Modelling dropouts for feature selection in scRNASeq experiments

Tallulah S. Andrews[1] and Martin Hemberg[1,2]
[1]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK
[2]corresponding Author, mh26@sanger.ac.uk

Abstract

A key challenge of single-cell RNASeq (scRNASeq) is the many genes with zero reads in some cells, but high expression in others. In full-transcript datasets modelling zeros using the Michaelis-Menten equation provides an equal or superior fit to existing scRNASeq datasets compared to other approaches and enables fast and accurate identification of features corresponding to differentially expressed genes without prior identification of cell subpopulations. For datasets tagged with unique molecular identifiers we introduce a depth adjusted negative binomial (DANB) to perform dropout-rate based feature selection. Applying our method to mouse preimplantation embryos revealed clusters corresponding to the inner cell mass and trophectoderm of the blastocyst. Our feature selection method overcomes batch effects to cluster cells from five different datasets by developmental stage rather than experimental origin.

Keywords
single cell RNASeq, feature selection, differential expression, dropouts, modelling

Author Summary

Feature selection is a powerful approach for improving the signal to noise ratio in high dimensional datasets. We present two unsupervised feature selection methods for single-cell RNASeq data which unlike all previous methods are based on dropout rate rather than variance: M3Drop, tailored to full-transcript scRNASeq protocols, and DANB, tailored to data tagged with unique molecular identifiers (UMI). Using differentially expressed genes defined from bulk RNASeq, we perform the first comparison of feature selection quality for both full-transcript and UMI-tagged scRNASeq data. We show that dropout based methods outperform their variance-based counterparts on both real and simulated data due to lower sampling errors. Finally we demonstrate the ability to merge mouse embryo datasets produced using different protocols by different research groups using only the combination of feature selection and library size normalization.

Introduction

Single-cell RNASeq (scRNASeq) has made it possible to analyze the transcriptome from individual cells. In a typical scRNASeq experiment for human or mouse, >10,000 genes will be detected. Most genes, however, are not relevant for understanding the underlying biological processes, and an important computational challenge is to identify the most relevant features. For some well-studied systems one can find the most important genes by searching the literature, but in most situations it would be more desirable to have an unsupervised approach for finding relevant features. However, unsupervised feature selection remains difficult due to the high technical variability and low detection rates of scRNASeq experiments.

In recent years, methods for identifying relevant features from scRNASeq data have been developed. scLVM [1] makes it possible to account for the contribution of distinct processes, e.g. cell-cycle or apoptosis, but the method requires *a priori* knowledge of the sets of genes related to each specific process. By contrast, the highly variable genes (HVG) method [2] is unsupervised as it automatically identifies the set of genes that have a higher degree of variability than expected based on the observed technical noise. HVG has a tendency to select many of very lowly expressed genes, which are unlikely to have reliable expression measurements [2]. Furthermore, the biological interpretation and relevance of HVGs remains incompletely understood, and a more straightforward concept is that of differentially expressed (DE) genes. However, differential expression methods, such as SCDE [3], require the predetermined homogeneous subpopulations. Applying these methods to scRNASeq data is often challenging due to the difficulty of identifying homogenous subpopulations.

Another approach to feature selection is the use of dimensionality reduction methods, e.g. principal component analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) [4], e.g [5–10], that extract a small number of new meta-features. However, these methods are often sensitive to systematic noise such as batch effects due to the large number of genes subject to technical noise relative to the number of genes influenced by biological effects in many datasets [11,12]. In addition, the biological interpretation of extracted meta-features is non-trivial.

A particularly challenging aspect of scRNASeq data is the presence of “dropouts”, i.e. genes that are not detected in some cells but highly expressed in others. We introduce two novel methods of feature selection based on the relationship between dropout-rate and mean expression across genes specific for full-transcript and UMI-tagged datasets. We compare our novel methods to existing method in the first systematic comparison of feature selection methods for single-cell RNASeq. Our new methods outperform their variance-based counterparts for detecting differentially expressed genes in real and simulated datasets. We show that feature selection methods are consistent across datasets and reduce batch effects within and across scRNASeq datasets enabling novel biological insights.

Results

Michaelis-Menten Modelling of Dropouts (M3Drop) fits full length transcript data

The results from a scRNASeq experiment can be represented as an expression matrix, where each row represents a gene and each column represents a cell. Two common approaches have emerged: deep sequencing of full-transcripts for a relatively small number of cells [1,6,13–18], and high-cell number, low-depth sequencing of 3’ or 5’ ends of transcripts tagged with unique molecular identifiers [9,10,12,19–21]. These two approaches produce data with very different characteristics, for instance UMI tagged data generally has a strong correlation between total counts and number of detected genes which is not evident in full-transcript data (**Figure S1)**, and thus different computational methods may be required for the analysis [2,19,20]. Feature

selection typically involves looking for genes with high variance relative to other genes based on some metric. The most salient characteristic of scRNASeq experiments, however, is the presence of a large number of zero values (i.e. dropout events). A typical scRNA-seq experiment has ~50% dropouts (**Table 1**) [22] the frequency of zero values (dropout rate) relative to the mean expression of the gene could equally be used. Dropout rate is subject to less sampling noise than variance (**Figure S2**) [23], and thus may be more accurate for small single-cell datasets.

It has been suggested that the large number of dropouts is due to transcripts being lost during the library preparation [3]. Based on studies of RT-qPCR assays [24,25], we hypothesize that the main reason for dropouts is due to failure of the reverse transcription (RT). Hence we model the relationship between dropout rate and mean expression using a Michaelis-Menten function:

$$P_{dropout} = 1 - \frac{S}{K_M + S}$$

where S is the average expression of the gene across all cells and K_M is the Michaelis constant, and $P_{dropout}$ represents the probability that the quantity of cDNA reaches some experiment-specific threshold of detection in any cell.

We compared the fit of the Michaelis-Menten model of dropouts (M3Drop) to two alternative models [3,26] across six full-transcript scRNASeq datasets (**Figure 1A, S3**). M3Drop fit as well or better than the alternatives for all datasets. The logistic model performs similarly to M3Drop because the M3Drop is a special case of the logistic model with a coefficient of one, however logistic regression is flattened by noise in the data resulting in coefficients less than one. Importantly due to the simplicity of the M3Drop model we can test the significance of identified features (**see Methods**), whereas there is no such procedure for the logistic model [3].

Depth-adjusted Negative Binomial fits UMI-tagged data

While M3Drop fits full-transcript scRNASeq datasets well, data quantified using unique molecular identifiers (UMI) to obtain raw transcript counts often does not fit the Michaelis-Menten function (**Figure 1C**). Thus we developed a different model specifically for UMI-tagged data. It has been previously shown that UMI-tagged data follows a negative binomial distribution [19,20]. However, due to the wide range of detection rates across cells (**Figure S1**) previous works have normalized counts prior to fitting the negative binomial. We introduce an alternative model which incorporates cell-specific detection rates into the negative binomial model. We refer to this approach as the Depth-Adjusted Negative Binomial (DANB) model. Unlike the estimation of sequencing efficiency presented by Grun et al. [20], the DANB model is not reliant on the inclusion of spike-in RNAs which are incompatible with some popular UMI-tagging scRNASeq platforms such as Drop-seq [9].

DANB models each observation (x_{ij}) with a negative binomial model with mean μ_{ij} equal to the product of the relative expression level of gene *j* (m_j), the relative detection rate for cell *i* (d_i), and the total molecules (*T*) in the dataset.

$$\mu_{ij} = m_j * d_i * T$$

Fitting the observed total gene- (t_j) and cell-specific molecule counts (t_i) results in an estimate of the mean given by:

$$\hat{\mu}_{ij} = \frac{t_j * t_i}{T}$$

Gene-specific dispersions were fit to the observed sample variances (Methods). We compared the fit of the DANB model to a basic negative binomial model fit to normalized data using dropout rates since they were not included directly in the fitting procedure (**Figure 1B, S4**). The depth-adjusted model fits all three UMI-tagged datasets better than the basic model.

Feature Selection

Feature selection is frequently performed to reduce dimensionality and noise in single-cell RNASeq datasets [1,6,8,13,27]. Unlike differential expression, feature selection is agnostic to biological groups, making it a more difficult problem. One approach to feature selection is to use dimensionality reduction methods, such as principal component analysis (PCA), to extract a small number of new features which optimally represent some characteristic of the original data, in the case of PCA this characteristic is variance. However, biological interpretation of these features is difficult and they are sensitive to batch effects and other technical noise because they combine information from a large number of genes [11,12]. An alternative approach, is to identify genes which are responding to the biological situation under examination. For instance, Brennecke et al. [2] identify genes with variation beyond the technical noise. Existing methods of feature selection exclusively rely on various measures of variance such as coefficient of variation squared [2,20], or the Gini Index [28]. In contrast, we consider the utility of dropout rate for feature selection since it suffers from less sampling noise thus should be more reliable than variance for datasets containing fewer than 1,000 cells (**Figure S2**).

To test feature selection methods on single-cell RNASeq data, we used two experiments where bulk RNASeq was available in addition to scRNASeq data and single-cell samples were relatively homogeneous [12,18]. The first examined human iPSCs derived from three different healthy donors [12]. The second considered mouse ESCs grown in different culture conditions [18]. Both of these datasets represent relatively homogeneous cell-cultures, thus differential expression between conditions is the most important source of biological variation in the single-cell experiment and differential expression between the bulk RNASeq experiments can be used as a ground truth to test feature selection methods. Importantly, these datasets use

very different scRNASeq protocols. Tung et al. [12] contains 3'UMI-tagged data; whereas Kolodziejczyk et al. [18] contains Smartseq full-transcript data, allowing us to compare methods for different data-types.

We compared nine different feature selection methods (**Methods**), each of which was used to rank genes from most to least biologically important and compared to the ground truth from the respective bulk-RNASeq data. These methods included ranking genes by the magnitude of their loadings in principal component analysis (PCA), by the strength of their most negative gene-gene correlation (Cor), by their relative Gini index (Gini), and how much of an outlier they were from the M3Drop dropouts-mean expression curve (M3Drop), the squared coefficient of variation (CV2) vs mean expression relationship (HVG) [2], the dispersion-mean expression relationship fit by DANB (NBDisp), or the dropouts-mean expression relationship fit by DANB (NBDrop). We did not include methods which require data to have predefined groups, whether from unsupervised clustering or *a priori* information, since the performance of these differential expression methods depends greatly on the quality/reliability of the predefined groups.

For the UMI-tagged data, PCA and correlation methods performed the best as measured by overall area under the ROC curve (AUC = 0.7), which summarizes false positive rates and true positive rates into a single quality score, and by the number of true positives amongst the top 2,000 ranked genes, i.e. precision (**Figure 2Ai,Bi**). In contrast, these were among the worst performing methods on full-transcript data (**Figure 2Aii,Bii**). M3Drop followed by the two DANB methods performed the best on full-transcript data and the latter were third and fifth when tested on UMI-tagged data (AUC = 0.69 and 0.65).

Interestingly, when the top 2,000 features were considered our novel dropout-based methods outperformed their variance based counterparts for both data-types, NBDrop > NBDisp and M3Drop > HVG (**Figure 2Bi, 2Bii, S5**). Since the goal of feature selection is typically to identify the most salient genes this comparison may be more relevant to actual usage of these methods than overall AUC. One potential explanation for why dropout-based methods perform better is the high sampling noise of estimates of variance compared to dropout-rates (**Figure S2**) [23].

To test the performance of feature selection methods for identifying differentially variable (DV), or generally highly variable (HV) genes, in addition to differentially expressed (DE) genes, we simulated data using either a depth-adjusted negative binomial model or a M3Drop-based zero-inflated negative binomial (**see Methods**). DV and HV genes are not detectable using bulk RNASeq and are generally poorly understood. Thus, no ground-truth is available for real scRNASeq datasets. We fit each model to three different UMI-tagged or full-transcript datasets respectively (**Figure 2C**). These simulations recapitulated many features of scRNASeq datasets (**Figure S6,S7**).

Patterns observed in real data were generally recapitulated in the DE simulated data (**Figure 2**). M3Drop performed much better on full-transcript data than on UMI-tagged data, whereas the Gini index performed better on UMI-tagged data than full-transcript. Two interesting differences

between simulated and real scRNASeq data were that gene-gene correlations and negative binomial dropouts were nearly tied for best performing method in simulations of both data-types. In addition, PCA performance is driven by the first principal component on simulated data as opposed to the second and third components in real data and performs better on full-transcript than UMI-tagged simulated data. Both discrepancies may be explained by the lack of technical noise in the simulated data compared to observed data, particularly for the full-transcript data which is known to be more prone to technical noise than UMI-tagged data [19].

In contrast to DE genes, differentially variable genes were not detected by correlations in either data-type. In addition, PCA was unable to detect DV gene in the UMI-tagged dataset, despite performing well on full-transcript data. The two DANB-based methods, and HVG performed similarly to each other and across data-types. Whereas the Gini index only performed well on UMI-tagged data; and M3Drop only performed well on full-transcript data. Identification of globally highly variable (HV) genes followed the same pattern as DV only with larger differences between methods, with the exception of PCA which did not capture HV genes in either datasets (**Figure 2**).

Feature selection makes it possible to overcome batch effects

Single-cell RNASeq is very sensitive to batch effects, which include various technical confounders between experimental replicates or between datasets such as sequencing efficiency, reagent quality, etc… [22,29]. One of the goals of feature selection is to reduce the impact of this technical noise on downstream analysis by focusing on only the most biologically relevant genes. If features capture the biologically relevant aspects of the data they should be consistent across datasets examining the same biological system, whereas features that are primarily affected by technical noise are expected to vary.

We tested if feature selection can help overcome batch effects by considering the reproducibility of the top 2,000 features across five datasets examining pre-implantation mouse embryos (**Table 1**). Datasets ranged in size from 17 cells to 255 cells covering the development from zygote to blastocyst.

All the feature selection methods show significant overlaps between the different developmental datasets, but variance-based methods are less reproducible than dropout-based methods (**Figure 3A**). In agreement with the full-transcript simulations (**Figure 2 Cii**), gene-gene correlations, M3Drop and NBDrop were the most reproducible methods. In addition, when the five datasets were combined and visualized using PCA, all feature selection methods greatly improved the clustering of cells by developmental time-point rather than by dataset of origin (**Figure 3 B,C, S8**). We quantified this removal by comparing clusterings of the data (Ward's hierarchical) following feature selection to the known sampling timepoints, adjusted rand index (ARI). M3Drop and NBDisp were tied for best performance with an ARI of 0.76 when cut to produce five clusters (**Figure S9**).

Examining the most reproducible genes across these five datasets we found Anxa2 and S100a10, which are known to form a hetero-tetramer in the endometrium involved in implantation [30,31]. Anxa2 was among the top 2,000 features from both M3Drop and NBDrop in four of the five datasets; and S100a10 was identified in all five datasets. The expression of these genes was significantly correlated starting from the 4 cell stage ($r=0.32$, $p = 0.0007$) increasing in strength in the 8- and 16- cell stage ($r = 0.43$, $p = 2x10^{-7}$) until diverging in the blastocyst to have a roughly linear relationship at the single-cell level in the trophectoderm and returning to being uncorrelated within inner-cell mass cells, $r= 0.49$, $p < 10^{-10}$ (**Figure 3 D-F**).

As a final demonstration of the importance of feature selection we considered the 133 single blastocyst cells profiled by Deng et al. [17]. The blastocyst develops following the first differentiation of embryonic cells into the trophectoderm (TE), which will form the placenta, and the inner-cell mass (ICM) which develops into the fetus. Clustering the blastocyst cells following M3Drop feature selection clearly indicates the presence of two cell-type which we identify as the TE and ICM using marker genes[32–34] (**Figure 4A**). We used M3Drop for this dataset since that model fit the data the best (**Figure 1C**). In contrast, clustering the cells without performing feature selection suggests anywhere from two to five cell-types of mixed identities (**Figure 4B**). Only M3Drop, DANB and HVG are able to distinguish ICM and TE in this dataset (**Figure S13**).

Discussion

We present two new models for single-cell RNASeq data and demonstrate their utility for feature selection using both real and simulated datasets. Unique among feature selection methods, these algorithms use dropout-rates as opposed to variance to evaluate biological relevance. M3Drop uses a Michaelis-Menten equation to model the zeros in full-transcript scRNASeq. The other model, DANB, extends the standard negative binomial model to account for differing sequencing depths or tagging efficiency of UMI-tagged scRNASeq data. These models fit the respective datasets better than available alternatives (**Figure 1**). Feature selection using these models outperforms existing methods on both real and simulated datasets (**Figure 2**). Furthermore, this feature selection can overcome batch-effects between datasets (**Figure 3**) and reduces technical noise within datasets (**Figure 4**) to reveal underlying biological processes.

We compared the performance of multiple feature selection methods on both real and simulated data from both full-transcript and UMI-tagged data (**Figure 2, 3A**). Overall, gene-gene correlations and our method for identifying genes with unusually high dropout rates based on a depth-adjusted negative binomial model were the most accurate methods for unsupervised feature selection of differentially expressed genes. Gene-gene correlations cannot detect DV or HV genes, because differences in variability does not produce the coordinated increase/decreases in expression detected by this method. In addition, gene-gene correlations are far more computationally intensive than the other methods, which will become important as datasets grow to tens of thousands of cells **(Figure S10)** [9,35].

Importantly, we show for the first time that generalizable computational methods can have significant differences in performance for different scRNASeq protocols (**Figure 2**). For instance the Gini Index consistently performs better on UMI-tagged data; whereas our M3Drop method consistently performs better on full-transcript data. This highlights the importance of tailoring models to account for differences between scRNASeq protocols. We expect that this effect is not limited to feature selection, but that the choice of protocol will affect many statistical analyses of scRNASeq data. Thus it is important to consider the type of data each method was designed to handle when choosing an analysis pipeline for scRNASeq datasets.

Contrasting variance-based and dropout-based feature selection methods, we found that the latter generally performed better. On full-transcript data, M3Drop outperformed HVG [2] on every measure with the exception of the identification of highly variable genes. This is not surprising since HVG is designed to detect such features. Yet the biological significance of generally highly variable genes remains unclear. On UMI-tagged data fit to our DANB model, detecting genes with high dropout rates (NBDrop) again outperformed detecting genes with high fitted dispersions (NBDisp) by most measures considered.

The main reason why variance-based methods perform worse is because estimates of variance are very sensitive to sampling noise, particularly for highly skewed distributions such as those that dominated single-cell RNASeq datasets [23]. This is a key finding since the most commonly used feature selection criterion for scRNASeq data is high gene expression variability, either in HVG [1,6,8,13,18,27] or in PCA [5–10]. This is particularly important for datasets of fewer than 1,000 cells (**Figure S2**).

Finally we demonstrated the importance of feature selection in reducing technical noise and batch effects within and across datasets (**Figure 3, 4**). This is important since batch effect correction is still an open question in scRNASeq. Unlike many other methods [1,20,36], M3Drop and NBDrop do not rely on technical spike-ins. Furthermore batch-effect correction can often drastically change the distribution of expression values, for instance regression-based methods may introduce negative values [12,37]. Such alterations are likely to violate the assumptions of downstream analyses such as differential expression methods or pseudotime inference [3,38]. In contrast, the feature selection methods presented here neither rely on technical spike-ins nor change the observed expression values.

Conclusion

We introduce two novel models of single-cell RNASeq data. We model dropout rates in full-transcript data using a Michaelis-Menten function (M3Drop). Whereas, for UMI-tagged data we extend the negative binomial model to account for sequencing depth/tagging efficiency (DANB). Using these models we show that feature selection based on identifying genes with unusually high dropout rates outperforms methods looking for highly variable genes on real and simulated data. Finally we demonstrate the benefits of feature selection for revealing biological signals otherwise hidden beneath technical noise.

Methods

Negative Binomial Models

Negative binomial models have been shown to fit normalized molecule counts from single-cell RNASeq data employing unique molecular identifiers, referred to as UMI-tagged data [19,20] . We introduce a modification of this model which we call the Depth-Adjusted Negative Binomial model (DANB) which eliminates the need to normalize molecule counts by explicitly modelling the tagging/sequencing efficiency for each cell. The DANB models each observation as drawn from a negative binomial model with mean equal to :

$$\hat{\mu}_{ij} = \frac{\sum_i t_{ij} * \sum_j t_{ij}}{\sum_{ij} t_{ij}} \tag{1}$$

where t_{ij} is the total molecules observed for gene *j* in cell *i*. The sum over cells is a measure of tagging efficiency, and the sum over genes divided by the total molecules in the dataset is a measure of the gene's relative expression level. Gene-specific dispersions (r_j) are fit by minimizing the difference between the sum of variances of each observation and sample variance of the difference between each observation and its respective expected value, μ_{ij}.

$$\arg\min_{r_j} |\sum_i ((t_{ij} - \mu_{ij})^2 - (\mu_{ij} + r_j \mu_{ij}^2))| \tag{2}$$

Genes with Poissonian behavior, which results in infinite dispersion, were assigned a maximum dispersion of 10^{10}. We use the sum of absolute differences between observed gene-specific dropout-rates (frequency of zeros) to the expectation for DANB to evaluate the fit of the model. For the basic negative binomial model we normalized molecule counts to the median t_i prior to fitting which results in a single negative binomial model fit to the expression of each gene.

Zero-Inflation Models

We fit three different models to the relationship between mean expression and dropout rate. The expression of each gene was averaged across all cells including those with zero reads for a particular gene (*S*). Dropout rate was calculated as the proportion of cells with zero reads for that gene ($P_{dropout}$). The three models were fit using these values. The Michaelis-Menten equation [39]:

$$P_{dropout} = 1 - \frac{S}{K_M + S} \tag{3}$$

was fitted using maximum likelihood estimation as implemented by the `mle2` function in the `bbmle` R package. The logistic regression [3]:

$$P_{dropout} = \frac{1}{1 + e^{-(a+b*\log(S))}} \quad (4)$$

was fitted using the glm R function. The double exponential model [26] was fit by log transforming the equation then using the lm R function to fit the coefficient to the resulting quadratic model:

$$P_{dropout} = e^{\lambda * S^2} \quad (5)$$

$$\ln(P_{dropout}) = \lambda * S^2 \quad (6)$$

We also considered a double MM model:

$$P_{dropout} = 1 - \frac{S^2}{(K_1 + S)(K_2 + S)} \quad (7)$$

where K_1, K_2 > 0. For all datasets except Zeisel, however, K_1 was almost identical to K_M while K_2 was < 0.001 (**Table S1**).

Feature Selection Methods

High Dropout Rate (M3Drop)

Rearranging the Michaelis-Menten equation (3) to solve for the Michaelis-Menten constant K gives:

$$K = \frac{P * S}{1 - P} \quad (8)$$

This is used to calculate the value specific to each gene, K_j; the variance for each K_j estimate was calculated using error propagation rules to combine errors on observed S and P:

$$\sigma_{K_j} = K_j * \sqrt{\left(\frac{\sigma_S}{S}\right)^2 + \left(\frac{\sigma_P}{P}\right)^2} \quad (9)$$

Where σ_S is the sample standard deviation of S and σ_P is the sample standard deviation of P. The K_j's were assumed to be log-normally distributed and we tested each one against the global K_M that was fit to the entire dataset using a one-sided Z-test:

$$Z = \frac{log(K_j) - log(K_M)}{\sqrt{\sigma^2_{log(K_j)} + \sigma^2_{log(K_M)}}} \quad (10)$$

$\sigma^2_{log(K_M)}$ was estimated as the standard error of the residuals and added to $\sigma^2_{log(K_j)}$

$$\sigma_{\log(K_M)} = \frac{sd(\log(K_j) - \log(K_M))}{\sqrt{N}} \tag{11}$$

$$\sigma_{\log(K_j)} = \log(K_j) - \log(K_j - \sigma_{Kj}) \tag{12}$$

Genes were ranked in order of decreasing significance (increasing p-value) of the Z-test. Since all computations are based on gene-level statistics the method scales linearly with the size of the expression matrix.

Highly Variable Genes (HVG)

Highly Variable Genes (HVG) were calculated using the published method [2]. Briefly, this involved fitting an inverse function to the relationship between the squared coefficient of variation and mean expression across genes:

$$CV2 = a_0 + a_1/\mu \tag{13}$$

A chi-squared test is used to determine the significance of high variance genes relative to this relationship. The original method fit equation (13) to spike-in RNAs which we refer to as HVG-E, but since many single-cell datasets do not include spike-ins we also consider an alternative which fit equation (13) to using all endogenous genes (HVG). Genes were ranked in order of decreasing significance (increasing p-value) of the chi-squared-test.

Depth-Adjusted Negative Binomial (NBDisp, NBDrop)

We identified features using both high variance (NBDisp) and high dropout rate (NBDrop) compared to the expectation based on the depth-adjusted negative binomial model. Gene-specific dispersions were fit as described as above. We fit power-law relationship between dispersion and mean expression using all genes with mean expression above 16 counts/cell and for which a finite dispersion was fit. High variance genes were identified using the log-residuals from the power-law relationship. High dropout rate genes were identified by comparing the observed frequency of dropouts with those expected from a negative binomial model with dispersion from the fitted power-law and depth-adjusted means. Significance was evaluated using the normal-approximation of the binomial distribution, and genes were ranked in order of decreasing significance (increasing p-value).

Principal Component Analysis (PCA)

Sparse matrix principal component analysis was performed on the log-transformed normalized expression matrix using the irlba R package as described in the monocle vignette (**Supplementary code**). Genes were ranked by the sum of absolute loadings for a set of

principal components. We considered both the top 3 components and just the second and third components since the first principal component is often dominated by technical variation [11,40].

Gini Index (Gini)

The Gini index evaluates the degree of inequality of values from a frequency distribution. It has a value of zero if all values are equal and a value of 1 if all expression comes from a single cell. Gene expression was normalized using counts-per-million, then the Gini index was calculated for each gene using the reldist R package. Gini index was strongly correlated with the dropout rate of genes (**Figure S11**). Hence we used the residuals from a linear regression of Gini index on dropout rate to rank genes from highest residual to lowest. This is similar to how the GiniIndex is used in GiniClust [28].

Correlation Strength (Cor)

In a mixed population of cells genes that are differentially expressed between conditions will be strongly correlated with each other. Thus, correlations for genes changing in the same direction will be positive; whereas genes changing in opposite directions will be negatively correlated. Many technical confounding factors such as batch effects or reads multimapping between homologous genes can lead to spurious positive correlations between genes. Hence we normalized the data using counts per million and then we ranked genes by the magnitude of their strongest negative pearson correlation with other genes. While often used in other data rich fields, we are not aware of any scRNASeq studies or methods which employ correlation-based feature selection.

Accuracy using bulk RNASeq ground truth

To evaluate the accuracy of feature selection methods we use two published datasets for which bulk RNASeq data was available in addition to single-cell data. For both datasets, the cell populations are relatively homogeneous. Tung et al. (2017) [12] considered iPSCs from three different individuals and performed three replicates of UMI-tagged scRNASeq and three replicates of bulk RNASeq for each. Read/UMI counts were obtained from the Gene Expression Omnibus (GSE77288).

For Kolodziejczyk et al. (2015)[18] we considered ESCs grown under two conditions: alternative 2i and serum for which there were three replicates of scRNASeq and two replicates of bulk RNASeq. Full-transcript single-cell RNASeq read counts were obtained from ArrayExpress (E-MTAB-2600). Raw reads for bulk samples were obtained from ArrayExpress and mapped to GRCm38 using STAR [41] and gene level read counts were obtained using featureCounts [42].

The ground truth differentially expressed genes were obtained for each dataset using three standard analysis methods: DESeq2 [43], edgeR [44], and limma-voom [45]. Genes identified as differentially expressed using all three methods at 5% FDR were considered ground-truth

positives. Genes not identified as DE by any of the methods at 20% FDR were considered ground-truth negatives. This resulted in 1,915 positives, and 8,398 negatives for the iPSCs; and 709 positives and 11,278 negatives for the ESCs (**Figure S12**).

Accuracy of feature selection methods was evaluated using the area under the receiver operator curve (AUC) using the RCOR R package, and the number of DE genes among the top 2000 ranked genes.

Single-cell RNASeq datasets

We considered thirteen public scRNASeq datasets (**Table 1**). These were chosen to reflect a range of different dataset sizes, sequencing methods and cell-types. Datasets where the expression matrix consisted of raw read counts (or UMI counts) were normalized using counts per million except for the Depth-adjusted negative binomial model-based methods. Quality control was performed prior to all analyses as follows. First, all genes annotated as processed pseudogenes in Ensembl (version 80) were removed and cells with fewer than 2000 detected genes were removed. Genes detected in fewer than 4 cells or with average normalized expression $< 10^{-5}$ were excluded from consideration. For the Deng data, only single mouse embryo cells analyzed using the SmartSeq protocol were considered to avoid technical artefacts. To facilitate the identification of true positive DE genes only the two replicates of Unstimulated and after 4h LPS stimulation were considered for the Shalek dataset; in addition technical artefacts as described by the authors were removed [15].

Simulated datasets

We simulated UMI-tagged data using the depth-adjusted negative binomial model fit to one of the three UMI-tagged datasets, Tung [12], Zeisel [21] and Klein [10]. Mean gene expression levels were drawn from a log-normal distribution fit to the respective dataset. Cell-specific sequencing efficiency was drawn from a gamma distribution. Finally, gene-specific dispersions were calculated from the mean expression level using the power-law relationship fit to the respective dataset. Each simulated dataset contained 25,000 genes and 500 cells.

We simulated full-transcript data using a zero-inflated negative binomial model fit to each of three full-transcript datasets, Pollen [7], Buettner [1], or Kolodziejczyk [18]. As before, mean gene expression levels were drawn from a log-normal distribution and gene-specific dispersions were calculated from the mean expression level using a power-law relationship. Simulated expression values were inflated with zeros according mean expression using the Michaelis-Menten equation fit to the respective dataset. Since full-transcript data is generally obtained from fewer cells than UMI-tagged data, each simulated full-transcript dataset contained 200 cells and 25,000 genes.

Differentially expressed (DE) genes were added by increasing/decreasing the mean expression of each gene in half the cells by a log-base 2 fold change draw from a normal distribution with

mean = 0 and sd = 2. Dispersions were adjusted in the differentially expressing cells according to the fitted power-law relationship. Alternatively differentially variable (DV) genes were added by increasing/decreasing the dispersion of the negative binomial model for half the cells by a similarly distributed log2 fold change. Finally highly variable (HV) genes were added in the same way as DV genes, except that all cells exhibited the changed dispersion. Three simulation replicates of DE, DV and HV genes were performed for each of the six datasets.

Genes with a greater than 5-fold increase or decrease in mean expression/dispersion in half the simulated cells were considered ground truth DE/DV genes respectively. For HV genes only genes with a greater than 5-fold increase in dispersion were considered ground truth positives. In all cases genes with less than absolute 1-fold change in mean or dispersion were considered unchanged.

Analysis of Developmental Datasets

Reproducibility

We considered five full-transcript single-cell RNASeq datasets examining mouse embryonic development from fertilization to blastocyst (**Table 1**). Any genes not detected in all five datasets were removed, and the top 2000 ranked genes for each feature selection method were compared across datasets. The expected number of genes identified in precisely *n* datasets was calculated as:

$$\binom{5}{n}(p)^n(1-p)^{5-n}$$

where *p* is the proportion of genes identified by each method i.e. *p* = 2,000/11,285 = 0.18. This was summarized across all *n* using a chi-square statistic :

$$\sum_n \frac{(obs-exp)^2}{exp}$$

Principal component analysis (PCA) was performed on log base two transformed counts per million (or transcripts per million if expression was reported as FPKMs/RPKMs) after adding a pseudocount of one. For each feature selection method the normalized and transformed expression matrix was restricted to genes identified in at least three out of five datasets before performing the PCA. Pearson correlations between Anxa2 and S100a10 were calculated using the log-transformed expression values.

Identification of TE and ICM

Unidentified blastocyst samples were assigned to ICM or TE by rescaling log transformed expression values to Z-scores. The average score over three ICM markers (Sox2, Pou5f1/Oct4, Nanog) and three TE markers (Elf5, Eomes, Cdx2) were calculated and cells assigned to the type for which they had a higher score. Finally, 133 blastocyst cells from the Deng dataset [17]

were clustered using Ward's hierarchical clustering [46] before and after M3Drop feature selection.

Code/Data Availability

Dataset accession codes are listed in **Table 1**.
R packages used: ROCR (version 1.0-7), gplots (version 3.0.0), bbmle (version 1.0.18), reldist (version 1.6-6), irlba (version 2.1.2), monocle (version 2.2.0)
M3Drop and DANB are freely available on github (contains code for highly variably gene detection) : https://github.com/tallulandrews/M3Drop

Author Contributions

TA and MH conceived of the project and wrote the manuscript. TA developed the method, produced the code, analyzed and interpreted the data.

Acknowledgements

The authors would like to thank: Vladimir Kiselev, Davis McCarthy, Simon Andrews, and Tomislav Ilicic for their comments and suggestions for improving this manuscript.

Competing Financial Interests

The authors declare they have no competing interests.

Tables

Table 1: Publicly available single-cell RNASeq datasets.

Dataset	Cell-types	Labels	Method	N	% Zero	Source
Tung	Human iPSCs	Donor individual	5' Seq UMIs -> (CPM)	768	49%	[12] GSE77288
Klein	Mouse ESC	Differentiation timepoint	CEL-Seq UMIs -> (CPM)	2448	63%	[10] GSE65525
Zeisel	Mouse brain	BackSPIN clustering	5' Seq UMIs -> (CPM)	2542	77%	[21] GSE60361
Usoskin	Mouse neurons	PCA-based clustering	5' Seq Counts -> (CPM)	530	73%	[5] GSE59739
Shalek	Mouse bone marrow	Stimulated & unstimulated	Smartseq FPKMs	173	66%	[15] GSE48968
Pollen	Human cell lines & tissues	Cell line identity	Smartseq FPKMs	301	60%	[7] SRP041736**
Buettner	Mouse ESC	Cell-cycle stage	Smartseq Counts -> (CPM)	279	51%	[1] E-MTAB-2805
Kolodziejczk (Kolo)	Mouse ESCs	Growth media	Smartseq Counts -> (CPM)	406	51%	[18] E-MTAB-2600
Deng	Mouse embryos	Development timepoint	Smartseq Counts -> (CPM)	255	50%	[17] GSE45719
Biase	Mouse embryos	Development stage	Smartseq FPKMs	56	38%	[47] GSE57249

Fan	Mouse embryos	Development stage	SUPeR-seq FPKMs	69	46%	[48] GSE53386
Xue	Mouse embryos	Development stage	Poly-A selected RPKMs	17	30%	[16] GSE44183
Goolam	Mouse embryos	Development stage	SmartSeq2 Counts -> (CPM)	124	43%	[14] E-MTAB-3321

**UMI = Unique Molecular Identifier; FPKM = fragments per kilobase per million; CPM = count per million*
*** Processed data was provided by the authors*

Figure Captions

Figure 1: Proposed models fit observed data better than alternatives. (A) The Michaelis-Menten (solid black), logistic (dashed purple), and double exponential (dotted blue) models are fit to the Deng dataset [17] (full-transcript). Expression (counts per million) was averaged across all cells for each gene (points) and the proportion of expression values that were zero was calculated. ERCC spike-ins are shown as open black circles. (B) The depth-adjusted negative binomial (DANB) and simple negative binomial were fit to the Zeisel dataset [21] (UMI-tagged). (C) Michaelis-Menten (M3Drop) consistently fits full-transcript datasets (Kolo, Shalek, Deng, Buettner, Pollen, Biase) despite having fewer parameters than the logistic regression, whereas DANB consistently fits UMI-tagged datasets (Tung, Klein, Zeisel). Error is the sum of absolute error of expected gene-specific dropouts.

Figure 2: Quality of feature selection methods. Nine feature selection methods were tested on real scRNASeq data for which bulk data was also available using overall area under the ROC curve (A), and number of true positives among the top 2,000 genes (B). Tung [12] and Kolodziejczyk [18] were the UMI-tagged and full-transcript datasets respectively. (C) Feature selection methods were tested on simulated from a zero-inflated negative binomial model and DANB model fit to three full-transcript and UMI-tagged datasets respectively.

Figure 3: Reproducibility of feature selection across developmental datasets. (A) The difference between observed and expected overlaps between features identified in exactly 1-5 developmental datasets. Methods with high values to the right of the plot were reproducible across most of the datasets. Negative values to the left indicate few genes were unique to a single dataset. Chi-squared statistics for each method are provided. All datasets used full-transcript sequencing with fewer than 10 detected spike-ins, thus HVG-E was not considered. (B & C) principal component analysis of the combined datasets using all genes (C) or restricted to only genes identified by M3Drop in 3 or more datasets, for the other feature selection methods see **Figure S8,S9** (B). (D-F) Combining datasets reveals an increasing correlation between Anxa2 and S100a10 as cells differentiate. Blastocyst cells were assigned to ICM (blue) or TE (green) using marker genes. Different datasets are indicated using symbols as in (B).

Figure 4: **Clustering blastocyst cells before and after feature selection** (A) Hierarchical clustering of cells using only genes identified by M3Drop in 3 or more datasets reveals two cell-types corresponding to ICM (blue) and TE (green) as shown by expression of marker genes. Whereas, when all genes are used, ICM and TE do not cleanly separate into two clusters (B). Black dots indicate mean expression over all blastocyst cells. Expression is measured as log2 CPMs. Equivalent plots for other feature selection methods are found in **Figure S13.**

Figures

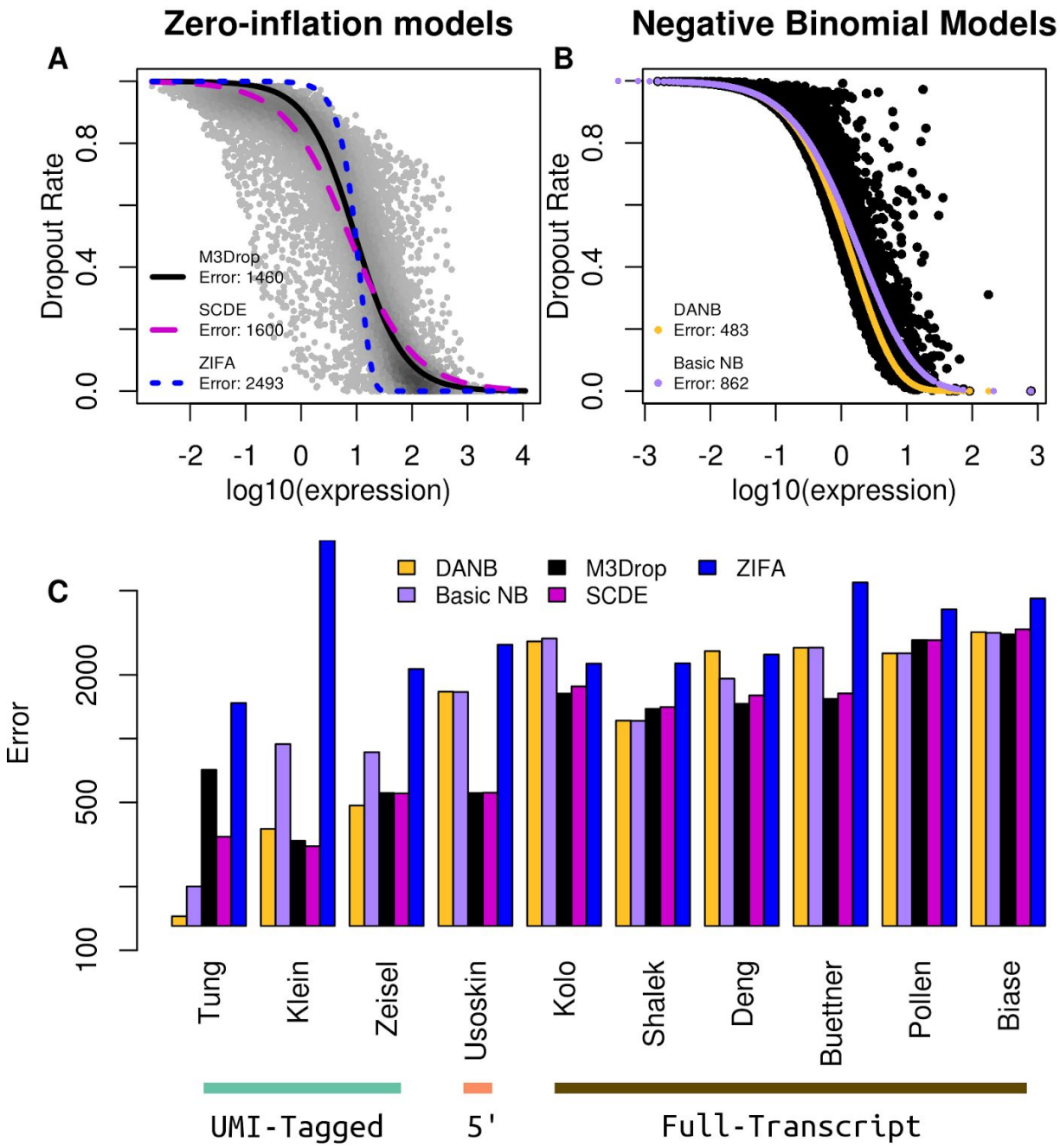
Zero-inflation models
Negative Binomial Models
A
B
C
Dropout Rate
log10(expression)
M3Drop
Error: 1460
SCDE
Error: 1600
ZIFA
Error: 2493
DANB
Error: 483
Basic NB
Error: 862
Error
DANB
Basic NB
M3Drop
SCDE
ZIFA
Tung
Klein
Zeisel
Usoskin
Kolo
Shalek
Deng
Buettner
Pollen
Biase
UMI-Tagged
5'
Full-Transcript

Figure 1

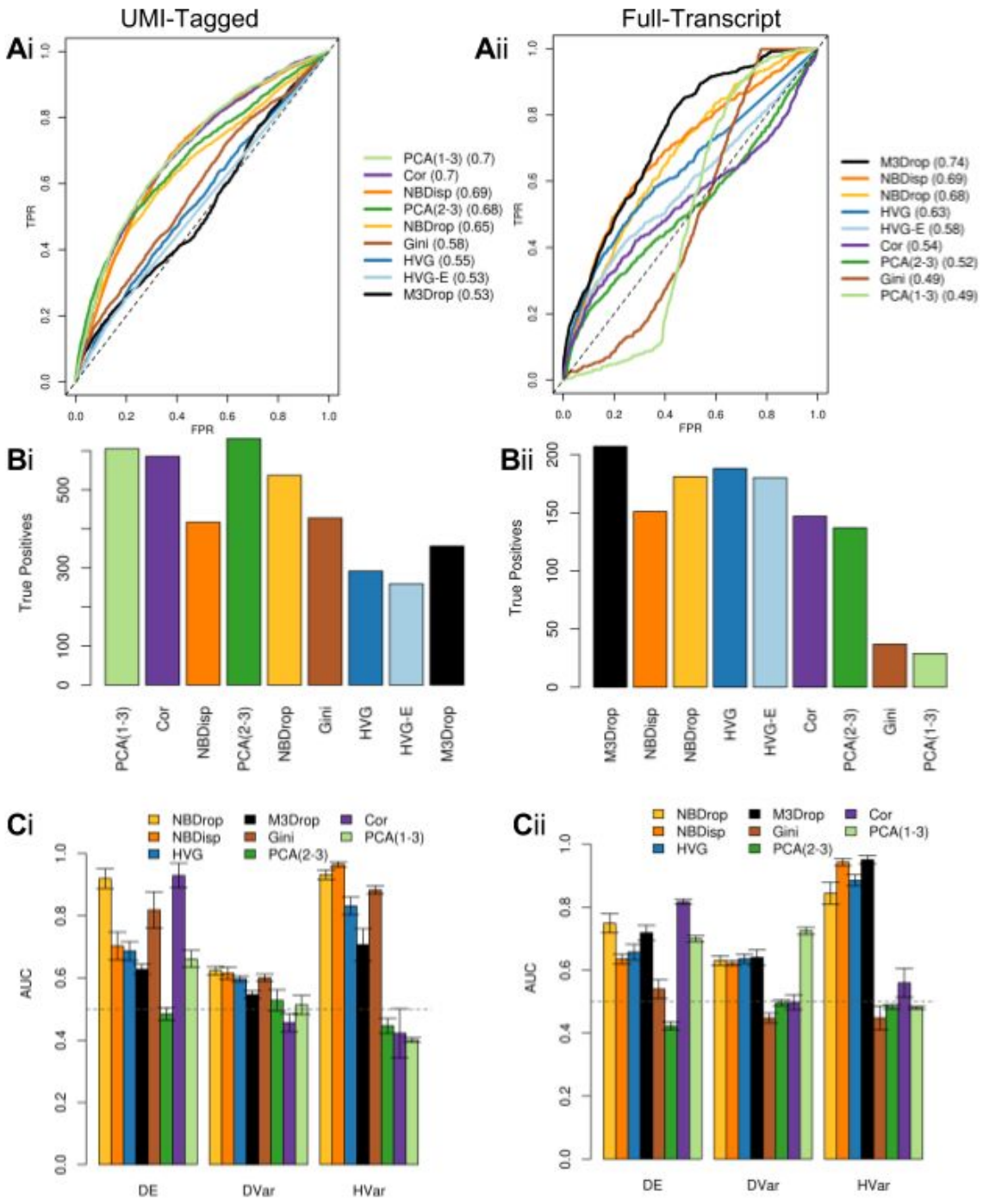
UMI-Tagged
Ai
TPR
FPR
PCA(1-3) (0.7)
Cor (0.7)
NBDisp (0.69)
PCA(2-3) (0.68)
NBDrop (0.65)
Gini (0.58)
HVG (0.55)
HVG-E (0.53)
M3Drop (0.53)
Full-Transcript
Aii
M3Drop (0.74)
NBDisp (0.69)
NBDrop (0.68)
HVG (0.63)
HVG-E (0.58)
Cor (0.54)
PCA(2-3) (0.52)
Gini (0.49)
PCA(1-3) (0.49)
Bi
True Positives
PCA(1-3)
Cor
NBDisp
PCA(2-3)
NBDrop
Gini
HVG
HVG-E
M3Drop
Bii
True Positives
M3Drop
NBDisp
NBDrop
HVG
HVG-E
Cor
PCA(2-3)
Gini
PCA(1-3)
Ci
NBDrop
NBDisp
HVG
M3Drop
Gini
PCA(2-3)
Cor
PCA(1-3)
AUC
DE
DVar
HVar
Cii
NBDrop
NBDisp
HVG
M3Drop
Gini
PCA(2-3)
Cor
PCA(1-3)
AUC
DE
DVar
HVar

Figure 2

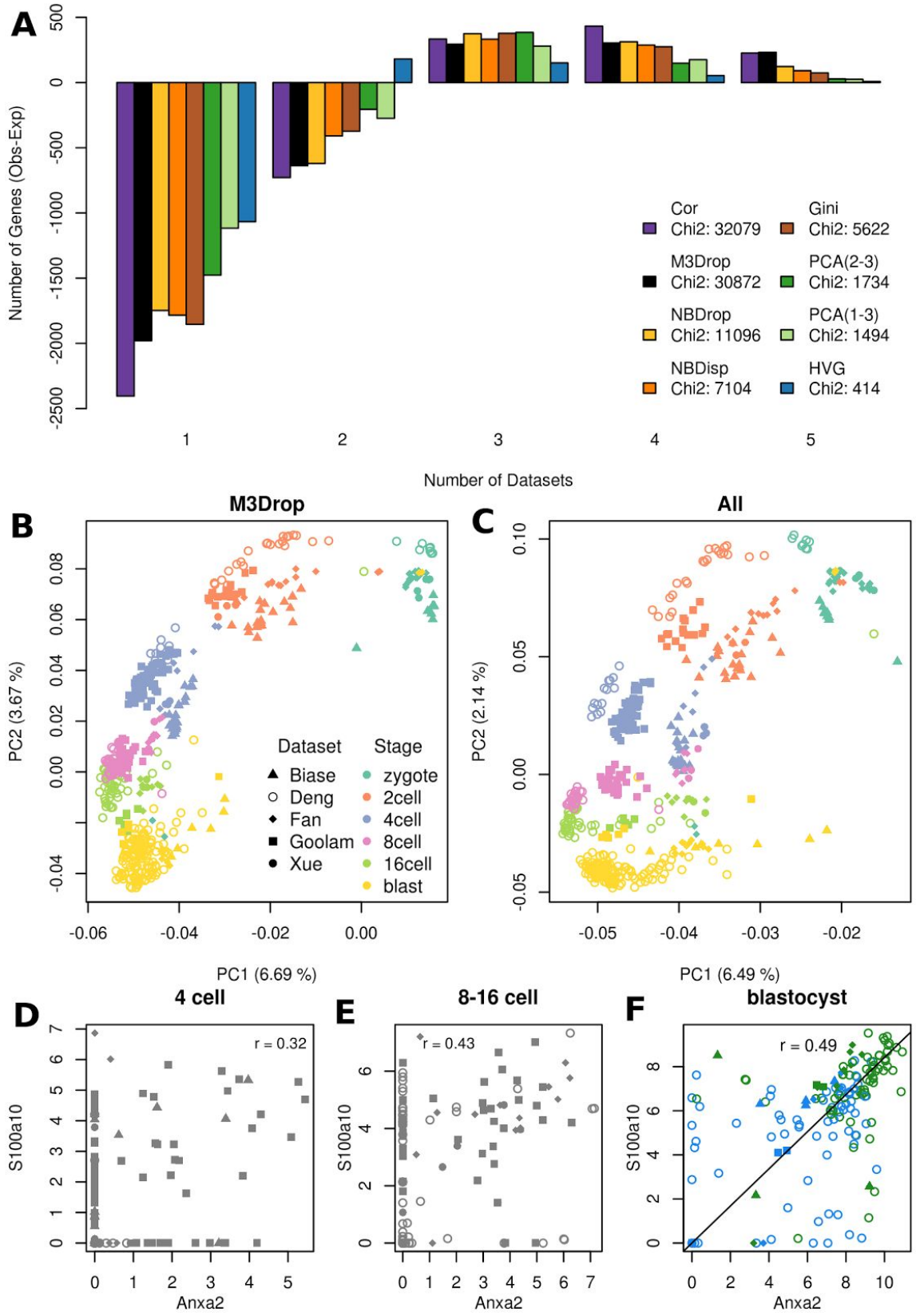
A
Number of Genes (Obs-Exp)
Number of Datasets
Cor Chi2: 32079
Gini Chi2: 5622
M3Drop Chi2: 30872
PCA(2-3) Chi2: 1734
NBDrop Chi2: 11096
PCA(1-3) Chi2: 1494
NBDisp Chi2: 7104
HVG Chi2: 414
B
M3Drop
PC2 (3.67 %)
PC1 (6.69 %)
Dataset
Biase
Deng
Fan
Goolam
Xue
Stage
zygote
2cell
4cell
8cell
16cell
blast
C
All
PC2 (2.14 %)
PC1 (6.49 %)
D
4 cell
r = 0.32
S100a10
Anxa2
E
8-16 cell
r = 0.43
S100a10
Anxa2
F
blastocyst
r = 0.49
S100a10
Anxa2

Figure 3

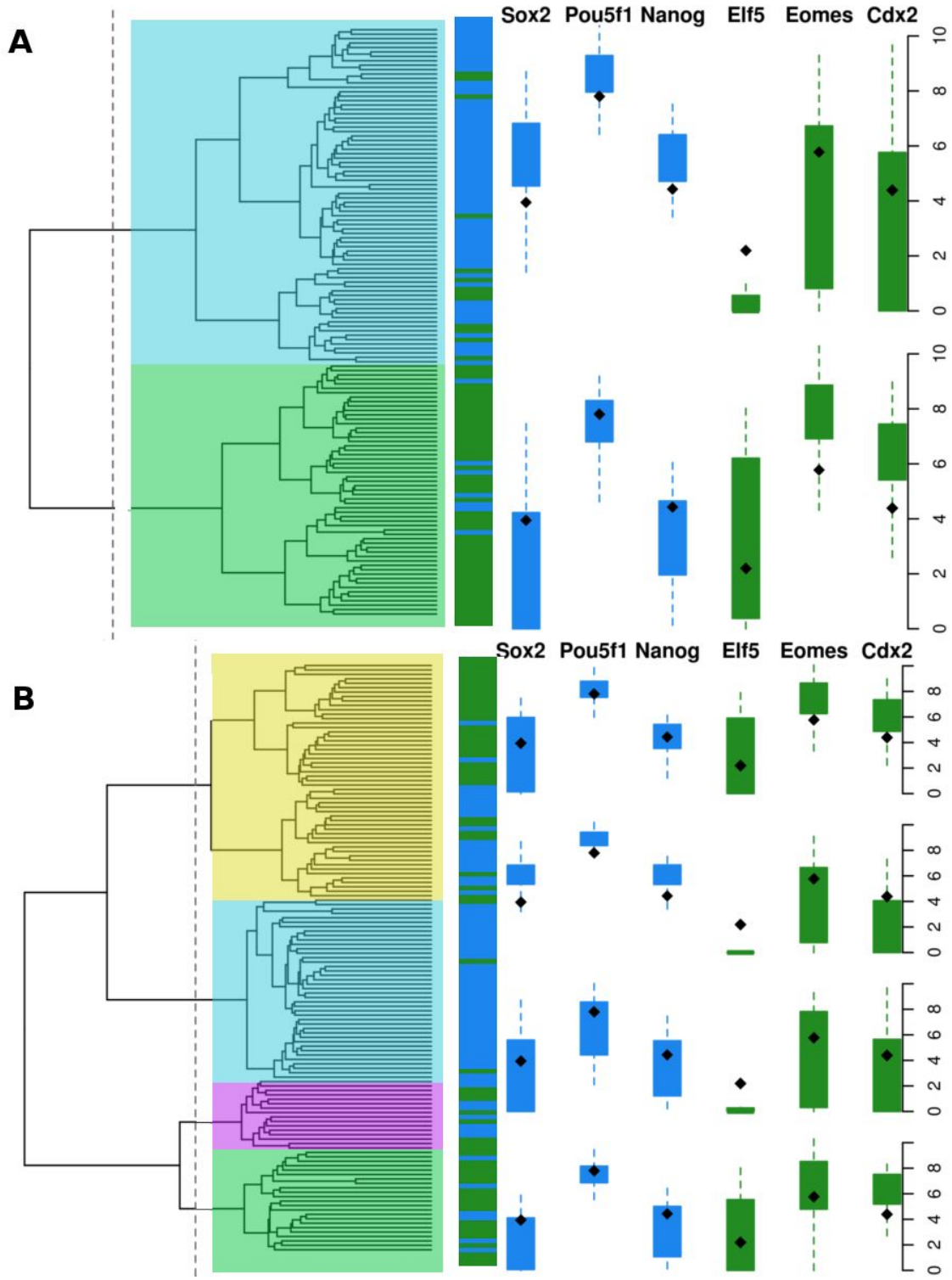
A
Sox2
Pou5f1
Nanog
Elf5
Eomes
Cdx2
B
Sox2
Pou5f1
Nanog
Elf5
Eomes
Cdx2

Figure 4

References

1. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol 33: 155–160. doi:10.1038/nbt.3102.
2. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 10: 1093–1095. doi:10.1038/nmeth.2645.
3. Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. Nat Methods 11: 740–742. doi:10.1038/nmeth.2967.
4. Maaten, Laurens van der, Hinton G (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research. Available: http://jmlr.org/papers/v9/vandermaaten08a.html.
5. Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, et al. (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci 18: 145–153. doi:10.1038/nn.3881.
6. Wilson NK, Kent DG, Buettner F, Shehata M, Macaulay IC, et al. (2015) Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell 16: 712–724. doi:10.1016/j.stem.2015.04.004.
7. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, et al. (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol 32: 1053–1058. doi:10.1038/nbt.2967.
8. Björklund ÅK, Forkel M, Picelli S, Konya V, Theorell J, et al. (2016) The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing. Nat Immunol 17: 451–460. doi:10.1038/ni.3368.
9. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161: 1202–1214. doi:10.1016/j.cell.2015.05.002.
10. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, et al. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161: 1187–1201. doi:10.1016/j.cell.2015.04.044.

11. Hicks SC, Teng M, Irizarry RA (2015) On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. BioRxiv. doi:10.1101/025528.

12. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, et al. (2017) Batch effects and the effective design of single-cell gene expression studies. Sci Rep 7: 39921. doi:10.1038/srep39921.

13. Tsang JCH, Yu Y, Burke S, Buettner F, Wang C, et al. (2015) Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. Genome Biol 16: 178. doi:10.1186/s13059-015-0739-5.

14. Goolam M, Scialdone A, Graham SJL, Macaulay IC, Jedrusik A, et al. (2016) Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. Cell 165: 61–74. doi:10.1016/j.cell.2016.01.047.

15. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, et al. (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature 510: 363–369. doi:10.1038/nature13437.

16. Xue Z, Huang K, Cai C, Cai L, Jiang C, et al. (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500: 593–597. doi:10.1038/nature12364.

17. Deng Q, Ramsköld D, Reinius B, Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343: 193–196. doi:10.1126/science.1245316.

18. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, et al. (2015) Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell 17: 471–485. doi:10.1016/j.stem.2015.09.011.

19. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, et al. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 11: 163–166. doi:10.1038/nmeth.2772.

20. Grün D, Kester L, van Oudenaarden A (2014) Validation of noise models for single-cell transcriptomics. Nat Methods 11: 637–640. doi:10.1038/nmeth.2930.

21. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, et al. (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347: 1138–1142. doi:10.1126/science.aaa1934.

22. Bacher R, Kendziorski C (2016) Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol 17: 63. doi:10.1186/s13059-016-0927-y.

23. Rose C, Smith MD (2002) Mathematical Statistics with Mathematica. New York: Springer-Verlag.

24. Bengtsson M, Hemberg M, Rorsman P, Ståhlberg A (2008) Quantification of mRNA

in single cells and modelling of RT-qPCR induced noise. BMC Mol Biol 9: 63. doi:10.1186/1471-2199-9-63.

25. Reiter M, Kirchner B, Müller H, Holzhauer C, Mann W, et al. (2011) Quantification noise in single cell experiments. Nucleic Acids Res 39: e124. doi:10.1093/nar/gkr505.

26. Pierson E, Yau C (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 16: 241. doi:10.1186/s13059-015-0805-z.

27. Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, et al. (2016) Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol 17: 67. doi:10.1186/s13059-016-0932-1.

28. Jiang L, Chen H, Pinello L, Yuan G-C (2016) GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol 17: 144. doi:10.1186/s13059-016-1010-4.

29. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, et al. (2016) Power Analysis of Single Cell RNA-Sequencing Experiments. BioRxiv. doi:10.1101/073692.

30. Garrido-Gómez T, Dominguez F, Quiñonero A, Estella C, Vilella F, et al. (2012) Annexin A2 is critical for embryo adhesiveness to the human endometrium by RhoA activation through F-actin regulation. FASEB J 26: 3715–3727. doi:10.1096/fj.12-204008.

31. Bharadwaj A, Bydoun M, Holloway R, Waisman D (2013) Annexin A2 heterotetramer: structure and function. Int J Mol Sci 14: 6259–6305. doi:10.3390/ijms14036259.

32. Marikawa Y, Alarcón VB (2009) Establishment of trophectoderm and inner cell mass lineages in the mouse embryo. Mol Reprod Dev 76: 1019–1032. doi:10.1002/mrd.21057.

33. Chen Y, Wang K, Gong YG, Khoo SK, Leach R (2013) Roles of CDX2 and EOMES in human induced trophoblast progenitor cells. Biochem Biophys Res Commun 431: 197–202. doi:10.1016/j.bbrc.2012.12.135.

34. Hamatani T, Carter MG, Sharov AA, Ko MSH (2004) Dynamics of global gene expression changes during mouse preimplantation development. Dev Cell 6: 117–131. doi:10.1016/S1534-5807(03)00373-3.

35. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8: 14049. doi:10.1038/ncomms14049.

36. Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS Comput Biol 11: e1004333. doi:10.1371/journal.pcbi.1004333.

37. Risso D, Ngai J, Speed TP, Dudoit S (2014) Normalization of RNA-seq data using

factor analysis of control genes or samples. Nat Biotechnol 32: 896–902. doi:10.1038/nbt.2931.

38. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, et al. (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 16: 278. doi:10.1186/s13059-015-0844-5.

39. Michaelis L, Menten ML (1913) Die Kinetik der Invertinwirkung. Biochem Z 49: 333–369.

40. Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 16: 133–145. doi:10.1038/nrg3833.

41. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21. doi:10.1093/bioinformatics/bts635.

42. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30: 923–930. doi:10.1093/bioinformatics/btt656.

43. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15. doi:10.1186/s13059-014-0550-8.

44. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, et al. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8: 1765–1786. doi:10.1038/nprot.2013.099.

45. Law CW, Chen Y, Shi W, Smyth GK (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15: R29. doi:10.1186/gb-2014-15-2-r29.

46. Ward JH (1963) Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 58: 236–244. doi:10.1080/01621459.1963.10500845.

47. Biase FH, Cao X, Zhong S (2014) Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Res 24: 1787–1796. doi:10.1101/gr.177725.114.

48. Fan X, Zhang X, Wu X, Guo H, Hu Y, et al. (2015) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol 16: 148. doi:10.1186/s13059-015-0706-1.