K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data

Florian Wagner[1+], **Yun Yan**[1], **and Itai Yanai**[1*]

[1]**School of Medicine, New York University, New York, NY, USA**
[+]**Email: florian.wagner@nyu.edu**
[*]**Email: itai.yanai@nyumc.org**

ABSTRACT

High-throughput single-cell RNA-Seq (scRNA-Seq) methods can efficiently generate expression profiles for thousands of cells, and promise to enable the comprehensive molecular characterization of all cell types and states present in heterogeneous tissues. However, compared to bulk RNA-Seq, single-cell expression profiles are extremely noisy and only capture a fraction of transcripts present in the cell. Here, we describe an algorithm to smooth scRNA-Seq data, with the goal of significantly improving the signal-to-noise ratio of each profile, while largely preserving biological expression heterogeneity. The algorithm is based on the observation that across platforms, the technical noise exhibited by UMI-filtered scRNA-Seq data closely follows Poisson statistics. Smoothing is performed by first identifying the nearest neighbors of each cell in a step-wise fashion, based on variance-stabilized and partially smoothed expression profiles, and then aggregating their UMI counts. For multiple datasets, the application of our algorithm resulted in more stable cell type-specific expression profiles, and recovered correlations between co-expressed genes. More generally, smoothing improved the results of commonly used dimensionality reduction and clustering methods, greatly facilitating the identification of cell subsets and clusters of co-expressed genes. Our work implies that there exists a quantitative relationship between the number of cells profiled and the potential accuracy with which individual cell types or states can be characterized, and helps unlock the full potential of scRNA-Seq to elucidate molecular processes in healthy and disease tissues.

Keywords: single-cell RNA-Seq, data analysis, k-nearest neighbors, Poisson distribution, algorithms

INTRODUCTION

Over the past decode, single-cell expression profiling by sequencing (scRNA-Seq) technology has advanced rapidly: After the transcriptomic profiling of a single cell (Tang et al. 2009), protocols were developed that incorporated cell-specific barcodes to enable the efficient profiling of tens or hundreds of cells in parallel (Islam, Kjällquist, et al. 2011; Hashimshony, Wagner, et al. 2012). scRNA-Seq methods were then improved by the incorporation of unique molecular identifiers (UMIs) that allow the identification and counting of individual transcripts (e.g., Islam, Zeisel, et al. 2014; Hashimshony, Senderovich, et al. 2016). More recently, single-cell protocols were combined with microfluidic technology (Klein et al. 2015; Macosko et al. 2015; Zheng et al. 2017), combinatorial barcoding (Cao et al. 2017; Rosenberg et al. 2017), or nanowell plates (Gierahn et al. 2017). These high-throughput scRNA-Seq methods allow the cost-efficient profiling of tens of thousands of cells in a single experiment.

Due to the typically very low amounts of starting material, and the inefficiencies of the various chemical reactions involved in library preparation, scRNA-Seq data is inherently noisy (Ziegenhain et al. 2017). This has motivated the development of many specialized statistical models, for example for determining differential expression (Kharchenko, Silberstein, and Scadden 2014), performing factor analysis (Pierson and Yau 2015), pathway analysis (Fan et al. 2016), or more general modeling of scRNA-Seq data (Risso et al. 2017). In addition, a diffusion method has been proposed to impute missing values and perform smoothing (Dijk et al. 2017). Finally, many authors of scRNA-Seq studies have relied on ad-hoc approaches for mitigating noisiness, for example by clustering and averaging cells belonging to each cluster (Shekhar et al. 2016; Baron et al. 2016).

Fundamental to any statistical treatment are the assumptions that are made about the data. For methods aimed at analyzing scRNA-Seq data, assumptions about the noise characteristics determine which approach can be considered the most appropriate. All aforementioned approaches have assumed an overabundance of zero values, compared to what would be expected if the data followed a Poisson or negative binomial distribution. However, in the absence of true expression differences, the analysis by Ziegenhain et al. (2017) has suggested that across scRNA-Seq protocols, there is little evidence of excess-Poisson variability when expression is quantified by counting unique UMI sequences instead of raw reads (see Figure 5B in Ziegenhain et al. (2017)). This is consistent with reports describing individual UMI-based scRNA-Seq protocols, which have demonstrated that in the absence of true expression differences, the mean-variance relationship of genes or spike-ins closely follows that of Poisson-distributed data (Grün, Kester, and Oudenaarden 2014; Klein et al. 2015; Zheng et al. 2017).

In this work, we propose a smoothing algorithm that makes direct use of the observation that after normalization to account for efficiency noise (Grün, Kester, and Oudenaarden 2014), the technical noise associated with UMI counts from high-throughput scRNA-Seq protocols is entirely consistent with Poisson statistics. Instead of adopting a model-based approach, we propose an algorithm that smoothes scRNA-Seq data by aggregating gene-specific UMI counts from the k nearest neighbors of each cell. To accurately determine these neighbors, we propose to use an appropriate variance-stabilizing transformation, and to proceed in a step-wise fashion using partially smoothed profiles. Conveniently, the noise associated with the smoothed expression profiles is again Poisson-distributed, which simplifies their variance-stabilization and downstream analysis. We demonstrate the improved signal-to-noise ratio of scRNA-Seq data processed with our method on several real-world examples.

RESULTS

The normalized UMI counts of replicate scRNA-Seq profiles are Poisson-distributed

To validate the Poisson-distributed nature of high-throughput scRNA-Seq data in the absence of true expression differences, we obtained data from control experiments conducted on three platforms: inDrop (Klein et al. 2015), Drop-Seq (Macosko et al. 2015), and 10x Genomics (Zheng et al. 2017). In these experiments, droplets containing identical RNA pools were analyzed. Assuming that the number of transcripts in each droplet was sufficiently large, there are no true expression differences among droplets, and all of the observed differences among droplets can be attributed to technical noise arising from library preparation and sequencing. As expected from published results (cf. Figure 5A in Klein et al. (2015), Supplementary Figure 2f in Zheng et al. (2017)), data from both the inDrop platform and the 10x Genomics platform followed the Poisson distribution (see Figure 1a,c; see Methods), with the exception of highly expressed genes, which is likely due to global droplet-to-droplet differences in capture efficiency, previously referred to as “efficiency noise” (Grün, Kester, and Oudenaarden 2014).

For the Drop-Seq data, Macosko et al. (2015) did not discuss the mean-variance relationship, but we observed a pattern consistent with inDrop and 10x Genomics data (see Figure 3b). Interestingly, the y axis intercept of the Drop-Seq CV-mean relationship was clearly above 0, suggesting that transcript counts followed a scaled Poisson distribution (see Methods). A possible explanation could be that the computational pipeline used to derive the Drop-Seq UMI counts generated artificially inflated transcript counts, but we did not explore this hypothesis further.

To test whether the larger-than-expected variance of highly expressed genes can indeed be explained by efficiency noise, we normalized the expression profiles in each dataset to the median UMI count across profiles (Model I in Grün, Kester, and Oudenaarden (2014); see Methods). This resulted in an almost perfectly linear CV-mean relationship (see Figure 1d-f), suggesting that efficiency noise is indeed the dominating source of variation for very highly expressed genes.

Finally, we directly compared the frequency of UMI counts of zero for each gene to that predicted by Poisson statistics, and found that for the inDrop and 10x Genomics data, the observed values matched the theoretical prediction almost perfectly (see Figure 3g,i). For the Drop-Seq data, the frequency of zeros was slightly shifted upwards across the entire expression range (see Figure 3h), which may be due to artificially inflated UMI counts.

In summary, we found that for all three high-throughput scRNA-Seq platforms examined, Poisson-distributed noise, in combination with the efficiency noise observed for very highly expressed genes, described virtually all of the observed technical noise, and that there was no evidence of substantial zero-inflation. We note that the recent publication describing the Quartz-Seq2 single-cell platform also reports a Poisson noise relationship (see Figure 2e in Sasagawa et al. (2017)), bringing the total number of high-throughput scRNA-Seq protocols with reported Poisson noise characteristics to four.

Aggregation of n replicate profiles results in Poisson-distributed values with the signal-to-noise ratio increased by a factor of $\sqrt{n}$

Since the sum of independent Poisson-distributed variables is again Poisson-distributed, we reasoned that the aggregation of normalized expression values from n independent measurements of the same RNA pool would result in Poisson-distributed values, with the signal-to-noise ratio increased by a factor of $\sqrt{n}$ (see Methods). Similarly, we predicted that averaging instead of aggregating (summing) would result in a scaled Poisson distribution with the same increased signal-to-noise ratio. We tested this idea on the inDrop pure RNA dataset previously shown in Figure 1a, which consisted of 935 expression profiles. Averaging randomly selected, non-overlapping sets of 16 profiles resulted in 58 new expression profiles, with genes exhibiting an almost exact four-fold increase in their signal-to-noise ratios, i.e., a four-fold reduction of their coefficients of variation, as expected (see Figure 2a). As an example, the UMI count distribution of the *GADPH* gene before and after averaging is shown in Figure 2b, and can be seen to closely match the theoretically predicted Poisson and scaled Poisson distributions, respectively. In summary, the results showed that independently of gene expression level, aggregating expression values from replicate profiles led to more accurate expression estimates that again exhibited Poisson-distributed noise profiles.

The Freeman-Tukey transform effectively stabilizes the technical variance of high-throughput scRNA-Seq data

Based on the aforementioned results, we conceived an algorithm to smooth single-cell RNA-seq data, with the following outline:

- For each cell *C*:
    1. Determine the k nearest neighbors of *C*.
    2. Calculate a smoothed expression profile for *C* by combining its UMI counts with those of the k nearest neighbors, on a gene-by-gene basis.
    3. (Optional) Divide *C*'s new expression profile by k, to retain the scale of the original data.

The main challenge in implementing this algorithm is to devise an appropriate approach for determining the k nearest neighbors of each cell, and to choose an appropriate k. We defer the question of how to choose k to the Discussion, and focus here on the problem of determining the k nearest neighbors.

Due to the Poisson-distributed nature of scRNA-Seq data, the technical variance (noise) associated with each gene is directly proportional to its expression level. This type of extreme heteroskedasticity poses a problem when attempting to calculate cell-cell similarities, because the noise of highly expressed genes can drown out the true expression differences of more lowly expressed genes, therefore strongly biasing the analysis towards the most highly expressed genes. One strategy to address this issue is the application of an appropriate variance-stabilizing transformation, designed to render the technical variance independent of the gene expression level (Love, Huber, and Anders 2014). For bulk RNA-Seq data, a log-TPM (or log-RPKM) transform is commonly used for this purpose, even though lowly expressed genes will still exhibit unduly large variances under this transformation (Love, Huber, and Anders 2014). Based on our results, we reasoned that for scRNA-Seq data, the *Freeman-Tukey transform* (FTT), $y = \sqrt{x} + \sqrt{x+1}$, would be a more appropriate choice, as it is designed to stabilize the variance of Poisson-distributed variables (Freeman and Tukey 1950).

To compare the abilities of the FTT and the log-CPM (counts per million) transform to stabilize the technical variance of scRNA-Seq data, we applied both transformations to the inDrop pure RNA dataset, and found that the FTT produced significantly better results (see Figure 3): With the log transform, genes with low-intermediate expression, which we considered to be those with expression values between the 60th and 80th percentiles (of all protein-coding genes, not only genes expressed by K562 cells), had between three- and ten-fold higher levels of variance than the 10% most highly expressed genes (see Figure 3b). In contrast, with the FTT, the difference was no larger than two-fold, and the variances of lowly expressed genes were biased downwards, not upwards (see Figure 3c). Moreover, we found that the FTT also stabilized the variance of the aggregated profiles (see Figure 3d-f), which was expected, given our earlier observation that the aggregated UMI counts are again Poisson-distributed. In particular, a greater share of genes now had variances close to 1. This closely mirrored theoretical results, according to which the variance Poisson-distributed variables with mean $\lambda \geq 1$ should be within 6% of the asymptotic value of 1 after FTT (Freeman and Tukey 1950). In summary, our analysis showed that distance calculations performed on Freeman-Tukey transformed UMI counts would give similar weight to genes with intermediate and high expression. Expression differences from lowly expressed genes would tend to be suppressed, but this suppression would become less severe for aggregated expression profiles.

A k-nearest neighbor algorithm for smoothing scRNA-Seq data

The previously discussed ideas suggested that a simple way to determine the k nearest neighbors for all cells would be to normalize their expression profiles, apply the FTT, and then find the k closest cells for each cell based on the Euclidean metric. However, we reasoned that this simple approach could be improved upon, because the noisiness of the data itself can interfere with the accurate determination of the k nearest neighbors. We therefore instead decided to adopt a step-wise approach, whereby initially, each profile is only minimally smoothed (using $k_1 = 1$). In the second step, a larger set of nearest neighbors (e.g., $k_2 = 3$) is identified for each cell based on those minimally smoothed profiles, and the raw data is then smoothed using these larger sets of neighbors. Additional steps using increasing k_i are performed until the desired degree of smoothing is reached (i.e., $k_i = k$). By choosing the i'th step to use $k_i = \min\{2^i - 1, k\}$, each step theoretically improves the signal-to-noise ratio by a factor of $\sqrt{2}$ — except for the last step, for which the improvement can be smaller —, and only a small number of steps are required even for large choices of k (e.g., six steps for $k = 63$). The resulting "kNN-smoothing" algorithm is formalized in Algorithm 1 (see Supplement for a reference implementation in Python). We found that in contrast to a simple "one-step" approach, the step-wise identification of neighbors gave significantly better results and avoided the generation of obvious smoothing artifacts (data not shown).

Application of kNN-smoothing to scRNA-Seq data of human pancreatic islets improves the signal-to-noise ratio of cell type-specific expression profiles

To test whether kNN-smoothing would improve the ability to distinguish between different cell types in a scRNA-Seq experiment, we applied our algorithm to a dataset of human pancreatic islets, containing various cell types (Baron et al. 2016). We performed principal component analyses and observed several improvements for the smoothed data (see Figure 4a,b): First, cell type clusters appeared significantly more compact in principal component space, indicating that the smoothed expression profiles were more similar than unsmoothed profiles for cells of the same type, but more different for cells from distinct types. Second, a single cluster of cells that contained alpha cells as well as other cells separated into two highly distinct clusters after smoothing. Notably, all alpha cells were still contained within a single cluster after smoothing. This suggested smoothing helped reveal important differences that were not previously captured by the first two principal components. Third, the proportion of cells of each type that could be identified using simple marker gene expression thresholds increased slightly, suggesting that the expression values of individual marker was less noisy in the smoothed data. Finally, a much greater share of total variation was explained by the first two principal components for the smoothed data than for the unsmoothed data (41.1% vs 23.9%), which would be consistent with a greater share of variation originating from true biological differences rather than technical noise. In addition to PCA, we also applied t-SNE to the data (Maaten and Hinton 2008), and similarly obtained more compact cell type clusters (see Figure 4c,d).

To obtain a more detailed view of the expression patterns of individual genes before and after smoothing, we applied hierarchical clustering to the expression values of the 1,000 most variable genes (after smoothing and variance-stabilization) across all 2,109 cells, which resulted in clearly discernible gene and cell clusters (see Figure 4e). To assess whether cell clusters delineated different cell types, we examined the expression patterns of known marker genes (Baron et al. 2016), and found that the hierarchical clustering of the smoothed expression profiles accurately grouped cells by their cell type (see Figure 4g). Moreover, the expression patterns in clusters appeared significantly more coherent in the smoothed data compared to the unsmoothed data (see Figure 4f), and marker genes exhibited much less noisy expression signatures (see Figure 4h). In summary, our analyses showed that kNN-smoothing significantly improved the signal-to-noise ratio of cell type-specific expression profiles, and led to improved results with dimensionality reduction and visualization techniques such as PCA and t-SNE.

Algorithm 1: K-nearest neighbor smoothing for UMI-filtered scRNA-Seq data

Input:

p, the number of genes.
n, the number of cells.
X, a $p \times n$ matrix containing the UMI counts for all genes and cells.
k, the number of neighbors to use for smoothing.

Output:

S, a $p \times n$ matrix containing the smoothed (aggregated) UMI counts.

```
procedure KNN-SMOOTH($p$, $n$, $X$, $k$)
    $S = \text{COPY}(X)$
    $steps = \lceil \log_2(k+1) \rceil$
    for $t = 1$ to $steps$ do
        $M = \text{MEDIAN-NORMALIZE}(S)$            // a new $p \times n$ matrix
        $F = \text{FREEMAN-TUKEY-TRANSFORM}(M)$     // a new $p \times n$ matrix
        $D = \text{PAIRWISE-DISTANCE}(F)$           // a new $n \times n$ matrix
        $A = \text{ARGSORT-ROWS}(D)$                // a new $n \times n$ matrix
        $k\_step = \text{MIN}(\{2^t - 1, k\})$
        for $j = 1$ to $n$ do        // empty matrix $S$
            for $i = 1$ to $p$ do
                $S_{ij} = 0$
            end for
        end for
        for $j = 1$ to $n$ do        // go over all cells
            for $v = 1$ to $k\_step + 1$ do        // go over all nearest neighbors (including self)
                $u = A_{jv}$
                for $i = 1$ to $p$ do        // aggregate original UMI counts for each gene
                    $S_{ij} = S_{ij} + X_{iu}$
                end for
            end for
        end for
    end for
    return $S$
end procedure
```

Notes: For a two-dimensional matrix X, X_{ij} refers to the element in the i'th row and j'th column of X. COPY(X) returns an independent memory copy of X (not a reference). MEDIAN-NORMALIZE(X) returns a new matrix of the same dimension as X, in which the values in each column have been scaled by a constant so that the column sum equals the median column sum of X. FREEMAN-TUKEY-TRANSFORM(X) returns a new matrix of the same shape as X, in which all values have been Freeman-Tukey transformed ($y = \sqrt{x} + \sqrt{x+1}$). PAIRWISE-DISTANCE$(X)$ computes the pair-wise distance matrix D from X, so that D_{ij} is the Euclidean distance between the i'th column and the $j'th$ column of X. For a matrix D with n columns, ARGSORT-ROWS(D) returns a matrix of indices A that sort D in a row-wise manner, i.e., $D_{jA_{j1}} \leq D_{jA_{j2}} \leq ... \leq D_{jA_{jn}}$ for all j.

Application of kNN-smoothing to scRNA-Seq data of human peripheral blood mononuclear cells improves correlations between cell type marker genes

We next applied our kNN-smoothing algorithm to a dataset containing peripheral blood mononuclear cells (PBMCs) (Gierahn et al. 2017), and examined the correlation between individual T cell and monocyte marker genes before and after smoothing, using $k = 15$ and $k = 63$ (see Figure 5). For the T cell receptor genes *CD3E* and *CD3G*, only weak correlation ($r = 0.20$) was observed for the counts before smoothing. However, after smoothing with $k = 63$, the correlation was extremely strong ($r = 0.90$). Similarly, the correlation between *CTSB* and *SOD2*, two markers used by Gierahn et al. (2017) to identify monocytes, improved from $r = 0.35$ to $r = 0.88$, revealing a clear bimodal pattern. As expected, the anti-correlation between the monocyte marker *CTSB* and the T cell marker *CD3E* changed from weak to very strong (Figure 5g-i). In summary, smoothing resulted in the effective recovery of strong yet previously undetectable co-expression patterns among marker genes.

Application of smoothing to scRNA-Seq data of mouse myeloid progenitor cells

To compare our method to a previously proposed approach (Dijk et al. 2017), we applied our smoothing algorithm to a scRNA-Seq dataset of mouse myeloid progenitor cells (Paul et al. 2015). We generated a heatmap of characteristic genes for 19 clusters identified by the authors of the original study, as well as for important cell surface markers, in a way that allows a direct comparison to the results obtained by Dijk et al. (2017) (see Figure 6a,b). We found that even though k-nearest neighbor smoothing is much simpler than their approach, our method performed similarly well in generating smooth expression profiles for cells belonging to the same cluster, while respecting cluster boundaries.

We similarly examined the pairwise correlations of cell surface markers, and obtained qualitatively similar results to Dijk et al. (2017) (see Figure 6c-e). As in their study, recovering cell type-specific co-expression patterns depended on the amount of smoothing applied. Some differences were observed in the precise shapes of the associations, but it was not clear how much of this was due to differences in normalization and/or scaling used for visualization. In summary, for this particular dataset, the diffusion-based approach by Dijk et al. (2017) and our algorithm gave qualitatively similar results, although there were some quantitative differences.

DISCUSSION

Comparison with previously reported methods

In this work, we have described a simple yet effective algorithm for smoothing single-cell RNA-Seq data. Our algorithm combines a previously proposed normalization method (Grün, Kester, and Oudenaarden 2014) with a standard variance-stabilizing transformation (VST) for Poisson-distributed data (Freeman and Tukey 1950). We are not aware of prior work suggesting the use of a VST in the context of smoothing scRNA-Seq data. Instead, most work has focused on parametric modeling (see Introduction). While these approaches can certainly be effective, our work suggests that they are not strictly necessary to effectively to address the issue of noise in scRNA-Seq data. Moreover, sophisticated models often require complex inference procedures, which can be difficult to implement correctly and efficiently. In contrast, our method requires only a few lines of code, while still being based on statistical theory.

Our approach relies on the basic notion of smoothing scRNA-Seq expression profiles by aggregating them with similar cells. Simple aggregation or averaging of scRNA-Seq expression profiles has been previously employed in specific contexts, for example for library size normalization (Lun, Bach, and Marioni 2016). Recently, La Manno et al. (2017) employed a simple version of k-nearest neighbor smoothing ("pooling") as part of a method designed to estimate the time derivative of mRNA abundance based on unspliced RNA sequences. The authors defined the most similar cells based on log-transformed data (for read counts from the SMART-Seq2 protocol), or PCA-transformed data (for UMI counts from inDrop and 10x Genomics protocols). However, they did not provide any justification for their choices of similarity metrics, nor a discussion of the statistical properties of the data before and after smoothing. Moreover, neither of these studies aimed to develop a general-purpose method to improve the signal-to-noise ratio of scRNA-Seq data, or employed a step-wise approach for defining the nearest neighbors, as we have done here. As a general method for smoothing, our work can be compared to a recently proposed diffusion-based approach (Dijk et al. 2017). However, van Dijk et al. aimed to apply the idea of manifold learning using diffusion maps to scRNA-Seq data, whereas we aimed to rely on a specific statistical property of scRNA-Seq data, namely its Poisson-distributed noise profile. Second, our method currently requires researchers to specify only a single parameter, k, which has a clear meaning (the number of neighbors to use for smoothing). The diffusion algorithm proposed by van Dijk et al. relies on three parameters (ka, $npca$, and t), and the extent to which different parameter combinations can give quantitatively or qualitatively different results is not always obvious, especially for different settings of t. Third, the algorithm proposed by Dijk et al. (2017) involves the calculation of *weighted* averages of expression profiles, which do not result in Poisson-distributed values (unless all the weights are equal). For example, if X_1 and X_2 are two independent Poisson-distributed variables, then $Y = 0.4 * X_1 + 0.6 * X_2$ is neither a Poisson nor a scaled Poisson variable. As a result, a simple transformation like the Anscombe transform will not be able to accurately stabilize the variance of data smoothed using the method proposed by van Dijk et al., which can make downstream analyses more challenging. We therefore believe that the method described here is unique in the sense that each step is motivated by the statistical properties of the data, and that it is guaranteed to retain its Poisson-distributed nature. This property facilitates downstream analyses using variance-stabilization transformations or parametric models.

How to choose k?

The choice of k directly affects the results obtained when smoothing a particular dataset using our method. Choosing k very small might not adequately reduce noise. On the other hand, choosing k too large incurs the risk of smoothing over biologically relevant expression heterogeneity. Moreover, large k can also lead to artifactual expression profiles that consist of averages of profiles belonging to different cell populations. Our method provides no guarantee that a smoothed expression profile accurately reflects an existing cell population. During the exploratory phase of data analysis, we therefore recommend to test different choices of k. When a signal of interest has been identified (such as a gene-gene correlation, a cluster of cells, an expression signature, etc.), it can be determined what minimum of value of k is required in order to obtain this signal. When this value is large, adequate controls should be performed to ensure that the observed signal is not a smoothing artifact.

An appropriate choice of k also depends on the particular application: When analyzing cells undergoing a highly dynamic process (e.g., differentiation), large values of k might result in an overly coarse picture of the transcriptomic changes. In contrast, when aiming to distinguish distinct cell types, larger choices of k can help identify robust expression profiles for each type.

Implications for study design

Based on the work described here, it appears tempting to speculate that in theory, there is no limit as to how accurately the average expression profile of individual cell populations and sub-populations can be determined using scRNA-Seq. Our analysis suggests that the signal-to-noise ratio can always be improved by aggregating more profiles from "biologically identical" cells. In practice, however, the number of cells that can be analyzed is limited by the protocol used, the cost of the experiment, the number of cells available, and/or the rarity of the population of interest. Furthermore, the accuracy with which "biologically identical" cells can be identified based on their noisy profile depends on several factors, including the granularity required (e.g., can cells in different cell cycle stages be considered identical for the purpose of the analysis?), and the precise measure of similarity adopted. When the transcriptomic differences between cell populations of interest become too small to allow a reliable identification of neighbors, it is not clear how to perform smoothing and extract the true biological signal. In this work, we have determined similarity on the basis of the expression of all genes, but restricting this calculation to a subset of genes could be more appropriate in certain settings.

More generally, the quadratic relationship between relationship between "cell coverage" (loosely defined as the average number of profiles obtained for each cell population) and quantification accuracy brings into focus the question of what constitutes an optimal number of sequencing reads per cell. While a quantitative treatment of this issue is beyond the scope of this work, it is clear that in certain cases, it would be more beneficial to sequence additional cells, rather than increase the read coverage per cell. The precise optimum likely depends on numerous factors, and is difficult to determine without an examination of all the experimental, statistical, and computational factors involved in scRNA-Seq studies. However, since sequencing often represents the single most expensive part of the experiment, this question clearly warrants further investigation.

Future directions

In this work, we have used multiple datasets to demonstrate that basic techniques for exploratory analysis of gene expression data (PCA, t-SNE, hierarchical clustering, correlation analysis) benefit strongly from our kNN-smoothing algorithm. In future work, we hope to explore the effect of smoothing for additional types of analyses, including differential expression analysis, gene set enrichment analysis, or exploratory analysis using prior knowledge (Wagner 2015). We anticipate that our kNN-smoothing algorithm will benefit all of these approaches, and generally enable the more effective analysis of scRNA-Seq data in wide variety of settings. It should be noted, however, that smoothed expression profiles of cells are no longer statistically independent, so smoothing should not be used naively in combination with statistical tests for differential expression.

The use of a global k could limit the effectiveness of our algorithm in cases where different cell populations are present at very different abundances. As an extreme example, if one population constitutes 5% of all cells, and another 95%, k should not be chosen larger than 5% of the total number of profiles, in order to avoid artifacts. However, the expression profile of the population present at 95% could benefit from larger choices of k. It would therefore seem useful to automatically adjust k for each cell. This is the approach chosen by Dijk et al. (2017), who use the distance of a cell to its ka'th neighbor as an important parameter in the calculation of the smoothed profile. However, a complication associated with this approach is that different expression profiles would exhibit distinct technical noise levels, since they would be the result of aggregating or averaging over different numbers of cells. Another way to address this issue would be to cluster cells by type before performing more aggressive smoothing. Ultimately, which strategy is more appropriate might depend on the specific application.

High-throughput scRNA-Seq technology is widely believed to hold enormous potential for the analysis of heterogeneous tissues and dynamic cellular processes in health and disease. However, the inherent noisiness of the data means that greater computational efforts are required in order to realize this potential. Fortunately, data from different protocols exhibit very similar statistical properties, presumably due to their shared reliance on 5'- or 3'-end counting and incorporation of UMI sequences. These properties should directly inform the design of effective algorithms for smoothing and analysis of scRNA-Seq data. We have described a generally applicable, easy-to-implement approach for improving the signal-to-noise ratio of single-cell expression profiles, which promises to significantly expand the realm of possibilities for downstream analyses of scRNA-Seq data.

METHODS

Download and processing of inDrop pure RNA replicate data

Raw sequencing data were downloaded from SRA (experiment accession SRX863258). In this experiment by Klein et al. (2015), droplets containing pure RNA extracted from K562 cells were processed using the inDrop protocol. The downloaded data were processed using a custom pipeline. Briefly, SRA data were converted to the FASTQ format using fastq-dump. Next, the “W1” adapter sequence of the inDrop RT primer were located in the barcode mate sequence (the first mate of the paired-end sequencing), by comparing the 22-mer sequences starting at positions 9-12 in the read with the known W1 sequence, allowing at most two mismatches. Reads for which the W1 sequence could not be located in this way were discarded. The start position of the W1 sequence was then used to infer the length of the first part of the inDrop cell barcode in each read, which can range from 8-11 bp, as well as the start position of the second part of the inDrop cell barcode, which always consists of 8 bp. Cell barcode sequences were mapped to the known list of 384 barcode sequences for each read, allowing at most one mismatch. The resulting barcode combination was used to identify the cell from which the read originated. Finally, the UMI sequence was extracted, and only with low-confidence base calls for the six bases comprising the UMI sequence (minimum PHRED score less than 20) were discarded. The mRNA mate sequences (the second mate of the paired-end-sequencing) were mapped to the human genome, release GRCh38, using STAR 2.5.3a with parameter ”–outSAMmultNmax 1” and default parameters otherwise. Testing the overlap of mapped reads with exons of protein-coding genes and UMI-filtering was performed using custom Python scripts. Droplets (barcodes) were filtered for having a total UMI count of at least 10,000, resulting in a dataset containing UMI counts for 19,865 protein-coding genes across 935 droplets.

Download of 10x Genomics ERCC spike-in expression data

UMI counts for ERCC spike-in RNA processed using the 10x Genomics scRNA-Seq protocol (Zheng et al. 2017) were downloaded from the 10x Genomic website. The dataset consisted of UMI counts for 92 spike-ins across 1,015 droplets.

Download of Drop-Seq ERCC spike-in expression data

UMI counts for ERCC spike-in RNA processed using the 10x Genomics scRNA-Seq protocol (Macosko et al. 2015) were downloaded from GEO accession number GSM1629193. The dataset consisted of UMI counts for 80 spike-ins across 84 droplets.

Download and processing of inDrop pancreatic islet data

Raw sequencing data were downloaded from SRA (experiment accession `SRX1935938`). In this experiment by Baron et al. (2016), inDrop was applied to pancreatic islet tissue from a human donor. Data was processed using the same pipeline used for the inDrop pure RNA data, and only profiles with a total UMI count of at least 1,000, resulting in a dataset containing UMI counts for 19,865 protein-coding genes across 2,109 cells.

Download of Seq-Well PBMC data

UMI counts were downloaded from nature.com (http://www.nature.com/nmeth/journal/v14/n4/extref/nmeth.4179-S2.zip from Gierahn et al. (2017)). The dataset consisted of UMI counts for 6,713 genes (pre-filtered by the authors) across 4,296 cells.

Download and processing of mouse myeloid progenitor data

UMI counts were downloaded from GEO, accession number GSE72857. The 19 clusters for cells are available at `MAGIC`'s (Dijk et al. 2017) code repository: https://github.com/pkathail/magic/issues/34. 27,297 cells with cluster labels were used for performing k-nearest neighbor smoothing (see Algorithm 1), and smoothed values were normalized to UCPM (UMI counts per million). For visualization as a heatmap in Figure 6a-b, the z-score of every gene across cells was calculated. For scatter plots in Figure 6c-e, the expression of each gene was $\log_2(\text{UCPM} + 1)$.

Prediction of scRNA-Seq noise characteristics based on Poisson statistics

In this paper, we initially focus on the technical variation observed in scRNA-Seq data for droplets containing identical pools of pure mRNA. Let u'_{ij} be the observed UMI count for the i'th gene (or ERCC spike-in) in the j'th droplet, for $i = 1, ..., p$ and $j = 1, ..., n$. Similarly, let U'_{ij} be a random variable representing the UMI count for the i'th gene in the j'th cell. We assume that U'_{ij} is Poisson-distributed with mean $\lambda'_{ij} = m_i e_j$, where m_i is the number of mRNA molecules present for the i'th gene, and e_j corresponding to the capture efficiency of the scRNA-Seq protocol for the j'th droplet (both m_i and e_j are unknown). We further assume that $U'_{i1}, ..., U'_{in}$ are independent, for all i. For the sake of simplicity, we assume that the read coverage (the number of reads sequenced per cell) is infinite, so that there are no cases in which a transcript is not observed due to limited read coverage. In practice, limited read coverage will not invalidate the Poisson assumption, but result in lower "effective" capture efficiencies.

If all e_j were identical (say, equal to e^{global}), then $U'_{i1}, ..., U'_{in} \overset{i.i.d}{\sim} \text{Poisson}(\lambda'_i)$, with $\lambda'_i = m_i e^{\text{global}}$. Grün, Kester, and Oudenaarden (2014) have proposed to normalize the expression profile of each cell to the median total UMI count across cells (Model I in Grün et al.), in order to counteract the differences in capture efficiency ("efficiency noise"). Median-normalization consists of calculating the total UMI count per profile (cell or droplet), $t_j = \sum_i u'_{ij}$, calculating the median $t^{\text{med}} = \text{median}\{t_1, ..., t_n\}$, and then multiplying each u'_{ij} by the factor t^{med}/t_j.

Based on the results by Grün et al., we hypothesized that median-normalized data would be approximately Poisson-distributed, as long as the differences in capture efficiency were not too extreme. Therefore, we let $N'_{i1}, ..., N'_{in}$ represent the UMI counts for the i'th gene after median-normalization, and assume them to be i.i.d. $\text{Poisson}(\lambda'_i)$.

For Poisson-distributed variables, the variance is always equal to the expectation (defined by λ). Let $N_i \sim \text{Poisson}(\lambda'_i)$. For the coefficient of variation (CV) of N_i, we have:

$$CV(N_i) = \frac{\sqrt{var(N_i)}}{E(N_i)} = \frac{\sqrt{E(N_i)}}{E(N_i)} = \frac{1}{\sqrt{E(N_i)}} = E(N_i)^{-0.5}$$

Taking the logarithm on both sides gives:

$$\log CV(N_i) = -0.5 * \log E(N_i)$$

Therefore, the relationship between $\log E(N_i)$ and $\log CV(N_i)$ is linear with a slope of -0.5. This is indicated by the gray lines in Figure 1a-f.

The probability of observing a count of zero for N_i is given by the Poisson PMF:

$$f(x) = \frac{\lambda_i^x e^{-\lambda_i}}{x!}$$

Therefore, $P(N_i = 0) = e^{-\lambda_i}$ values are shown as the orange lines in Figure 1g-i.

If a computational pipeline used to determine UMI counts reports systematically inflated values, then the median-normalized UMI counts for the i'th gene can be approximately represented by a scaled Poisson variable $N_i^{\text{inf}} = cN_i'$, where c is the inflation factor. N_i^{inf} then has mean $c\lambda_i'$ and variance $c^2\lambda_i'$, so for $CV(N_i^{\text{inf}})$, we have:

$$CV(N_i^{\text{inf}}) = \frac{\sqrt{var(N_i^{\text{inf}})}}{E(N_i^{\text{inf}})} = \frac{\sqrt{cE(N_i^{\text{inf}})}}{E(N_i^{\text{inf}})} = \sqrt{c}\frac{1}{\sqrt{E(N_i^{\text{inf}})}} = \sqrt{c}E(N_i^{\text{inf}})^{-0.5}$$

Taking the log on both sides gives:

$$\log CV(N_i^{\text{inf}}) = -0.5 \log E(N_i^{\text{inf}}) + 0.5 \log c$$

Therefore, the relationship between $\log E(N_i^{\text{inf}})$ and $\log CV(N_i^{\text{inf}})$ will still be linear, but with an y-axis intercept of $0.5 \log c$ instead of 0, which is consistent with Figure 3b,e.

Prediction of the effect of aggregating scRNA-Seq expression profiles from technical replicates

We again assume that for droplets containing identical pools of pure mRNA, the median-normalized UMI counts $N_{i1}', ..., N_{in}' \overset{i.i.d}{\sim} \text{Poisson}(\lambda_i)$. Let $S_i' = \sum_j N_{ij}'$, and $N_i \sim \text{Poisson}(\lambda_i')$. It is clear that $CV(S_i') = CV(N_i')/\sqrt{n}$:

$$CV(S_i') = \frac{\sqrt{var(S_i')}}{E(S_i')} = \frac{\sqrt{n * var(N_i)}}{nE(N_i)} = \frac{1}{\sqrt{n}}CV(N_i)$$

Similarly, for averaged UMI counts $A_i' = \sum_j N_{ij}/n$:

$$CV(A_i') = \frac{\sqrt{var(A_i')}}{E(A_i')} = \frac{\sqrt{(1/n^2) * var(N_i)}}{E(N_i)} = \frac{1}{\sqrt{n}}CV(N_i)$$

This effect is demonstrated in Figure 2.

Smoothing of scRNA-Seq expression profiles from biological samples based on Poisson statistics

In real data, genes can exhibit differential expression across cells. Therefore, we define $\lambda_{ij} = m_{ij}e_j$, where m_{ij} is the number of mRNA molecules present for the i'th gene in the j'th cell, and e_j is the capture efficiency of the scRNA-Seq protocol for the j'th cell. Let U_{ij} be a random variable representing the UMI count for the i'th gene in the j'th cell. We again assume that U_{ij} is Poisson-distributed with mean λ_{ij}, and that $U_{i1}, ..., U_{in}$ are independent, for all i. Let $\mathcal{Z}_j = \{z_{j1}, ..., z_{jk}\}$ be the set of k nearest neighbors of the j'th cell, as determined in Algorithm 1. Let $\lambda_{ij}^{\text{smooth}} = \lambda_{ij} + \sum_{z \in \mathcal{Z}_j} \lambda_{ij}$. We then define the aggregated expression level $A_{ij} = U_{ij} + \sum_{z \in \mathcal{Z}_|} U_{iz}$, and note that $A_{ij} \sim \text{Poisson}(\lambda_{ij}^{\text{smooth}})$. From the aforementioned discussion, it follows that if the k neighbors have transcriptomes that are sufficiently similar to that of the j'th cell, and if the efficiency noise is not too strong, then $CV(A_{ij}) \approx CV(U_{ij})/\sqrt{k+1}$. Similarly, we can calculate the averaged expression level $S_{ij} = A_{ij}/(k+1)$. Then S_{ij} is a Poisson variable with mean $\lambda_{ij}^{\text{smooth}}$, scaled by a factor of $1/(k+1)$, and therefore has the same CV as A_{ij}. The point here is that even if the U_{ij} are not identically distributed (due to expression differences and/or efficiency noise), simple aggregation or averaging will always result in Poisson-distributed smoothed values. The same is not true for weighted sums or averages. Let $\{w_{j0}, w_{j1}, ..., w_{jk}\}$ represent weights (all positive), and let $W_{ij} = w_{j0}U_{ij} + \sum_{z \in \mathcal{Z}_|} w_{j1}U_{jz}$. Then the weighted sum W_{ij} is neither a Poisson nor a scaled Poisson variable, unless all weights are identical.

ACKNOWLEDGMENTS

We would like to thank Bo Xia, Maayan Baron, and Dr. Gustavo França for helpful discussions.

REFERENCES

Baron, Maayan et al. (2016). “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure”. In: *Cell Systems* 3.4, 346–360.e4. DOI: 10.1016/j.cels.2016.08.011.

Cao, Junyue et al. (2017). “Comprehensive single-cell transcriptional profiling of a multicellular organism”. In: *Science (New York, N.Y.)* 357.6352, pp. 661–667. DOI: 10.1126/science.aam8940.

Dijk, David van et al. (2017). “MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data”. In: *bioRxiv*. DOI: 10.1101/111591.

Fan, Jean et al. (2016). “Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis”. In: *Nature Methods*. DOI: 10.1038/nmeth.3734.

Freeman, Murray F. and John W. Tukey (1950). “Transformations Related to the Angular and the Square Root”. In: *The Annals of Mathematical Statistics* 21.4, pp. 607–611. DOI: 10.1214/aoms/1177729756.

Gierahn, Todd M. et al. (2017). “Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput”. In: *Nature Methods* 14.4, pp. 395–398. DOI: 10.1038/nmeth.4179.

Grün, Dominic, Lennart Kester, and Alexander van Oudenaarden (2014). “Validation of noise models for single-cell transcriptomics”. In: *Nature Methods* 11.6, pp. 637–640. DOI: 10.1038/nmeth.2930.

Hashimshony, Tamar, Naftalie Senderovich, et al. (2016). “CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq”. In: *Genome Biology* 17, p. 77. DOI: 10.1186/s13059-016-0938-8.

Hashimshony, Tamar, Florian Wagner, et al. (2012). “CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification”. In: *Cell Reports* 2.3, pp. 666–673. DOI: 10.1016/j.celrep.2012.08.003.

Islam, Saiful, Una Kjällquist, et al. (2011). “Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq”. In: *Genome Research* 21.7, pp. 1160–1167. DOI: 10.1101/gr.110882.110.

Islam, Saiful, Amit Zeisel, et al. (2014). “Quantitative single-cell RNA-seq with unique molecular identifiers”. In: *Nature Methods* 11.2, pp. 163–166. DOI: 10.1038/nmeth.2772.

Kharchenko, Peter V., Lev Silberstein, and David T. Scadden (2014). “Bayesian approach to single-cell differential expression analysis”. In: *Nature Methods* 11.7, pp. 740–742. DOI: 10.1038/nmeth.2967.

Klein, Allon M. et al. (2015). “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells”. In: *Cell* 161.5, pp. 1187–1201. DOI: 10.1016/j.cell.2015.04.044.

La Manno, Gioele et al. (2017). “RNA velocity in single cells”. In: *bioRxiv*. DOI: 10.1101/206052.

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. DOI: 10.1186/s13059-014-0550-8.

Lun, Aaron T. L., Karsten Bach, and John C. Marioni (2016). “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts”. In: *Genome Biology* 17, p. 75. DOI: 10.1186/s13059-016-0947-7.

Maaten, Laurens van der and Geoffrey Hinton (2008). “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (Nov), pp. 2579–2605.

Macosko, Evan Z. et al. (2015). “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. In: *Cell* 161.5, pp. 1202–1214. DOI: 10.1016/j.cell.2015.05.002.

Paul, Franziska et al. (2015). “Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors”. In: *Cell* 163.7, pp. 1663–1677. DOI: 10.1016/j.cell.2015.11.013.

Pierson, Emma and Christopher Yau (2015). “ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis”. In: *Genome Biology* 16, p. 241. DOI: 10.1186/s13059-015-0805-z.

Risso, Davide et al. (2017). “ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data”. In: *bioRxiv*. DOI: 10.1101/125112.

Rosenberg, Alexander B et al. (2017). “Scaling single cell transcriptomics through split pool barcoding”. In: *bioRxiv*. DOI: 10.1101/105163.

Sasagawa, Yohei et al. (2017). “Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads”. In: *bioRxiv*. DOI: 10.1101/159384.

Shekhar, Karthik et al. (2016). "Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics". In: *Cell* 166.5, 1308–1323.e30. DOI: 10.1016/j.cell.2016.07.054.

Tang, Fuchou et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6.5, pp. 377–382. DOI: 10.1038/nmeth.1315.

Van Der Maaten, Laurens (2014). "Accelerating t-SNE Using Tree-based Algorithms". In: *J. Mach. Learn. Res.* 15.1, pp. 3221–3245.

Wagner, Florian (2015). "GO-PCA: An Unsupervised Method to Explore Gene Expression Data Using Prior Knowledge". In: *PloS One* 10.11, e0143196. DOI: 10.1371/journal.pone.0143196.

Zheng, Grace X. Y. et al. (2017). "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8, p. 14049. DOI: 10.1038/ncomms14049.

Ziegenhain, Christoph et al. (2017). "Comparative Analysis of Single-Cell RNA Sequencing Methods". In: *Molecular Cell* 65.4, 631–643.e4. DOI: 10.1016/j.molcel.2017.01.023.

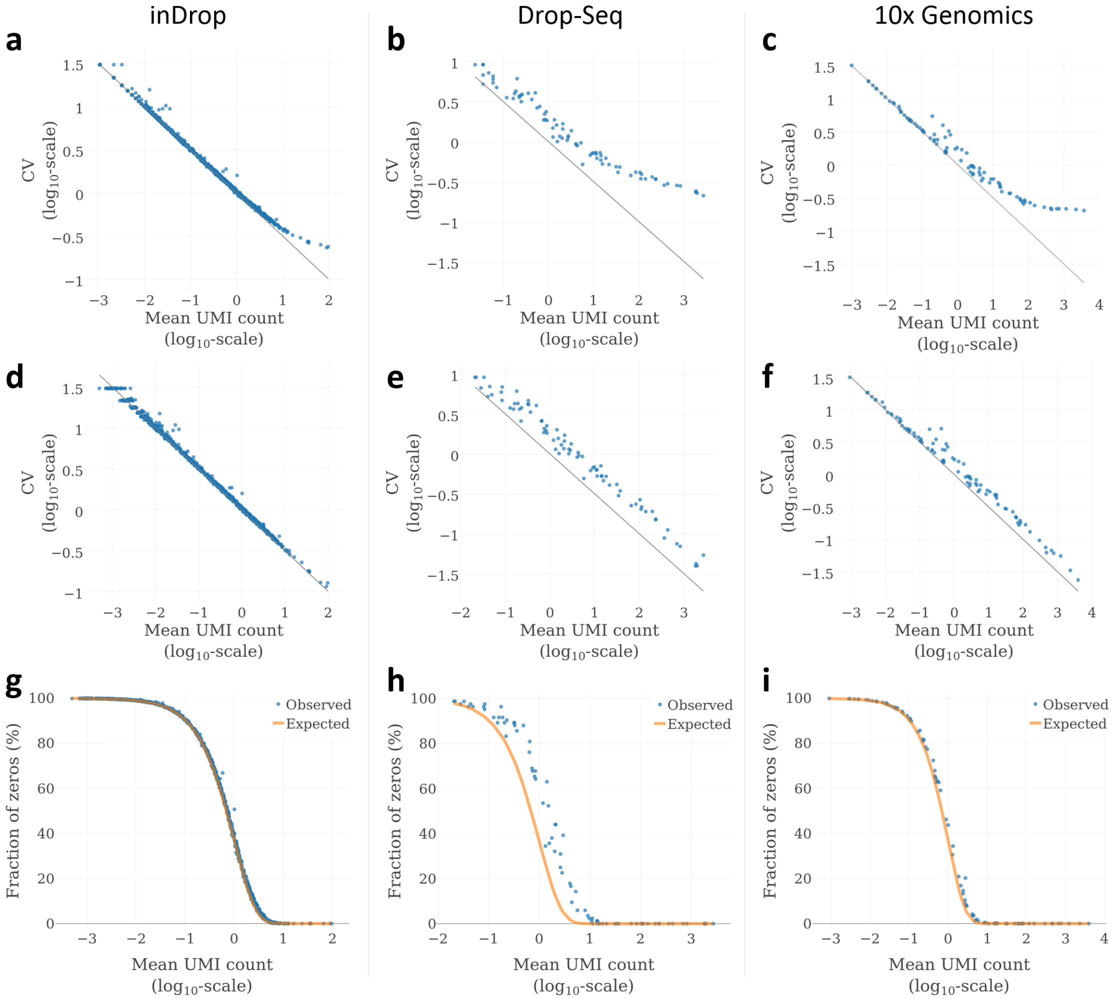

Figure 1. Noise profiles of three high-throughput single-cell RNA-Seq platforms. (**a-c**) Relationship between mean UMI count and coefficient of variation (CV) in pure RNA replicates, analyzed using inDrop (**a**) Drop-seq (**b**), and 10x Genomics (**c**). For inDrop, RNA was extracted from cultered cells (Klein et al. 2015). For Drop-Seq and 10x Genomics, ERCC spike-in RNA was analyzed (see Macosko et al. (2015) and Zheng et al. (2017)). (**d-f**) The same relationship after normalizing each profile to the median transcript count (see Methods). (**g-i**) Expected vs. observed fraction of zeros, as a function of mean expression (after median-normalization). For inDrop data (**a**, **d** and **g**), a randomly sampled subset of 1,000 genes is shown for better readability.

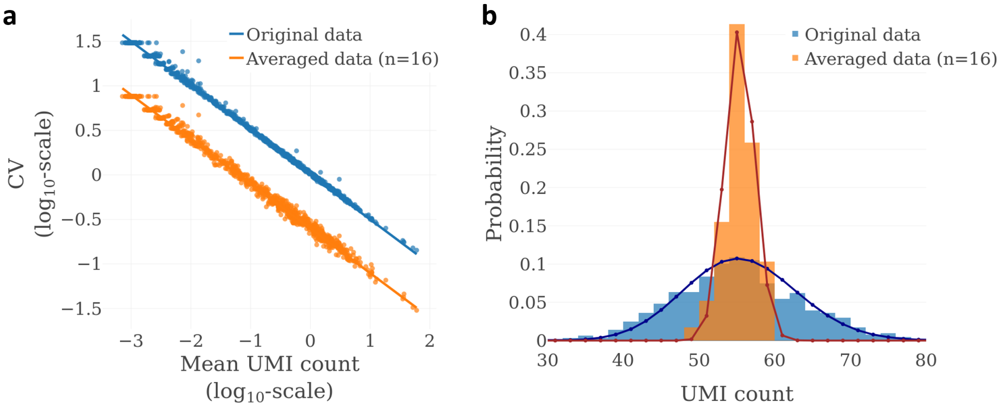

Figure 2. Simple averaging of scRNA-Seq expression profile replicates reduces the coefficient of variation in a manner predicted by Poisson statistics. (**a**) Effect of averaging on the coefficient of variation, for 1,000 randomly selected genes in the inDrop pure RNA dataset (Klein et al., 2015). Solid lines represent the theoretical relationship based on the Poisson distribution. After averaging of 16 profiles at a time, the CV can be seen shifted downwards by about 0.6 units, which corresponds to a factor of 4 on the $\log_{10}$-scale used. (**b**) Distribution of UMI counts for the *GAPDH* gene, before and after averaging. Bars show the observed UMI distributions. The solid lines show the theoretical distributions for a Poisson-distributed variable representing the original values (blue), and a scaled Poisson-distributed variable representing the averaged values (orange). To eliminate efficiency noise, both original and averaged profiles were normalized to the median transcript count (Grün et al., 2014).

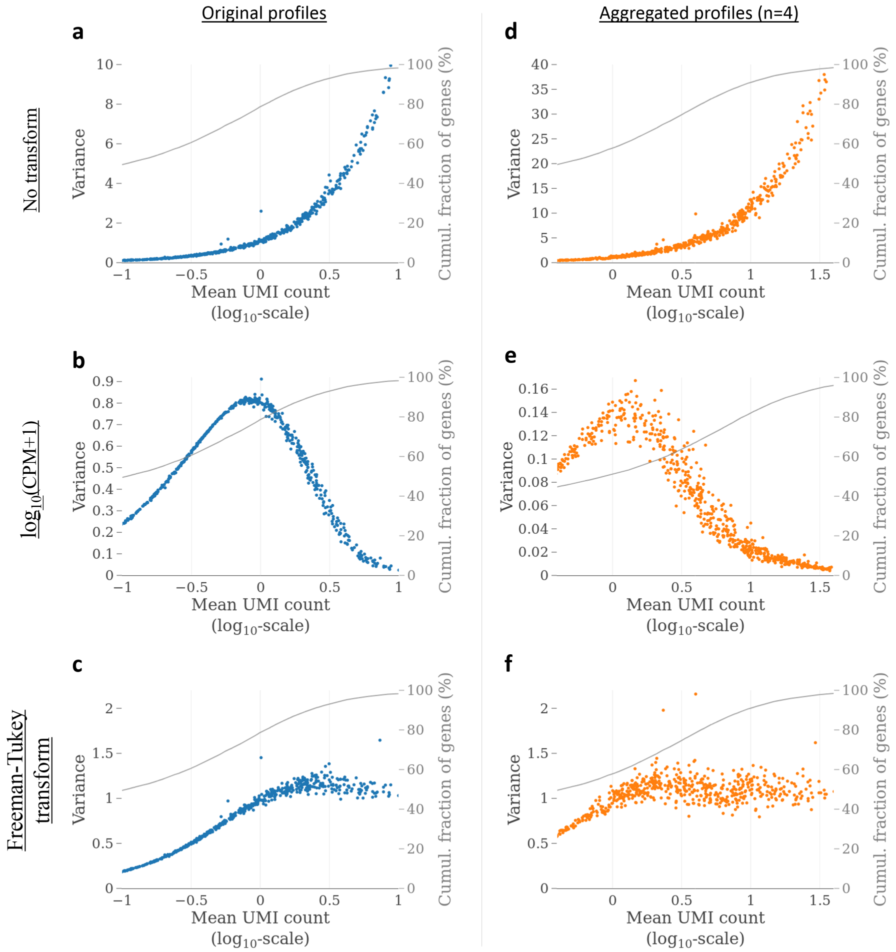

Figure 3. Effect of scRNA-Seq data transformations on mean-variance relationships in technical replicates from the inDrop protocol. (**a-c**) Gene mean-variance relationships in the pure RNA samples (Klein et al., 2015) without transformation, with log(CPM+1) transform, and with Freeman-Tukey transform ($y = \sqrt{x} + \sqrt{x+1}$), respectively. (**d-f**) Mean-variance relationships after aggregating the expression profiles of randomly selected, non-overlapping batches of 4 cells, for the same transformations. All plots show data for the same 1,000 randomly selected genes.

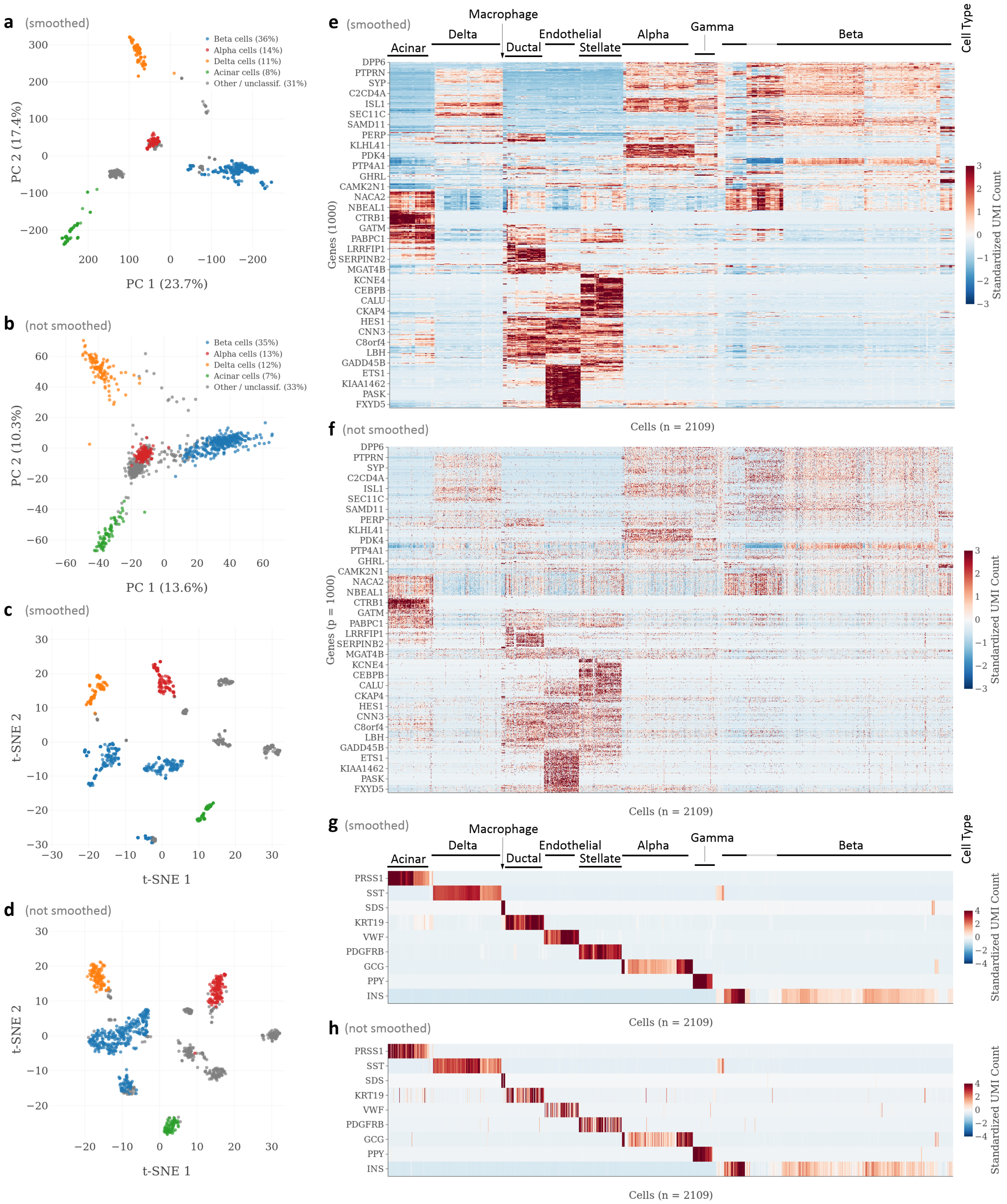

Figure 4. Application of k-nearest neighbor smoothing to scRNA-Seq data from human pancreatic islet tissue. Shown is inDrop data from Baron et al. (2016). Smoothing was performed using $k = 15$. (**a, b**) Principal component analysis (PCA) with (**a**) and without (**b**) smoothing. (**c, d**) t-SNE analysis with (**c**) and without (**d**) smoothing. PCA and t-SNE were performed on Freeman-Tukey transformed (FTT'ed) data of all 19,865 protein-coding genes, and cell types were identified based on ad-hoc expression thresholds for the same marker genes used by Baron et al. (2016). Beta cells were defined as having expression of *INS* $\geq$ 40,000 CPM (UMI counts per million); alpha cells, *GCG* $\geq$ 5,000 CPM; delta cells, *SST* $\geq$ 40,000 CPM; acinar cells, *CPA1* $\geq$ 1,000 CPM. Cells that exceeded none of the thresholds, or more than one, were labeled as "other / unclassified". For t-SNE, the Barnes-Hut algorithm (Van Der Maaten 2014) was applied with perplexity=100 and default parameters otherwise. (**e, f**) Hierarchical clustering of genes and cells with (**e**) and without (**f**) smoothing. Clustering was performed using correlation distance on genes and Euclidean distance on cells, both with average linkage, on smoothed and FTT'ed data, filtered for the 1,000 most variable genes. (**g, h**) Expression of cell type-specific marker genes (Baron et al. 2016) with (**g**) and without (**h**) smoothing. Cells are ordered as in (**e, f**).

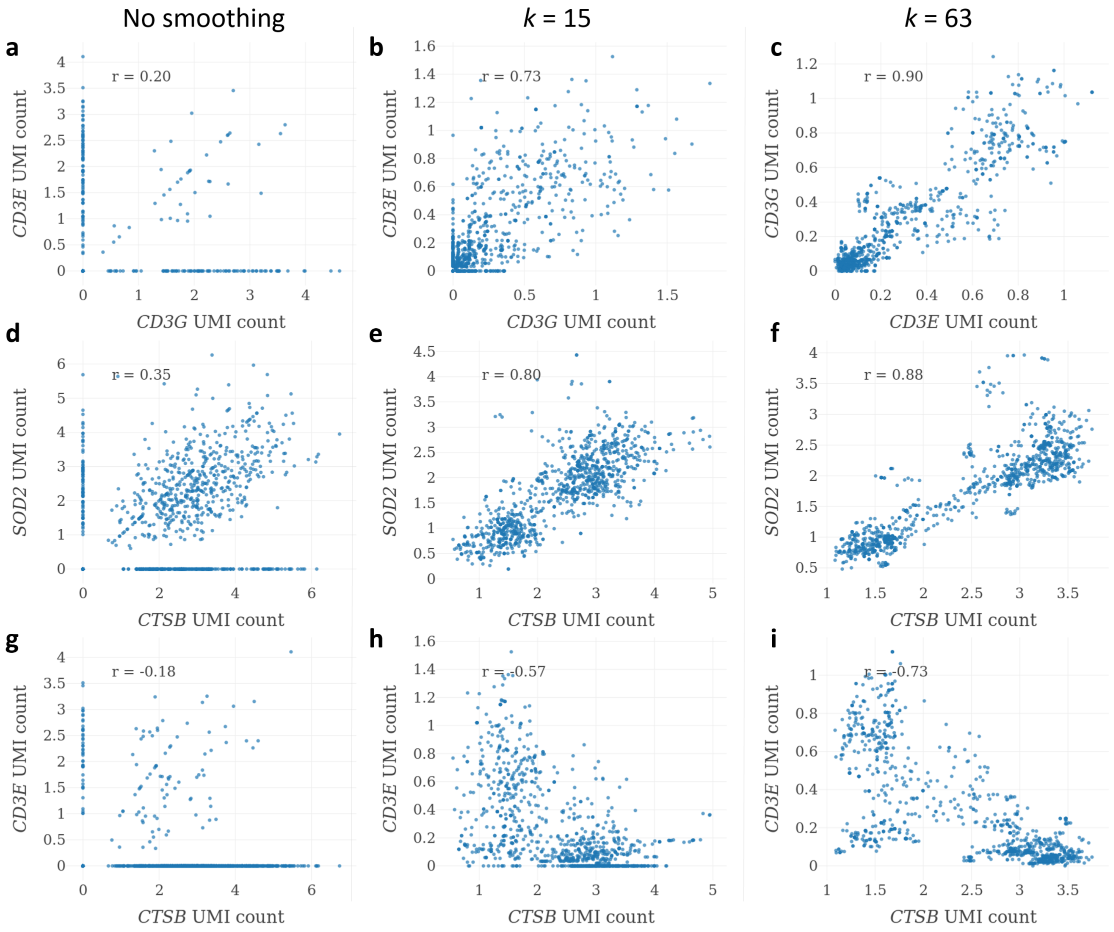

Figure 5. Effect of k-nearest neighbor smoothing on correlations between peripheral T cell and monocyte marker genes. Shown is Seq-Well data from human peripheral blood mononuclear cells (Gierahn et al. 2017). All panels show data for the same randomly selected sample of 1,000 cells (out of 4,296), but smoothing was performed on the full dataset. (**a-c**) Correlations of the T cell receptor genes *CD3G* and *CD3E*, for different degrees of smoothing. (**d-f**) Correlations of *CTSB* and *SOD2* that were used by Gierahn et al. (2017) as monocyte marker genes. (**g-i**) Correlations between *CTSB* and *CD3E*.

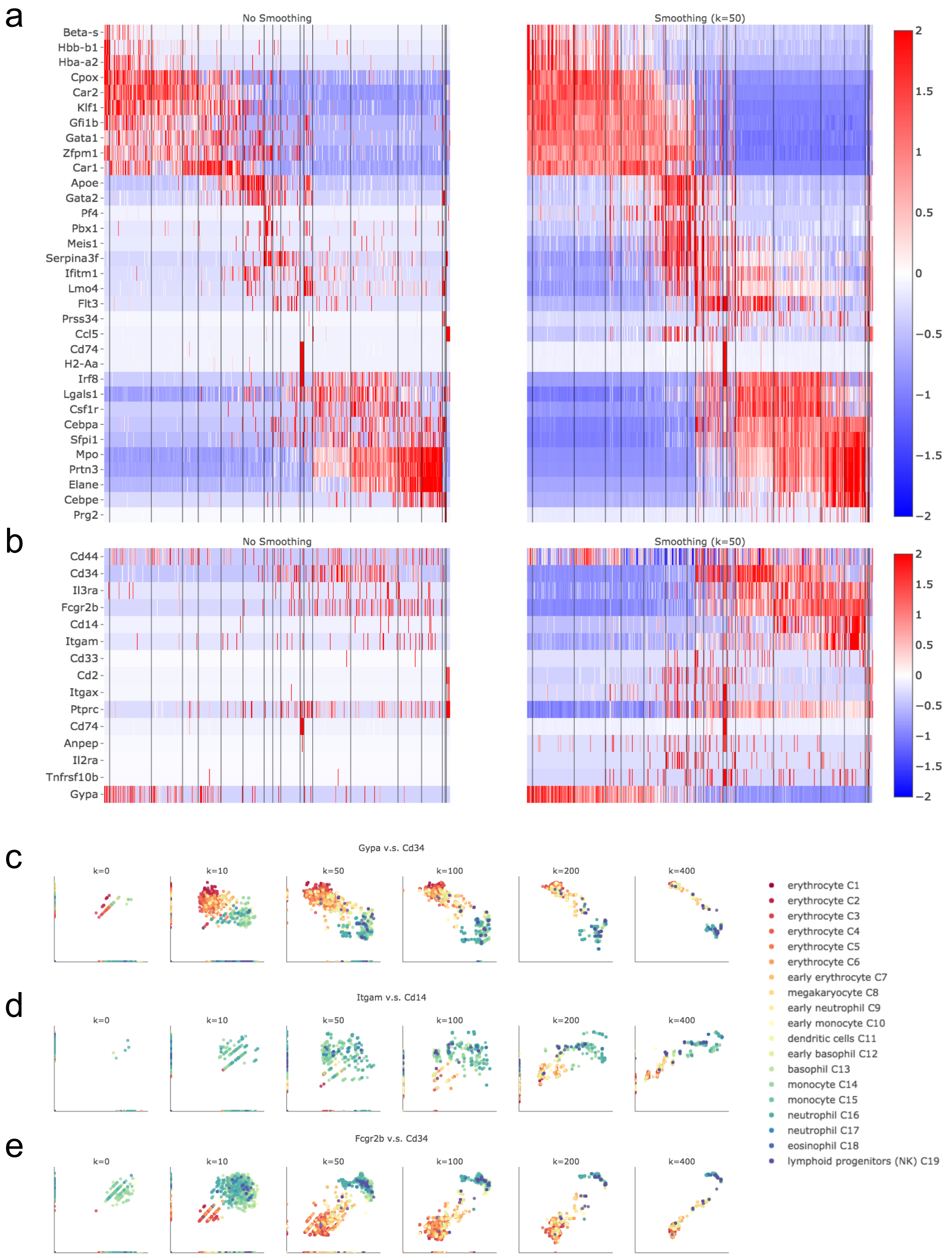

Figure 6. Application of k-nearest neighbor smoothing to scRNA-Seq data of mouse myeloid progenitors. This figure is directly comparable to Figure 3 from Dijk et al. (2017). (**a, b**) Heatmaps of the expression matrices for (**a**) 33 key hematopoietic genes, and **b**) 15 surface marker genes of immune cells, as defined in Paul et al. (2015), before smoothing (left) and after smoothing (right). Gene are ordered as same as shown in Dijk et al. (2017), Figure 3. Cells from left to right are ordered in clusters (C1-C19) as defined in Paul et al. (2015). **c-e**) Scatter plots of expressions showing the recovery of relationships of three pairs of immune marker genes after smoothing with different k (k=0, 10, 50, 100, 200, 400). Each dot is an individual cell colored by the 19 clusters used in **a**. See Methods for details.