ground-truth," and performed denoising on this dataset using SAVER-X (without pretraining on external data sets). First, we compared the scales of the plug-in distances $\sum_g (\hat{X}_{gc} - \hat{X}_{gc'})^2$ to the added cell uncertainty terms $\sum_g v_{gc}$. As one might expect, we want to examine not only the absolute magnitude but also the level of variation in these terms: If the variation in $\sum_g v_{gc}$ is, on the whole, much smaller than the variation in $\sum_g (\hat{X}_{gc} - \hat{X}_{gc'})^2$, then the analytical correction would be less likely to affect visualization. On the other hand, if $E[d^2(c, c')|Y, \Lambda = \hat{\Lambda}]$ is dominated by the variation in the cell uncertainty terms, then we would expect visualization based on corrected distances to be very different from that based on plug-in estimates. As shown in Figure 4, the correction terms are small in magnitude and also have much smaller variation across cell pairs as compared to the distance terms computed directly from the denoised matrix, and this trend holds true for the three

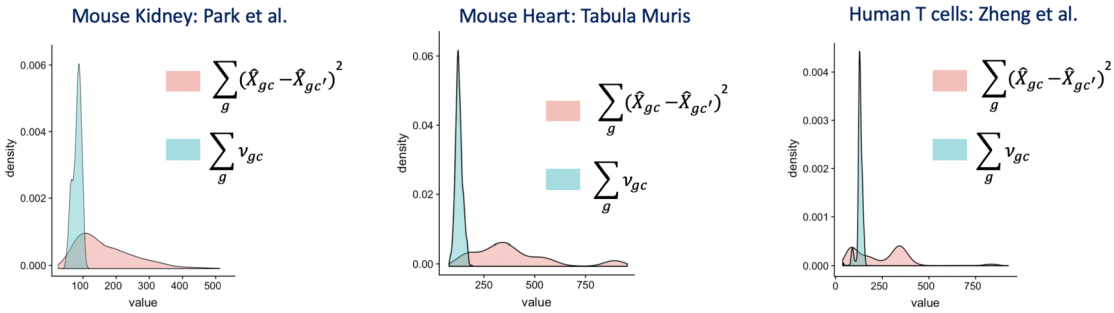

FIG. 4. *Density plots showing the distribution of the plug-in distances and the correction terms for three datasets. Across the different datasets, we observe that the variance term that accounts for uncertainty (blue) is, in general, much smaller in magnitude compared to the corresponding signal (red).*

datasets shown in the figure. Hence, we do not expect the uncertainty adjustment to change broad, strong patterns in the visualization, but may affect more local, subtle patterns.

Figure 5 shows the tSNE plots for the three data sets in Figure 4, constructed from: the raw data (without denoising), the denoised data without adjustment, and the denoised data with two types of adjustments: analytical via (5.5) and Monte Carlo sampling form posterior. As expected, in terms of macro-structures, for these three datasets neither analytical correction nor sampling from the posterior produced tSNE plots that have appreciable differences from the one derived from plug-in estimates.

Yet denoising correction can sometimes remove artifacts or highlight local structures that are not evident without the correction. Consider the second data set in Figures 4 and 5 consisting of cells from the mouse heart (The Tabula Muris Consortium, 2018). Although the five large clusters representing the five major cell type groups (erythrocyte, cardiomyocyte, endothelial, fibroblast, and endocardial) don't change, the corrected distances (both analytical and posterior-subsampled) allow for the separation of a fibroblast subcluster that is not visible in the

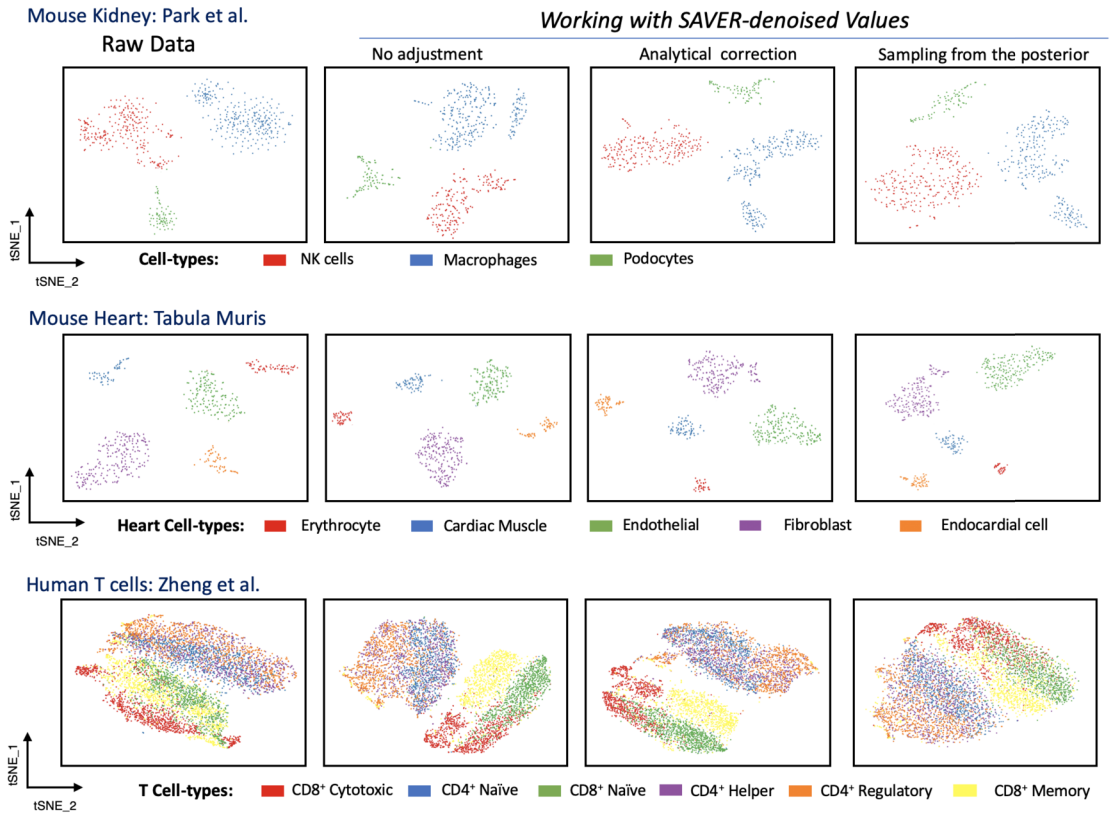

FIG. 5. *Visualizations obtained from three types of denoised distance matrices. tSNE visualizations of three immune cell types in the mouse kidney* (Park et al., 2018), *five major cell types that constitute the mouse heart* (The Tabula Muris Consortium, 2018), *and human T cells* (Zheng et al., 2017) *using distance matrices computed from raw and denoised data are shown.*

tSNE plots derived from raw or uncorrected denoised distances. Although one needs to be careful that embedding artifacts in tSNE plots can produce spurious structures, the fact that this subcluster is evident after both the analytical and posterior-subsampling corrections makes it more likely to be real. Follow-up, either through annotation of differentially expressed genes or wet lab experiments, would be needed to confirm the biological meaning of this subpopulation.

Now consider the third data set in Figures 4 and 5 consisting of human T cells from Zheng et al. (2017). Compared to the previous two examples, this is a more "difficult" data set, as T cell subtypes are much more similar to each other than are the major cell type groups in the mouse heart and kidney. Thus, denoising produces more drastic changes, for example, We see that the separation of the memory and naive CD8+ T cells, and between the CD4+ and CD8+ major groups, is improved. See Wang et al. (2019) for an extensive analysis of this data set. Here, we want to point out that the sampling-based correction seems to perform worse than the analytical correction, as it blurs the distinction between subpopulations.

5.1.3 *Clustering*. Clustering, usually coming hand-in-hand with visualization, is a recurring scRNA-seq analysis. For any clustering method, the cell membership labels can be viewed as functions of X. Here, we evaluate whether post-denoising adjustment would help give more accurate labels.

As in dimension reduction, most clustering algorithms start with cell-to-cell distances, which are also functions of X. In Section 5.1.2, we discussed how estimation of these distances could be biased if we were to simply use $\hat{X}$ as a plug-in for X. However, as for visualization, cell-to-cell distances are only the means to an end, and how slightly biased estimates of cell-to-cell distances affect clustering accuracy is unclear. Here, we evaluate clustering accuracy on the aforementioned four simulation datasets using Seurat (Stuart et al., 2019).

We utilize the Adjusted Rand Index (ARI) to compare clustering results obtained by using denoised values, versus those obtained by using the "true" X. We perform the sampling-based post-denoising correction by generating a new $X^\star$ from the estimated posterior distribution of X given Y and $\Lambda = \hat{\Lambda}$. As shown in Figure 6, all of the denoising methods improve clustering accuracy. For SAVER-X, there is not much difference between the ARIs obtained by clustering using $\hat{X}$ and those obtained by using $X^\star$. This leads us to conclude that, at least for this set of 4 benchmark data sets, post-denoising correction does not appear to give appreciable difference.

This is only an exploratory discourse on how denoised matrices can be used for visualization and clustering. More research in how to account for uncertainty in a computationally efficient way would be necessary as such analyses become mainstream. As a practical guide, we do indeed find denoising to improve clustering and visualization for low-coverage scRNA-seq data. If the downstream algorithm allows, analytical corrections such as (5.5) or posterior-subsampling have the potential to enhance visualization.

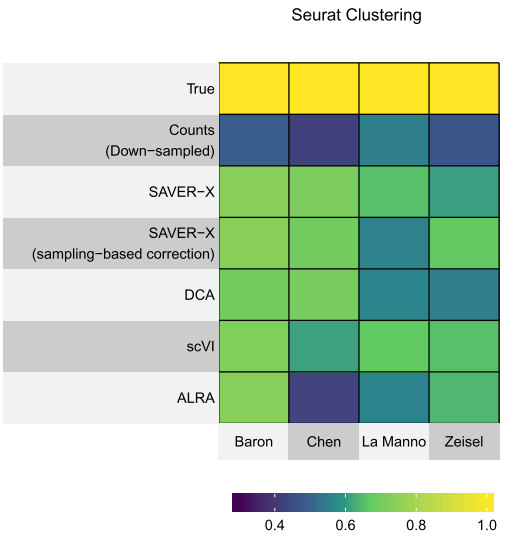

FIG. 6. *Effect on cell clustering. Heatmap of the adjusted rand index for Seurat clustering of the four data sets (Baron, Chen, La Manno and Zeisel) starting from the denoised values obtained from 7 methods (Original data treated as truth, down-sampled raw counts, SAVER-X, SAVER-X with posterior sampling, DCA, scVI, and ALRA).*

5.2 Gene Expression Differential Testing

In scRNA-seq, we are often faced with the task of finding genes that have different distributions between two groups of cells, and testing for differentially expressed genes is a common analysis. A relevant concept here is that of "marker genes," which are genes that are highly expressed in only *select* cell types, as opposed to housekeeping genes, which may be expressed in several or even all cell types. See the Supplementary Material (Agarwal, Wang and Zhang, 2020) a more detailed review and Soneson and Robinson (2018) for a detailed comparisons of some existing DE methods. Many denoising methods can heighten the contrast of marker genes between cell clusters in low dimensional visualizations, thus aiding the identification and labeling of cell types. However, beyond improving visualization, does denoising improve the power of differential expression testing, while keeping the false positive rate controlled? As reviewed in Section 4.1, all scRNA-seq denoising methods tend to inflate the false positive rate beyond the nominal value. This is expected as, intuitively, denoising introduces correlation across cells. Entries of the denoised matrix $\widehat{X}$ cannot simply be treated as i.i.d. across cells, thereby violating a common assumption made by differential testing methods. Here, we examine the severity of false positive rate inflation under various denoising methods, and propose strategies to minimize false discoveries.

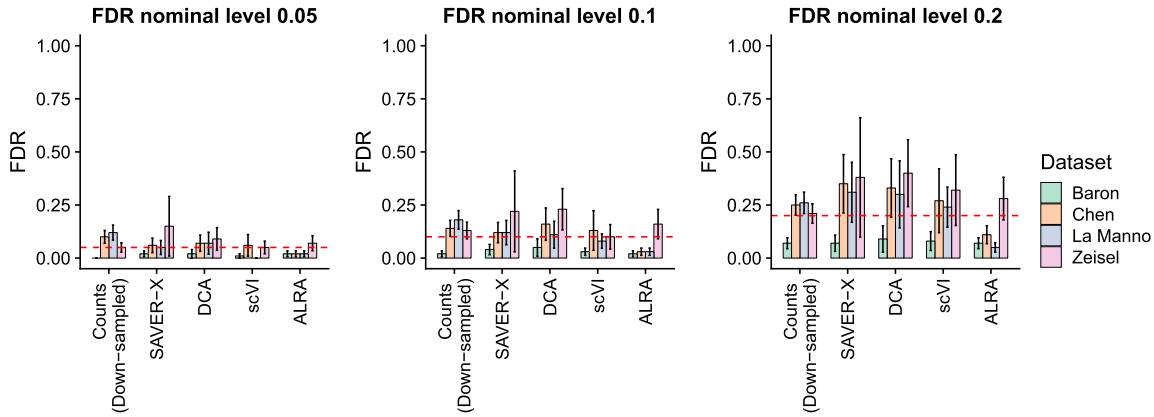

FIG. 7. *Complete null permutation simulation.*

5.2.1 *In silico experiments to evaluate power and type I error control.* To accurately assess false positive rate and power, datasets where the set of true positives is known are needed. For real scRNA-seq data, we are, at best, certain of only a small set of known marker genes, and the full set of true differentially expressed genes is unknown and perhaps not meaningful (practically, does a tiny ϵ shift in expression matter?) Thus, the common strategy is to create artificial data sets through some combination of permutation, data splitting, subsampling, or simulation. Through our engagement of this topic, we learned of the difficulty in creating a *realistic* dataset that allows us to give realistic evaluations of both the type I and type II error of methods, and we have yet to find a satisfactory procedure.

First, let's begin by appreciating the subtle pitfalls of the three existing simulation designs as they have appeared in the single cell literature. The first strategy is a complete null scenario, implemented in Huang et al. (2018). In this design, a population of relatively homogeneous cells is randomly split into two groups, creating a situation where all genes are null, and thus any rejection counts as a false positive. However, this design would not allow us to assess power. Furthermore, valid type I error control under the complete null setting does not imply that the validity would hold as signals (i.e., true non-null genes) are introduced. In fact, any denoising method coupled with a permutation-based significance testing procedure would give valid p-values under such design, so long as the group labels are ignored while denoising. As shown in Figure 7, all methods indeed effectively control the false positive rate at the nominal level under this complete null scenario.

A second approach, as used in Soneson and Robinson (2018), adds simulated shifts to a set of genes in the complete null design described above. Here, starting with a permutation of a real dataset, a fraction of genes are then randomly selected to be the non-null genes. For these genes, the directions and magnitudes of their between-group fold changes are randomly generated, and their observed counts are modified to give their target fold-change. The problem with this design is that the non-null genes are randomly chosen, and and thus are not correlated with each other. This is not true in real data, where marker genes are involved in similar processes and often have highly correlated expression. Denoising methods all rely on such correlations between true signals to boost power, and thus, in the absence of such correlations this scheme does not accurately reflect the sensitivity boost allowed by denoising.

A third design, used in Andrews and Hemberg (2018), is based on permutation of only a proportion of the genes of a real dataset. Starting with two cell populations that are known to have biological differences, a set of differentially expressed (DE) genes is identified. For genes that are not identified as DE, their values are permuted across cells to guarantee that they are true nulls. The identified DE genes are kept fixed. Such simulations detected greatly inflated false positive rates after denoising (Andrews and Hemberg, 2018). However, since most methods are testing for DE in relative expression, where the expressions of all genes in a given cell are normalized to sum to 1, the permuted genes are not guaranteed to be true nulls after normalization unless the fold changes of the non-null genes are balanced (they usually are not). Thus, false positive inflation in real data should be much less severe than what Andrews and Hemberg (2018) described.

Given the above considerations, we compared existing denoising methods using a modification of the third design. First, we selected two cell types—using the labels provided in the original papers—from each of the four datasets:

- Baron et al. (2016): acinar, ductal
- Chen et al. (2017): Ependy, Tany
- La Manno et al. (2016): hRgl2a, hRgl3

- Zeisel et al. (2015): Oligo3, Oligo6

For each data set, we chose cell types that have at least 100 cells and that are very similar to each other, so as to create a difficult DE testing scenario. Let X denote the true relative expression matrix where $\sum_{g=1}^{G} X_{gc} = 1$. First, we select from X a set of differentially expressed genes $\mathcal{G}$ by identifying those genes whose Wilcoxon Rank sum test p-value is less than 0.05 after Benjamini–Hochberg (BH) adjustment. The rest of the genes ($\mathcal{G}^C$) are considered as nulls. To guarantee that $\mathcal{G}^C$ are truly null after permutation and normalization, we perform rescaling for each cell:

$$X_{gc}^{\text{new}} = \begin{cases} kX_{gc} \Big/ \Big(\sum_{g \in \mathcal{G}^C} X_{gc} \Big) & \text{if } g \in \mathcal{G}^C, \\ (1-k)X_{gc} \Big/ \Big(\sum_{g \in \mathcal{G}} X_{gc} \Big) & \text{if } g \in \mathcal{G}, \end{cases} \tag{5.6}$$

where k is the mean of $(\sum_{g \in \mathcal{G}^C} X_{gc})$ across c.

With such rescaling, the *relative expressions* are guaranteed to be the same between the two groups for the null genes after permutation. The genes in G^C are permuted jointly so as to keep their gene-gene correlations intact. We down-sample X^{new} to get the sparse "observed" Y^{new}, which is the input to the denoising method. Such a process is repeated 10 times to quantify the uncertainty of the false discovery proportion (FDP).

As shown in Figure 8, all four denoising methods (SAVER-X, DCA, scVI, ALRA) indeed increase the FDP beyond the pre-selected nominal value of 0.05. This observation raises concerns in drawing strong inferences based on the denoised data in real applications, and motivated us to consider whether, and how, one might be able to reduce the FDP inflation.

5.2.2 *Adjusted Wilcoxon-rank sum test for denoised data*. Conducting valid statistical tests using denoised data is an important point of inquiry, especially as more and more applications involve datasets that are first denoised before downstream inference. We propose a modification to the Wilcoxon rank sum test that appears to effectively alleviate type 1 error inflation in our experiments.

Suppose that we run the standard Wilcoxon rank sum test for each gene between two groups of cells in X. Let the indices of cells in group 1 be $\{i_1, i_2, \ldots, i_m\}$, and the indices for group 2 be $\{j_1, j_2, \ldots, j_n\}$.

In the SAVER/SAVER-X models, we have assumed that the underlying Λ is fixed. However, for DE testing we need to allow Λ to be random, independent and identically distributed according to one distribution for group 1, and another (possibly the same) distribution for group 2. In other words, we need to impose i.i.d. assumptions on Λ thus X within each group for the differential testing

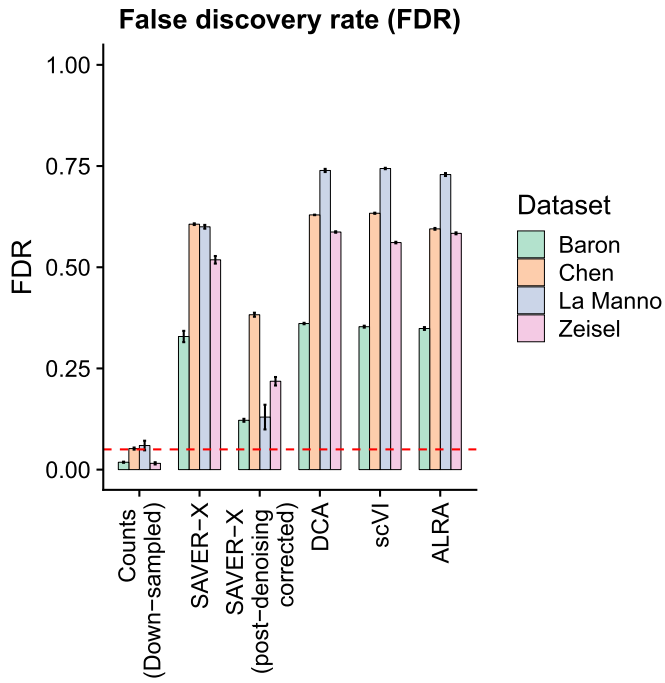

FIG. 8. *Permutation simulation with true differentially expressed genes obtained from the DE analysis of two similar cell types. The true data for each of the four datasets was obtained as described in the main text.*

problem to be well defined, though such an assumption is not needed in denoising.

Now we can assume that

$$X_{gi_1}, X_{gi_2}, \ldots, X_{gi_m} \overset{\text{i.i.d.}}{\sim} F_{g1}$$

for group 1, and

$$X_{gj_1}, X_{gj_2}, \ldots, X_{gj_n} \overset{\text{i.i.d.}}{\sim} F_{g2}$$

for group 2. Were we observing X, assume that we have some standardized test statistic $U(\cdot)$ for the null hypothesis

$$H_{0g} : F_{g1} = F_{g2}.$$

By the large sample approximation,

$$U(X_{g\cdot}) \sim N(0, 1).$$

This is the case when $U(X_{g\cdot})$ is the Wilcoxon Rank sum test statistic for gene g computed using the matrix X. However, X is not observed, and the test statistic using denoised values $U(\widehat{X}_{g\cdot})$ can largely deviate from $N(0, 1)$, that explains why we would have type 1 error inflation after denoising.

With SAVER and SAVER-X, we can approximate the posterior distribution of $U(X_{g\cdot}) \mid Y, \Lambda$ by sampling from the estimated posterior distribution of X. Our simulations show that empirically, the posterior distribution of the test statistic is also approximately Gaussian. In other words:

$$U(X_{g\cdot}) \mid Y, \Lambda \overset{\cdot}{\sim} N(\widehat{\mu}_g, \widehat{\sigma}_g^2),$$

TABLE 1

Differential testing between regulatory CD4+ T cells and other cells on pre-selected marker genes for regulatory CD4+ T cells. P-values are adjusted using BH on all measured genes. Significant p-values (FDR level : 0.05) with a positive logFC are highlighted in red

	logFC (raw data)	logFC (SAVER-X)	adjusted p-values (raw data)	adjusted p-values (post-denoising)
FOXP3	0.14	2.3	2.6e-01	**8.2e-03**
IL2RA	0.66	2.3	**2.0e-04**	**6.1e-04**
DUSP4	0.42	2.5	**7.4e-04**	**6.8e-04**
RGS1	0.17	1.3	6.8e-01	**3.3e-02**
IL2RB	0.48	1.2	**8.1e-03**	**5.1e-04**

where $\widehat{\mu}_g$ and $\widehat{\sigma}_g^2$ are obtained from the posterior samples. Thus, the unobserved quantity $U(X_{g\cdot})$ can be written as

$$U(X_{g\cdot}) \approx \widehat{\mu}_g + \widehat{\sigma}_g(Y, \Lambda)\epsilon, \tag{5.7}$$

where $\epsilon \sim N(0, 1)$.

It is hard to approximate the null distribution of $U(\widehat{X}_{g\cdot})$ but we may have information on the null distribution of $\widehat{\mu}_g$ under H_{0g}. With (5.7), we have

$$\begin{aligned}\operatorname{Var}(\widehat{\mu}_g - U(X_{g\cdot})) &= E\big(\operatorname{Var}[\widehat{\mu}_g - U(X_{g\cdot}) \mid Y, \Lambda]\big) \\ &\quad + \operatorname{Var}(\widehat{\mu}_g - U(X_{g\cdot})) \approx \widehat{\sigma}_g^2,\end{aligned}$$

where $\mu_g = E[U(X_{g\cdot}) \mid Y, \Lambda]$. Under H_{0g}, the unobserved $U(X_{g\cdot}) \sim N(0, 1)$. Thus, $\widehat{\mu}_g = U(X_{g\cdot}) + \widehat{\mu}_g - U(X_{g\cdot})$ can be approximated by a Gaussian distribution

$$\widehat{\mu}_g \dot{\sim} N\big(0, \tau^2\big)$$

with τ^2 conservatively estimated by

$$\widehat{\tau}^2 = 2\big(1 + \widehat{\sigma}_g^2\big).$$

This leads to a post-denoising adjusted p-value.

If (5.7) is exact, then it always holds that

$$\operatorname{Var}(\widehat{\mu}_g) \le \operatorname{Var}\big(U(X_g)\big) = 1$$

under H_{0g} as ϵ is independent of Y and Λ.

In simulations we find that $\widehat{\mu}_g$ across the null genes always has a dispersion larger than 1, indicating that the uncertainty in estimating the posterior distribution of $U(X_g)$ cannot be ignored.

As shown in Figure 8, type 1 error inflation is effectively reduced with this post-denoising adjustment, although it still does not achieve the nominal value. More effective ways to obtain exact error control post-denoising are needed.

5.2.3 *Power in marker gene identification*. We showed in Wang et al. (2019) that denoising substantially enhances the contrast of marker genes across cell clusters and facilitates the labeling of cell types. Yet, often we would like to discover new marker genes through a multiple hypothesis test. How does the above correction to the Wilcoxon Rank Sum test affect marker gene identification in such a formal test? Are known marker genes significant in the denoised data after the correction? To address this, we re-examined the 500 PBMC T cells analyzed in Wang et al. (2019), where the raw data was taken from Zheng et al. (2017). SAVER-X, pretrained on a large set of publicly available PBMC T cells, improved the visual contrast for a set of known marker genes across T cell subtypes. We compared the DE p-values of these known marker genes computed on raw data to those computed on the denoised data with our proposed correction.

We focus on the pre-selected known markers genes (displayed in the third panel of Figure 2b in Wang et al. (2019)), and limit our discussion here to two cell types—regulatory CD4+ T cells and naive CD4+ T cells—where the known marker genes are among the hardest to identify in raw data. For each cell type, differential testing is conducted using Wilcoxon rank sum test between cells of the given type and all remaining cells. For both the raw p-values and the post-denoising corrected p-values, we control false discover rate (FDR) using the BH procedure at the nominal level of 0.05.

Results of the comparisons between using the raw data without denoising and using post-denoising correction are shown in Table 1 for regulatory CD4+ T cells, and in

TABLE 2

Differential testing between naive CD4+ T cells and other cells on pre-selected known marker genes for naive CD4+ T cells. P-values are adjusted using BH on all measured genes. Significant p-values (FDR level : 0.05) with a positive logFC are highlighted in red

	logFC (raw data)	logFC (SAVER-X)	adjusted p-values (raw data)	adjusted p-values (post-denoising)
LEF1	0.273	0.638	7.19e-01	**4.28e-04**
TCF7	0.106	0.42	8.04e-01	**1.21e-02**
SATB1	0.143	0.695	8.04e-01	8.48e-02
SELL	0.615	0.748	**4.01e-04**	**6.84e-07**

Table 2 for naive CD4+ T cells. First, note that denoising followed by Wilcoxon rank sum test (with adjustment as proposed) allows us to detect all but one of the marker genes; some of them were insignificant in the raw data. Also, note that the log fold-change (logFC) computed from the denoised data are typically larger. This is a surprising insight because we expect denoising to reduce standard errors but keep the effect sizes approximately the same. The increase in logFC is likely due to the fact that denoising allows the computing of fold changes without the addition of pseudocounts, which would be otherwise necessary due to the extreme sparsity of the observed count matrix. Denoising obviates the need of this ad hoc hack, and therefore gives more realistic fold-change estimates.

6. CONCLUSION

We set out in this paper with three main goals—introduce the burgeoning field of single cell transcriptomics to the statistics community, review the statistical framework (SAVER/SAVER-X) for data denoising in the context of scRNA-seq, and explore the challenges and opportunities that lie in statistical inference using the denoised values. The latter, in particular, harbors interesting challenges and invites novel methodological ideas. Since denoising, imputation, and/or data-smoothing is now a routine pre-processing step in many applications, it is important to understand how these steps impact downstream statistical estimation and testing.

We explored a plug-in strategy (5.2) for denoising-uncertainty correction in obtaining more accurate estimates of functions of the data, such as gene dispersion, gene-gene correlations, and cell-cell distances (for clustering and visualization). For achieving unbiased estimates of gene-level functions, we found (5.2) to work well. For clustering and visualization, (5.2) gives more precise estimates of cell-cell distances. Although, in our analyses, this has not affected the detection of macrostructures in the data, it may allow more accurate identification of more subtle subpopulations.

Furthermore, we demonstrated that directly using the denoised values for differential testing may severely inflate the false positive rate, corroborating the findings of Andrews and Hemberg (2018) and Zhang and Zhang (2018). We developed a post-denoising correction that mitigates the type-I inflation, but there is still ample room for improvement. Further work needs to be done to find strategies to control type-1 error, while preserving power, for general tests using denoised data.

ACKNOWLEDGMENTS

D. A. is grateful to *Vecteezy* for their vector graphics and illustration toolkit, which was used to generate Figure 1.

SUPPLEMENTARY MATERIAL

Supplement to "Data Denoising and Post-Denoising Corrections in Single Cell RNA Sequencing" (DOI: 10.1214/19-STS7560SUPP; .pdf). Supplementary information.

REFERENCES

ANDREWS, T. S. and HEMBERG, M. (2018). False signals induced by single-cell imputation. *F1000Res* **7**.

AGARWAL, D., WANG, J. and ZHANG, N. R (2020). Supplement to "Data Denoising and Post-Denoising orrections in Single Cell RNA Sequencing." https://doi.org/10.1214/19-STS7560SUPP.

ARKIN, A., ROSS, J. and MCADAMS, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics* **149** 1633–1648.

BADSHA, M. B., LI, R., LIU, B., LI, Y. I., XIAN, M., BANOVICH, N. E. and FU, A. Q. (2018). Imputation of single-cell gene expression with an autoencoder neural network. *BioRxiv* 504977.

BARON, M., VERES, A., WOLOCK, S. L., FAUST, A. L., GAUJOUX, R., VETERE, A., RYU, J. H., WAGNER, B. K., SHEN-ORR, S. S. et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems* **3** 346–360.

BARROSO, G. V., PUZOVIC, N. and DUTHEIL, J. Y. (2018). The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics* **208** 173–189.

BRENNECKE, P., ANDERS, S., KIM, J. K., KOŁODZIEJCZYK, A. A., ZHANG, X., PROSERPIO, V., BAYING, B., BENES, V., TEICHMANN, S. A. et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10** 1093–1095.

BUTLER, A., HOFFMAN, P., SMIBERT, P., PAPALEXI, E. and SATIJA, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36** 411–420.

CHEN, M. and ZHOU, X. (2018). VIPER: Variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol.* **19** 196.

CHEN, R., WU, X., JIANG, L. and ZHANG, Y. (2017). Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Reports* **18** 3227–3241.

CHIRON, L., VAN AGTHOVEN, M. A., KIEFFER, B., ROLANDO, C. and DELSUC, M.-A. (2014). Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry. *Proc. Natl. Acad. Sci. USA* **111** 1385–1390.

CLEVERS, H., RAFELSKI, S. and ELOWITZ, M. et al. (2017). What is your conceptual definition of 'cell type' in the context of a mature organism? *Cell Systems* **4** 255–259.

DEGRELLE, S. A., HENNEQUET-ANTIER, C., CHIAPELLO, H., PIOT-KAMINSKI, K., PIUMI, F., ROBIN, S., RENARD, J.-P. and HUE, I. (2008). Amplification biases: Possible differences among deviating gene expressions. *BMC Genomics* **9** 46.

DI GREGORIO, A., BOWLING, S. and RODRIGUEZ, T. A. (2016). Cell competition and its role in the regulation of cell fitness from development to cancer. *Developmental Cell* **38** 621–634.

EBERWINE, J., YEH, H., MIYASHIRO, K., CAO, Y., NAIR, S., FINNELL, R., ZETTEL, M. and COLEMAN, P. (1992). Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. USA* **89** 3010–3014.

ELDAR, A. and ELOWITZ, M. B. (2010). Functional roles for noise in genetic circuits. *Nature* **467** 167–173.

ELOWITZ, M. B., LEVINE, A. J., SIGGIA, E. D. and SWAIN, P. S. (2002). Stochastic gene expression in a single cell. *Science* **297** 1183–1186.

ENGE, M., ARDA, H. E., MIGNARDI, M., BEAUSANG, J., BOTTINO, R., KIM, S. K. and QUAKE, S. R. (2017). Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171** 321–330.

ERASLAN, G., SIMON, L. M., MIRCEA, M., MUELLER, N. S. and THEIS, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10** 390.

GONG, W., KWAK, I.-Y., POTA, P., KOYANO-NAKAGAWA, N. and GARRY, D. J. (2018). DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinform.* **19** 220.

GOSSETT, D. R., HENRY, T., LEE, S. A., YING, Y., LINDGREN, A. G., YANG, O. O., RAO, J., CLARK, A. T. and DI CARLO, D. (2012). Hydrodynamic stretching of single cells for large population mechanical phenotyping. *Proc. Natl. Acad. Sci. USA* **109** 7630–7635.

HAFEMEISTER, C. and SATIJA, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. BioRxiv 576827.

HAGHVERDI, L., LUN, A. T. L., MORGAN, M. D. and MARIONI, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36** 421–427.

HAN, X., WANG, R., ZHOU, Y., FEI, L., SUN, H., LAI, S., SAADATPOUR, A., ZHOU, Z., CHEN, H. et al. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell* **172** 1091–1107.

HEDLUND, E. and DENG, Q. (2018). Single-cell RNA sequencing: Technical advancements and biological applications. *Mol. Aspects Med.* **59** 36–46.

HICKS, S. C., TOWNES, F. W., TENG, M. and IRIZARRY, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19** 562–578.

HSU, L., SELF, S. G., GROVE, D., RANDOLPH, T., WANG, K., DELROW, J. J., LOO, L. and PORTER, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6** 211–226.

HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J. I., RAJ, A., LI, M. et al. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15** 539.

HWANG, B., LEE, J. H. and BANG, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine* **50** 1–14.

ISLAM, S., ZEISEL, A., JOOST, S., MANNO, G. L., ZAJAC, P., KASPER, M., LÖNNERBERG, P. and LINNARSSON, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11** 163–166.

KIM, J. K., KOLODZIEJCZYK, A. A., ILICIC, T., ILLICIC, T., TEICHMANN, S. A. and MARIONI, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6** 8687.

KIM, T., CHEN, I. R., LIN, Y., WANG, A. Y.-Y., YANG, J. Y. H. and YANG, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.* **20** 2316–2326.

KLEIN, A. M., MAZUTIS, L., AKARTUNA, I., TALLAPRAGADA, N., VERES, A., LI, V., PESHKIN, L., WEITZ, D. A. and KIRSCHNER, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161** 1187–1201.

KOLODZIEJCZYK, A. A., KIM, J. K., SVENSSON, V., MARIONI, J. C. and TEICHMANN, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell* **58** 610–620.

LA MANNO, G., GYLLBORG, D., CODELUPPI, S., NISHIMURA, K., SALTO, C., ZEISEL, A., BORM, L. E., STOTT, S. R., TOLEDO, E. M. et al. (2016). Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167** 566–580.

LA MANNO, G., SOLDATOV, R., ZEISEL, A., BRAUN, E., HOCHGERNER, H., PETUKHOV, V., LIDSCHREIBER, K., KASTRITI, M. E., LÖNNERBERG, P. et al. (2018). RNA velocity of single cells. *Nature* **560** 494–498.

LI, W. V. and LI, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9** 997.

LINDERMAN, G. C., ZHAO, J. and KLUGER, Y. (2018). Zero-preserving imputation of scRNA-seq data using low-rank approximation. *BioRxiv* 397588.

LOPEZ, R., REGIER, J., COLE, M. B., JORDAN, M. I. and YOSEF, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15** 1053–1058.

LOSICK, R. and DESPLAN, C. (2008). Stochasticity and cell fate. *Science* **320** 65–68.

MARTINEZ-JIMENEZ, C. P., ELING, N., CHEN, H.-C., VALLEJOS, C. A., KOLODZIEJCZYK, A. A., CONNOR, F., STOJIC, L., RAYNER, T. F., STUBBINGTON, M. J. T. et al. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355** 1433–1436.

MCADAMS, H. H. and ARKIN, A. (1997). Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* **94** 814–819.

NOVICK, A. and WEINER, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. USA* **43** 553–566.

PAPALEXI, E. and SATIJA, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev., Immunol.* **18** 35–45.

PAREKH, S., ZIEGENHAIN, C., VIETH, B., ENARD, W. and HELLMANN, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6** 25533. https://doi.org/10.1038/srep25533

PARK, J., SHRESTHA, R., QIU, C., KONDO, A., HUANG, S., WERTH, M., LI, M., BARASCH, J. and SUSZTÁK, K. (2018). Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360** 758–763.

PEARSON, K. (1982). *The Grammer of Science*. Cambridge Univ. Press, Cambridge.

RAJ, A. and VAN OUDENAARDEN, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135** 216–226.

REGEV, A., TEICHMANN, S. A., LANDER, E. S., AMIT, I., BENOIST, C., BIRNEY, E., BODENMILLER, B., CAMPBELL, P., CARNINCI, P. et al. (2017). Science forum: The human cell atlas. *eLife* **6** e27041.

ROZENBLATT-ROSEN, O., STUBBINGTON, M. J., REGEV, A. and TEICHMANN, S. A. (2017). The human cell atlas: From vision to reality. *Nature News* **550** 451.

SAELENS, W., CANNOODT, R., TODOROV, H. and SAEYS, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37** 547–554.

SKINNIDER, M. A., SQUAIR, J. W. and FOSTER, L. J. (2019). Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16** 381–386.

SONESON, C. and ROBINSON, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15** 255–261.

SONG, R., SARNOSKI, E. A. and ACAR, M. (2018). The systems biology of single-cell aging. *IScience* **7** 154–169.

STUART, T. and SATIJA, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* **20** 257–272.

STUART, T., BUTLER, A., HOFFMAN, P., HAFEMEISTER, C., PAPALEXI, E., MAUCK, W. M., HAO, Y., STOECKIUS, M., SMIBERT, P. et al. (2019). Comprehensive integration of single-cell data. *Cell* **177** 1888–1902.

SVENSSON, V., VENTO-TORMO, R. and TEICHMANN, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13** 599–604.

SVENSSON, V., NATARAJAN, K. N., LY, L.-H., MIRAGAIA, R. J., LABALETTE, C., MACAULAY, I. C., CVEJIC, A. and TEICHMANN, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14** 381–387.

THE TABULA MURIS CONSORTIUM (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562** 367.

TANG, F., BARBACIORU, C., WANG, Y., NORDMAN, E., LEE, C., XU, N., WANG, X., BODEAU, J., TUCH, B. B. et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6** 377.

TESCHENDORFF, A. E. and ENVER, T. (2017). Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. Commun.* **8** 15599.

TIAN, L., DONG, X., FREYTAG, S., LÊ CAO, K.-A., SU, S., JALALABADI, A., AMANN-ZALCENSTEIN, D., WEBER, T. S., SEIDI, A. et al. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16** 479–487.

TRAPNELL, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.* **25** 1491–1498.

TUNG, P.-Y., BLISCHAK, J. D., HSIAO, C. J., KNOWLES, D. A., BURNETT, J. E., PRITCHARD, J. K. and GILAD, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7** 39921.

VAN DIJK, D., SHARMA, R., NAINYS, J., YIM, K., KATHAIL, P., CARR, A. J., BURDZIAK, C., MOON, K. R., CHAFFER, C. L. et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* **174** 716–729.

VAN GELDER, R. N., VON ZASTROW, M. E., YOOL, A., DEMENT, W. C., BARCHAS, J. D. and EBERWINE, J. H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. USA* **87** 1663–1667.

WAGNER, F., YAN, Y. and YANAI, I. (2017). K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *BioRxiv* 217737.

WANG, J., HUANG, M., TORRE, E., DUECK, H., SHAFFER, S., MURRAY, J., RAJ, A., LI, M. and ZHANG, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA* **115** E6437–E6446.

WANG, J., AGARWAL, D., HUANG, M., HU, G., ZHOU, Z., YE, C. and ZHANG, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16** 875–878.

ZAPPIA, L., PHIPSON, B. and OSHLACK, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14** e1006245.

ZEISEL, A., MUÑOZ-MANCHADO, A. B., CODELUPPI, S., LÖNNERBERG, P., LA MANNO, G., JURÉUS, A., MARQUES, S., MUNGUBA, H., HE, L. et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347** 1138–1142.

ZHANG, L. and ZHANG, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* https://doi.org/10.1109/TCBB.2018.2848633

ZHENG, G. X., TERRY, J. M., BELGRADER, P., RYVKIN, P., BENT, Z. W., WILSON, R., ZIRALDO, S. B., WHEELER, T. D., MCDERMOTT, G. P. et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** 14049.

ZIEGENHAIN, C., VIETH, B., PAREKH, S., REINIUS, B., GUILLAUMET-ADKINS, A., SMETS, M., LEONHARDT, H., HEYN, H., HELLMANN, I. et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell* **65** 631–643.