Modelling dropouts improves 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

Abstract

A key challenge of single-cell RNASeq (scRNASeq) is the many genes with zero reads in some cells, but high expression in others. Modelling zeros using the Michaelis-Menten equation provides a 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. 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

Background

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 the user to define the sets of genes related to each specific process. By contrast, the highly variable genes (HVG) method [2] is unsupervised as it does not require *a priori* knowledge of gene sets, and it automatically identifies the set of genes that have a higher degree of variability than expected based on the observed technical noise. The biological interpretation and relevance of HVGs remains incompletely understood, and a more straightforward concept is that of differentially expressed (DE) genes. SCDE [3] and NODES [4] compare two sets of cells to identify DE genes, similarly to bulk RNASeq methods such as DESeq2 [5] and EdgeR [6]. Applying these methods to scRNASeq data is often challenging due to the difficulty of identifying homogenous subpopulations.

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 demonstrate that the dropout rate can be related to the expression level via the Michaelis-Menten equation. We utilize the model to develop an unsupervised feature selection algorithm, Michaelis-Menten model of dropouts (M3Drop), and we demonstrate that the identified genes correspond to DE

genes. Unlike other DE methods, M3Drop does not require prior identification of groups of cells, making it significantly easier to apply. Using publicly available data, we demonstrate that clustering using the genes identified by M3Drop can reveal new, biologically meaningful groups. Finally, we demonstrate that the features identified by M3Drop make it possible to overcome batch effects to cluster data from different labs and protocols by biological group rather than experiment.

Results

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. The most salient characteristic of scRNASeq data is the presence of a large number of zero values (i.e. dropout events). A typical scRNA-seq experiment has ~50% dropouts (**Table 1**) [3], and it has been suggested that the large number of dropouts is due to transcripts being lost during the library preparation [7]. Based on studies of RT-qPCR assays [8] [9], we hypothesize that the main reason for dropouts is due to failure of the reverse transcription (RT). Since RT is an enzyme reaction we adapt the standard model of enzyme kinetics, the Michaelis-Menten (MM) equation [10], to propose a phenomenological model between the frequency of dropouts and the expression level of genes:

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

where S is the average expression of the gene across all cells and K_M is the Michaelis constant which corresponds to the mean expression level required for a gene to be detected in half of the cells (**Supplementary Note 3**), and $P_{dropout}$ represents the probability that the quantity of cDNA reaches some experiment-specific threshold of detection in any cell.

Fitting Published Datasets

To evaluate the MM model, we fit it to eight recent scRNASeq datasets (**Table 1**) with different biological origins that were analyzed using different experimental and computational procedures. We find that the MM model performs better than two previous models (**Figure 1 A,B**). Pierson and Yau [11] proposed modeling dropout probability as a function of the squared average expression which we find is a poor fit for all the datasets. Kharchenko et al. [7] proposed a logistic model which provides a fit similar to the MM model. The logistic model is equivalent to:

$$P_{dropout} = 1 - \frac{S^n}{K_d + S^n} \quad (2)$$

(i.e. the Hill equation), and by comparing to Eq (1) it is clear that the MM model can be obtained by setting n = 1. Due to the high noise levels the logistic regression fits result in a less steep curve (**Figure S3H**), and thus the sum of absolute residuals is lowest for the MM model (**Figure 1**) while the sum of squared residuals is lowest for the logistic regression (**Figure S3I**). The

similarity of the MM model and the logistic model is most evident for the Zeisel and Klein datasets where the noise is reduced through the use of unique molecular identifiers (**Figure S3**).

Differential Expression

By considering deviations from the MM model, biological insights about a scRNASeq dataset can be obtained. One assumption of the MM model is that each gene has the same expression in all cells. Consequently, genes located to the right of the fitted curve indicate differences in expression across two or more cell subpopulations. The reason why the MM model can be used to identify differentially expressed genes is because the dropout rate is a convex function of the expression level whereas calculating mean expression and dropout rate are linear functions (**Figure 1C, D**). Thus, the fitted curve can serve as a null model, and we can test for genes that are further to the right of the curve than expected by chance (**Methods**). We refer to this method as M3Drop and it allows us to identify genes that are DE between subsets of cells.

We compared the performance of M3Drop to three existing DE methods: DESeq [12], DESeq2 [5], and SCDE [7], in addition to a method for identifying highly variable genes (HVG, [2]). HVG is not designed to identify DE genes, but it is similar to M3Drop in that it does not require pre-defined groups to select genes. M3Drop and HVG are much faster than the three DE methods, each took only a minute to compute for datasets of thousands of cells and over ten thousand genes using a single processor (**Figure 2A**) and can be computed for the 44,808 cell mouse retina dataset [13] in less than two minutes. In contrast, DESeq2 and SCDE took several hours for datasets with >200 cells. For 5 datasets, DESeq, DESeq2 and SCDE report ~40% of genes as DE, many of which are known housekeeping genes (e.g. Psmb2 and Snrpd3) or external spike-ins, whereas M3Drop and HVG tend to be more conservative and report ~10% of the genes (**Figure 2B**).

To assess the quality of the predictions we defined a set of true positive genes for each dataset based on Gene Ontology categories relevant for each context. Similarly, we defined a set of true negatives based on spike-ins and housekeeping genes (**Methods**). Our results show that M3Drop and HVG have higher enrichments of true positive genes and lower false positive rates for six of the datasets (**Figure 2C,D**). The only two datasets where DESeq, DESeq2 and SCDE performed better (Buettner and Shalek), examined fewer than 300 cells and relatively small biological differences, cell cycle and stimulation with lipopolysaccharide respectively. We conclude that M3Drop is a fast and accurate method for identifying DE genes.

Reproducibility using pseudoreplicates

To investigate the consistency of the different methods we took advantage of the fact that some of the datasets derive from similar tissues; two were mouse neuronal tissue (Usoskin and Zeisel), and two come from mouse preimplantation embryos (Deng and Biase). We treat these studies as pseudoreplicates, and we asked if similar genes were identified by each method.

Genes identified by M3Drop were the most consistent across datasets which examined the same biological system (**Figure 3A, B**). The gene sets reported by M3Drop, DESeq2, DESeq, and SCDE overlap more than expected by chance within datasets ($p < 10^{-50}$). Whereas, genes identified by HVG were only significantly consistent with the other methods for the Deng and Usoskin datasets ($p < 10^{-20}$) and there was even a significant depletion of common genes between HVG and DESeq2 for the Zeisel and Biase dataset ($p < 10^{-30}$). M3Drop was significantly more reproducible across datasets compared to the other methods with [2.37, 3.29]-fold enrichment (95% CI) of genes found in both Deng and Biase datasets compared to chance, and [1.99, 2.91]-fold enrichment between Usoskin and Zeisel datasets. DESeq, DESeq2, and SCDE were only 1.069 to 1.634 fold enriched between datasets. HVG showed high reproducibility across the developmental datasets, [1.63, 2.96]-fold enrichment, but very low reproducibility across the neuronal datasets, [1.01, 1.16]-fold enrichment.

Testing on simulated data

Testing computational methods on real data is difficult since we have very limited knowledge of the ground truth. Unlike bulk RNASeq there are no large scale benchmarked datasets available [14,15]. Thus, we generated simulated data using a zero-inflated negative binomial model fitted to all genes present in the Buettner dataset (**Table 2, Methods**). This dataset was chosen since it appears to be relatively homogenous and it has the fewest DE genes according to all five methods (**Figure 2B**). Simulated data consisted of two subpopulations with a subset of genes differentially expressed between them (**see Methods**). We simulated ten replicates for each of 46 sets of model parameters, varying the size of the cell populations, degree of differential expression, number of differentially expressed genes, and degree of dispersion.

Simulations show that M3Drop is more accurate than HVG at detecting differentially expressed genes (**Figure 4**). Consistent with our results for the real data, we find that M3Drop is conservative as it finds no false positives using a 5% FDR. M3Drop fails to detect DE genes with very high or very low expression, but our method has good performance for genes detected with mean count of 30-2000 (**Figure 4A**). In contrast, HVG identifies many false positives among low to moderately expressed genes (0.5-200 counts), which accounts for it's overall poor performance since most genes in scRNASeq experiments fall into this regime (**Figure 4A, S3**). We found that HVG only detected DE genes with relatively high overall expression (>100 counts)

We find that both methods have low sensitivity, with poor performance when the fold change in mean expression is less than 5 (**Figure 4B**). Surprisingly, HVG exhibits a decrease in performance at very high fold changes (> 50). This may be due to the weakening relationship between dispersion and mean expression which counters the increase in variability due to difference in mean expression. By contrast, the performance of M3Drop improves with increasing fold change.

Both methods' performance improves as the number of cells increases (**Figure S5**). However, when considering the proportion of cells with high/low expression of the DE genes, M3Drop is symmetric with equal performance, both when only a few cells overexpress the genes and when only a few cells have reduced expression. By contrast, HVG is biased towards genes overexpressed in a small number of cells (**Figure 4C**). We also found that high transcriptional variability may result in an increase in dropouts, even in the absence of differential expression. Using simulated data where 10% of genes have 3-fold inflated variability across all cells, we find that M3Drop also detects highly variable genes but is less sensitive than the HVG method (**Figure S6**).

These results show that a method for detecting highly variable genes fails to identify differential expression in a number of situations, such as low expression in a small number of cells or very high fold changes, and is prone to identifying false-positives. Although differentially expressed genes often exhibit high variability, these two qualities are not synonymous. Genes may exhibit high transcriptional variability without being differentially expressed and genes may be differentially expressed without exhibiting high overall variability (**Figure 4D**). Indeed, we find that observed variability is only weakly correlated with fold-change of differentially expressed genes (Pearson correlation = 0.33, mean expression = 100, N = 100) since most of the observed variability is due to within-population variability rather than between population differences in mean expression. In addition, there is high intrinsic noise in estimating sample variability [16]. In contrast, dropout rate is highly correlated with fold-change (Pearson correlation = 0.78) and exhibits much lower levels of intrinsic noise (**Figure 4E, S7, S8**). Hence, by using dropouts as opposed to variance, M3Drop outperformed HVG in detecting differentially expressed genes.

De novo identification of trophectoderm and inner cell mass

The main advantage of using M3Drop to identify DE genes, as opposed to traditional DE methods, is that no assumptions about the underlying groups is required. Indeed, M3Drop is best described as an unsupervised feature selection method, and the previous section shows that the features correspond to DE genes. One application of feature selection is noise reduction when identifying subpopulations of cells, and others have used high variability to select genes for this purpose [1,17–20].

We show that the DE genes identified by M3Drop are informative for identifying cell subpopulations by applying M3Drop to a set of 255 cells across early mouse development [21]. There were 1,478 significantly DE genes identified at 1% FDR (**Figure 1D**). Clustering the cells using Ward's hierarchical clustering [22] recapitulates the known sampling times of the cells (**Figure 5A**). However, the three timepoints of blastocysts (earlyblast, midblast, and lateblast) cluster together into two distinct groups. Based on what is known about embryonic development, we hypothesized that the two groups correspond to the inner cell mass (ICM), which develops into the fetus, and trophectoderm (TE), which develops into the placenta. To test this hypothesis, we compared the expression levels of six known marker genes (**Figure 5**

B-G). One group had significantly higher expression of Sox2, Oct4, and Nanog which are known to be important for the ICM; whereas the other group had significantly higher expression of the TE marker Cdx2 and its targets Eomes and Elf5 [23,24,25] . This dichotomy was not observed when blastocyst cells were grouped by time-point (**Figure S11**).

We identified the top 100 marker genes which distinguish the two blastocyst groups based on the extent of overlap between the distribution of expression values across cells in each group [26]. Gene Ontology analysis of these marker genes showed that the putative TE cells expresses genes involved in the actin cytoskeleton, cellular adhesion, and vasculature development - processes which are important for implantation of the embryo and development of the placenta (**Figure 5H, Table S4**). The putative ICM cells expresses genes involved in DNA methylation/demethylation, stem cell maintenance, and gastrulation (**Figure 5I, Table S4**), consistent with evidence that DNA methylation is reset during the blastocyst stage [27].

Combining developmental datasets

As an additional application of M3Drop feature selection, we consider the problem of merging datasets from different labs. Merging datasets generated by different labs is complicated not only by the use of different sample preparation methods and protocols, but also by the presence of batch effects that can affect samples that were processed in the same way [28,14]. One consequence of this technical variability is that samples will often show statistically significant differences, even though the biological origins are identical.

We obtained three additional mouse preimplantation datasets [29-31] and we pooled those cells with the Deng and the Biase datasets to obtain a set of 521 cells (**Methods**). Using only the 316 genes identified as DE by M3Drop in both the Deng and the Biase dataset (**Table S5**), we found that cells cluster by embryonic stage rather than dataset of origin (**Figure 6**). By contrast, using the genes identified by HVG failed to cleanly separate the stages and using the full set of genes results in a clustering where the cells are grouped by batch instead (**Figure 6,S13,S14**). Consistent with these results, we find that using SC3 the correct number of clusters can be estimated from the M3Drop DE genes, 7 clusters, whereas using all genes gives an estimate of 14 clusters, and HVG genes gives an estimate of 4 clusters. We conclude that batch effects are not a significant barrier provided that one can focus on the most biologically informative genes.

Importantly, the 316 genes identified by M3Drop include many that are known to be important for early embryonic development, such as: Zscan4d, which is known to have highly specific expression in 2-cell embryos [32], H1foo, an oocyte-specific histone, Tdgf1 which is required for gastrulation [33], the oocyte-specific growth factor Bmp15 [34] and Bhmt which are involved in mouse embryonic methylation [35]. Four of the six Obox gene family members which have been previously associated with oogenesis [36] were amongst this set of genes, Obox1, Obox3, Obox5, and Obox6. However, only Obox1 and Obox5 were preferentially expressed among zygotes and oocytes (**Figure S16**). By contrast, Obox6 was most highly expressed among 4-cell to 8-cell embryos and Obox3 was more than 20-fold more highly expressed among 2-cell

embryos than any other stage. Taken together, these results highlight the utility of M3Drop for identifying reproducible features that make it possible to integrate different scRNASeq experiments.

Discussion

M3Drop is a powerful new unsupervised feature selection method for scRNASeq data. We have demonstrated that the genes identified by M3Drop correspond to DE genes. Unlike traditional DE methods, M3Drop can identify differentially expressed genes without *a priori* identification of biological conditions. For many single-cell experiments the biologically relevant groups are unknown, making it hard to apply traditional DE methods. We demonstrated that the genes identified by M3Drop can reveal biologically meaningful clusters and that they allow us to overcome batch effects when merging datasets.

Our results reveal the utility of modelling scRNASeq as an enzyme reaction where a fixed but unknown amount of product (cDNA) is required for detection. The MM model identifies consistent differences between different protocols and our results show that when the reverse transcription and pre-amplification occur in small volumes prior to pooling (e.g. Fluidigm C1), then detection rates are higher compared to methods where one or both of these steps occur after pooling multiple cells together, e.g. CEL-Seq, Drop-Seq (**Figure S2**) [37,38]. This result highlights an intrinsic advantage of using small reaction volumes, and it should be an important consideration when evaluating different protocols.

The most commonly used feature selection criterion for scRNASeq data is high gene expression variability [1,17–20]. Our simulation studies reveal that M3Drop is less prone than HVG [2] to identifying false-positives among lowly expressed genes, and that the number of false positives can be more than an order of magnitude lower (**Figure 4A**). On average 60% of genes detected by HVG were false positives whereas only 0.1% of genes detected by M3Drop were false positives. Since false positives occur due to random chance it is unlikely that the same genes will be identified across different datasets which explains the low concordance observed between the pseudoreplicates. When comparing to the HVG method on published datasets, we found that M3Drop is more reproducible (**Figure 3A,B**).

For both high variance genes and differentially expressed genes we observe that high sampling noise affects observed variances (**Figure S7, S9**). High sampling noise is an intrinsic property of sample variance [16]. Its effect evident in the weak of correlation between variance and fold change for samples of fewer than 1,000 cells (**Figure S8, S10**). Dropout rates, are much more robust to sampling noise and they consistently show high correlations with fold change even for small sample sizes (**Figure S8, S10**).

The fundamental reason why the HVG method performs poorly is due to the fact that (normalized) variance is a poor metric for discriminating features in scRNASeq data (**Figure S7,S8**). Current protocols and sequencing depths result in the majority of expression

measurements being dropouts. As a result, the first principal component of variability across genes is more strongly correlated with the dropout rate than with the variance. In addition, for most genes, the within population variance is similar to the between population variance even for large fold-changes in mean expression. Thus, dropout rate is a better indicator of different subsets of cells since it has a cleaner relationship with differential expression.

Conclusion

We introduce a novel unsupervised feature selection method, M3Drop, based on a Michaelis-Menten model of the dropout rate in scRNASeq datasets. We show M3Drop can reproducibly identify DE genes in both real and simulated single-cell RNASeq datasets without *a priori* determination of experimental groups. Moreover, the features identified by M3Drop allows us to overcome batch effects when combining data from different studies.

Methods

Single-cell RNASeq datasets

We considered eight 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 converted to counts per million. 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 [39].

Fitting Dropout Models

The expression of each gene was averaged across all cells including those with zero reads for a particular gene (S) (See **Supplementary Note 2**). 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:

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

was fitted using maximum likelihood estimation as implemented by the `mle2` function in the `bbmle` R package. The logistic regression [7]:

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

was fitted using the glm R function. Comparing to Equation (2), we note that *a* = log(K_d) and *b* = -n.The double exponential model [11] was fit by log transforming the equation then using the lm R function to fit the coefficient to the resulting quadratic model:

$$P_{dropout} = e^{\lambda * S^2}$$

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

We also considered a double MM model:

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

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**).

Differential Expression (DE)

Rearranging the general Michaelis-Menten equation to solve for the Michaelis-Menten constant K gives:

$$K = \frac{P * S}{1 - P} \tag{1}$$

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:

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

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:

$$Z = \frac{log(K_j) - log(K_M)}{\sqrt{\sigma^2_{log(K_j)} + \sigma^2_{log(K_M)}}} \tag{3}$$

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

$$\sigma_{\log(K_M)} = \frac{sd(\log(K_j) - \log(K_M))}{\sqrt{N}} \tag{4}$$

$$\sigma_{\log(K_j)} = \log(K_j) - \log(K_j - \sigma_{Kj}) \quad (5)$$

Differential expression was also calculated using DESeq (version 1.22.1) and DESeq2 (version 1.10.1) and SCDE (version 1.2.1) for every pair of groups for each dataset (**Supplementary Note 3**). DESeq was run on a single processor and all cells in each dataset. DESeq2 was run on 10 processors and datasets containing more than 1,000 cells were downsampled to 1,000 cells to reduce runtimes. SCDE was run on 10 processors and datasets containing more than 500 cells were downsampled to 500 cells, while ensuring equal coverage for all groups. SCDE was run with `max.pairs` set to 500, `min.pairs.per.cell` set to 1, using threshold segmentation, and the original fitting method. All other methods were run with default parameters. DE genes were corrected for multiple testing using the Benjamini-Hochberg procedure with a 1% FDR.

To reduce the number of pairwise comparisons between groups tested using classical DE methods we merged cell subpopulations into coarse-level labels (**Table S2**). The eleven different cell-lines of the Pollen dataset were merged into 4 groups based on the tissue of origin. Similarly, the nine different neuronal subtypes of the Zeisel dataset: Interneurons (N), S1 Pyramidal (N), CA1 Pyramidal (N), Oligodendrocyte (G), Astrocyte (G), Ependymal (G), Microglia, Endothelial (Non) and Mural (Non), were merged into four broad categories: Neuron (N), Glia (G), Microglia, and Non-Neuronal (Non). Sequential developmental stages in the Deng dataset were merged to increase sample sizes, e.g. the dataset contained only 14 cells in the 4 cell stage after quality control.

Highly Variable Genes (HVG) were calculated using the published method [2]. However, rather than fitting the model to the spike-ins, of which there were fewer than 10 in many datasets, the model was fit using all genes. Significantly variable genes were those passing a 1% FDR and minimum 50% biological dispersion.

Quality of Differentially Expressed Genes

Accuracy of DE was evaluated using the known marker genes for each of the 6 cell-type datasets: Deng, Usoskin, Klein, Zeisel, Pollen and Biase (**Table S3**). For the Shalek dataset we compared the two replicates of unstimulated cells to the two replicates of LPS stimulated cells and used all 990 genes with the Gene Ontology annotation “immune response” as true positives. For the cell-cycle dataset (Buettner, [1]) we used the list of 892 known cell-cycle genes from the Gene Ontology and Cyclebase provided with the original publication. For true negatives any spiked-in RNAs for each dataset was combined with 11 lowly variable housekeeping genes identified by Eisenberg and Levanon [40].

Reproducibility of DE genes was measured as the number of genes identified by two different methods applied to the same dataset or by the same method applied to two different datasets. The intersection of the two lists of DE genes was compared to the number that would be

expected by chance given the number of all detected genes (N) that were identified as DE in each case (n_k), where the index *k* represents either a dataset or a method.

$$E(overlap_{ij}) = \frac{n_i * n_j}{N}$$

Significance was evaluated using Fisher's exact test.

Synthetic Data

Synthetic data was generated using a zero-inflated negative binomial distribution. The relationship between mean and dispersion parameters (CV2 as defined in [2]) was estimated using a power law from the Buettner dataset, $CV2 = 9285 * \mu^{-1.85}$. We chose to use this dataset since it exhibits only minor biological variability due to the cell cycle ([1,41], **Figure S3**). Based on a binomial distribution where the parameter is determined by the MM model fit to the experimental data (K = 10.3, **Table S1**). The resulting synthetic expression matrices approximated observed single-cell data (**Figure S4**). Genes where every entry was a zero or where no entries were zero were excluded since neither can be detected by M3Drop.

When comparing overall accuracy of the different methods (**Figure 4B & C,S6**), gene-specific mean expression was drawn from a truncated log normal distribution (mean=sd=1) which approximated observed values (**Figure S4D**). Genes were ranked by reported p-values from M3Drop and HVG [2]. Ranks were used to calculate the area under the ROC curve (AUC) for each method.

When considering accuracy across expression levels (**Figure 4A,S7**), log mean expression values for 25,000 genes were drawn from a uniform distribution over the interval [-2, 4]. A 5% FDR multiple testing correction was applied to the p-values reported by M3Drop and HVG. Accuracy was evaluated for bins spanning one order of magnitude. We calculated the observed false discovery rate (FDR), the proportion of genes deemed significant by the respective method (positives) that were not simulated with differential expression/high variance (true positives), and false negative rate (FNR), the proportion of true positives not found to be significant by the respective method (negatives).

To further examine the effect of differential expression on observed variance and dropout rate, we simulated a single gene differentially expressed in two equally sized subpopulations with fixed overall mean expression while varying fold-change from 1-100 across the two subpopulations. Observed sample variance and dropout rate and Pearson correlations were calculated using standard methods.

Preimplantation Mouse Development

M3Drop was applied to the 255 cells of the Deng dataset. Cells were clustered using Ward's hierarchical clustering [22], the resulting tree was cut to give five groups, which corresponds to

the first division of the blastocyst group. Normalized expression of three markers of the Inner cell mass (ICM) and trophectoderm (TE) was compared for the two groups of blastocyst cells using a Wilcoxon rank-sum test. Novel markers were ranked according the the area under the ROC curve when each gene was used to predict to which of the blastocyst groups each cell belonged [26]. Gene Ontology annotations were obtained from Ensembl80, terms annotated to more than 50% of detected genes were excluded from consideration. Enrichments among the top 100 marker genes for the ICM and TE clusters were determined using a hypergeometric test and Bonferroni multiple testing correction.

Combining Datasets

We combined the two mouse development datasets, Deng and Biase, with three additional publicly available datasets [29-31] . Each dataset was filtered for quality using identical criteria as described above : all genes annotated as processed pseudogenes in Ensembl (version 80) were excluded, cells with fewer than 2000 detected genes were removed, and genes detected in fewer than 4 cells or with average normalized expression < 10^{-5} were excluded from consideration. The Deng and Goolam data was normalized using CPM, all other datasets were already FPKM/RPKM normalized. No other correction for batch effects was applied.

We compared feature selection using M3Drop or HVG on both the Deng and Biase datasets to no feature selection. Only genes detected in 4 or more cells and with expression > 10^{-5} in all four datasets were considered. This left 316 M3Drop genes, 67 HVG genes, and 11,440 genes total across the datasets. Cluster number estimation on the combined dataset was performed using SC3 version 1.1.8 [26].

Individual cells from all datasets were clustered together using complete linkage hierarchical clustering. The resulting hierarchy was cut at every level and the resulting clusters were compared to the consensus stages: zygote (incl. oocytes), 2-cell, 4-cell, 16-cell (incl. morula), and either general blastocyst or TE and ICM stages. ICM and TE labels for the Deng data were derived from the clustering described above. The similarity between the stage labels and each clustering was quantified using the adjusted Rand index (ARI) [42].

ARI is calculated as the proportion of all possible pairs of cells which are consistently in the same group (n_{11}) or in different groups (n_{00}) in both clusterings. This is then normalized to account for the proportions expected by chance given the number and size of the detected clusters. The maximum adjusted Rand index across all seven possible clusterings was reported:

$$Index = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

$$ARI = \frac{Index - Expected}{MaxIndex - Expected}$$

Code/Data Availability

Dataset accession codes are listed in **Table 1**.

Bioconductor packages used: DESeq (version 1.22.1), DESeq2 (version 1.10.1), SCDE (version 1.2.1).

M3Drop is freely available on github : 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.

References

1. Buettner F, Florian B, Natarajan KN, Paolo Casale F, Valentina P, Antonio S, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33: 155–160.

2. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10: 1093–1095.

3. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 2016;17: 63.

4. Sengupta D, Debarka S, Rayan NA, Michelle L, Bing L, Shyam P. Fast, scalable and accurate differential expression analysis for single cells [Internet]. 2016. doi:10.1101/049734

5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

6. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8: 1765–1786.

7. Kharchenko PV, Lev S, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11: 740–742.

8. Bengtsson M, Martin B, Martin H, Patrik R, Anders S. Quantification of mRNA in single cells

and modelling of RT-qPCR induced noise. BMC Mol Biol. 2008;9: 63.

9. Reiter M, Kirchner B, Müller H, Holzhauer C, Mann W, Pfaffl MW. Quantification noise in single cell experiments. Nucleic Acids Res. 2011;39: e124.

10. Michaelis L, Menten ML. Die Kinetik der Invertinwirkung. Biochem Z. 1913;49: 333–369.

11. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16: 241.

12. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11. doi:10.1186/gb-2010-11-10-r106

13. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161: 1202–1214.

14. Tung P-Y, Po-Yuan T, Blischak JD, Chiaowen H, Knowles DA, Burnett JE, et al. Batch effects and the effective design of single-cell gene expression studies [Internet]. 2016. doi:10.1101/062919

15. Rapaport F, Franck R, Raya K, Yupu L, Mono P, Azra K, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14: R95.

16. Cho E, Cho MJ, Eltinge J. THE VARIANCE OF SAMPLE VARIANCE FROM A FINITE POPULATION. International Journal of Pure and Applied Mathematics. 2005;21: 387–394.

17. Björklund ÅK, Forkel M, Picelli S, Konya V, Theorell J, Friberg D, et al. The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing. Nat Immunol. 2016;17: 451–460.

18. Wilson NK, Kent DG, Buettner F, Shehata M, Macaulay IC, Calero-Nieto FJ, et al. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell. 2015;16: 712–724.

19. Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. 2016;17: 67.

20. Tsang JCH, Yu Y, Burke S, Buettner F, Wang C, Kolodziejczyk AA, et al. Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. Genome Biol. 2015;16: 178.

21. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343: 193–196.

22. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. 1963;58: 236–244.

23. Chen Y, Wang K, Gong YG, Khoo SK, Leach R. Roles of CDX2 and EOMES in human

induced trophoblast progenitor cells. Biochem Biophys Res Commun. 2013;431: 197–202.

24. Hamatani T, Toshio H, Carter MG, Sharov AA, Ko MSH. Dynamics of Global Gene Expression Changes during Mouse Preimplantation Development. Dev Cell. 2004;6: 117–131.

25. Marikawa Y, Alarcón VB. Establishment of trophectoderm and inner cell mass lineages in the mouse embryo. Mol Reprod Dev. 2009;76: 1019–1032.

26. Kiselev VY, Kristina K, Schaub MT, Tallulah A, Tamir C, Natarajan KN, et al. SC3 - consensus clustering of single-cell RNA-Seq data [Internet]. 2016. doi:10.1101/036558

27. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev. 2014;28: 812–828.

28. Hicks SC, Mingxiang T, Irizarry RA. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data [Internet]. 2015. doi:10.1101/025528

29. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015;16: 148.

30. Goolam M, Mubeen G, Antonio S, Graham SJL, Macaulay IC, Agnieszka J, et al. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. Cell. 2016;165: 61–74.

31. Xue Z, Zhigang X, Kevin H, Chaochao C, Lingbo C, Chun-yan J, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013;500: 593–597.

32. Falco G, Lee S-L, Stanghellini I, Bassey UC, Hamatani T, Ko MSH. Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. Dev Biol. 2007;307: 539–550.

33. Jin J-Z, Ding J. Cripto is required for mesoderm and endoderm cell allocation during mouse gastrulation. Dev Biol. 2013;381: 170–178.

34. Mottershead DG, Sugimura S, Al-Musawi SL, Li J-J, Richani D, White MA, et al. Cumulin, an Oocyte-secreted Heterodimer of the Transforming Growth Factor-β Family, Is a Potent Activator of Granulosa Cells and Improves Oocyte Quality. J Biol Chem. 2015;290: 24007–24020.

35. Lee MB, Kooistra M, Zhang B, Slow S, Fortier AL, Garrow TA, et al. Betaine Homocysteine Methyltransferase Is Active in the Mouse Blastocyst and Promotes Inner Cell Mass Development. J Biol Chem. 2012;287: 33094–33103.

36. Rajkovic A, Yan C, Yan W, Klysik M, Matzuk MM. Obox, a family of homeobox genes preferentially expressed in germ cells. Genomics. 2002;79: 711–717.

37. Ziegenhain C, Christoph Z, Beate V, Swati P, Björn R, Martha S, et al. Comparative

analysis of single-cell RNA sequencing methods [Internet]. 2016. doi:10.1101/035758

38. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014;11: 41–46.

39. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014;510: 363–369.

40. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29: 569–574.

41. McDavid A, Andrew M, Greg F, Raphael G. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. Nat Biotechnol. 2016;34: 591–593.

42. Rand WM. Objective Criteria for the Evaluation of Clustering Methods. J Am Stat Assoc. 1971;66: 846.

43. Usoskin D, Dmitry U, Alessandro F, Saiful I, Hind A, Peter L, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2014;18: 145–153.

44. Klein AM, Linas M, Ilke A, Naren T, Adrian V, Victor L, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015;161: 1187–1201.

45. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347: 1138–1142.

46. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32: 1053–1058.

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

Tables

Table 1: Eight publicly available datasets.

Dataset	Cell-types	Original Labels	Method	N	% Zero	Source
Buettner	Mouse ESC	Cell-cycle stage	Smartseq Counts -> (CPM)	279	51%	[1] E-MTAB-2805
Deng	Mouse embryos	Developmental timepoint	Smartseq Counts -> (CPM)	255	50%	[21] GSE45719
Usoskin	Mouse neurons	PCA-based clustering	5’ Seq Counts -> (CPM)	530	73%	[43] GSE59739
Klein	Mouse ESC	Differentiation	CEL-Seq	2448	63%	[44]

		timepoint	UMIs -> (CPM)			GSE65525
Zeisel	Mouse brain	BackSPIN clustering	5' Seq UMIs -> (CPM)	2542	77%	[45] GSE60361
Shalek	Mouse bone marrow	Stimulated & unstimulated	Smartseq FPKMs	173	66%	[39] GSE48968
Pollen	Human cell lines & tissues	Cell line identity	Smartseq FPKMs	301	60%	[46] SRP041736**
Biase	Mouse embryos	Developmental stage	Smartseq FPKMs	56	38%	[47] GSE57249

**UMI = Unique Molecular Identifier; FPKM = fragments per kilobase per million; CPM = count per million*

*** Processed data was provided by the authors*

Table 2: Simulation parameters

Parameter	Description	Default value
n	Number of genes	25,000
N	Number of cells	300
Fc	Fold change in expression between subpopulations	10
Pc	Proportion of genes that are DE	10%
Ps	Proportion of cells in the subpopulation (higher expression)	50%
d	Fold increase in dispersion over fitted values	1 (no change)
K	Parameter of Michaelis-Menten dropouts	10.3

Figure Captions

Figure 1: Michaelis-Menten model of dropouts. (A) The Michaelis-Menten (solid black), logistic (dashed purple), and double exponential (dotted blue) models are fit to the Deng dataset [21]. 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) Michaelis-Menten had the smallest sum of absolute residuals (SAr) across all eight datasets considered. (C) Since the Michaelis-Menten equation (black) is a convex function, genes that are expressed at different levels in different cell types (light grey points) become outliers (dark grey point) when averaging across a mixed population. (D) 1,478 significant outliers (purple) from the Michaelis-Menten equation (black line) were identified at 1% FDR in the Deng dataset.

Figure 2: Quality of differential gene detection on real datasets. (A) HVG and M3Drop are the most computationally efficient methods. SCDE was limited to 500 cells and DESeq2 was limited to 1000 cells both were run in parallel on 10 processors. DESeq, HVG and M3Drop were run on the full datasets on a single processor. (B) Proportion of all detected genes that were called as DE (1% FDR). (C) Fold enrichment of true positive (cell-type markers or members of relevant pathways) among DE genes for each dataset. (D) Proportion of housekeeping genes and spike-ins called as DE (conservative estimate of false positive rate) for each method.

Figure 3: Michaelis-Menten identifies reproducibly differentially expressed genes. (A & B) The ratio of Observed/Expected overlaps (given the number of genes called as DE) between DE genes identified using different methods/datasets. Ratios of observed to expected overlap controlling for the number of detected genes and number of DE genes for each method are presented for the Biase (A green lower triangle), Deng (A blue upper triangle), Zeisel (B green lower triangle), and Usoskin (B blue upper triangle). The diagonals (purple) show the value for comparisons for the same DE method across two datasets on early embryonic development in mice (A) and two datasets collected from mouse neuronal tissue (B). Cells in white are not significantly different from chance (obs/exp = 1) based on Fisher's exact test.

Figure 4: **Dropouts are more informative of DE than variability.** (A) Quality of calls for M3Drop and HVG for genes 10-fold differentially expressed in 50% of cells across a range of expression values representative for many datasets. Grey indicates distribution of observed mean expression values in Buettner dataset. Red dashed line indicates FDR used in multiple-testing correction. Note that the solid black line is at 0 for all expression levels. (B-C) Area under the ROC curve when genes are ranked by p-value for HVG (green) and M3Drop (black) as it relates to fold-change of DE genes (B) and proportion of cells overexpressing the genes (C). (D-E) Relationship between fold change and observed variance (D) and observed dropout rate (E) when overall mean expression is fixed at 100, in a population of 50 low and 50 high expression cells.

Figure 5: Identification of biologically relevant clusters from a development time-course. (A) Expression of the 1,478 DE genes identified by M3Drop (1% FDR) in the Deng dataset reveals two clusters of cells across all three blastocyst timepoints using Ward's hierarchical clustering. (B-D) Expression of markers of the inner cell mass (ICM) for the two blastocyst clusters, all show significant upregulation in Group 5 (Wilcox rank-sum test, $p < 0.0001$). Star indicates that Sox2 was identified as DE by M3Drop. (E-G) Expression of markers of the trophectoderm (TE), all show significant upregulation in Group 4 (Wilcox rank-sum test, $p < 0.00001$) and all were identified as DE by M3Drop. (H & I) Gene Ontology enrichments among the top 100 marker genes for group 4 and 5 respectively, all categories are significant with $p < 0.01$ using hypergeometric test and Bonferroni correction.

Figure 6: M3Drop marker genes makes it possible to merge datasets from different experiments. (A) Clustering 521 cells from four different datasets using genes selected by M3Drop (black), HVG (red) or all detected genes (blue). Clusters were compared to the true developmental stages: zygote, 2-cell, 4-cell, 8-cell, 16-cell, and blastocyst. Random chance: ARI = 0, identical clustering: ARI = 1. (B-D) Principal component analysis based on the genes detected by M3Drop (B) or HVG (C) in both Deng and Biase datasets or all detected genes (D).

Figures

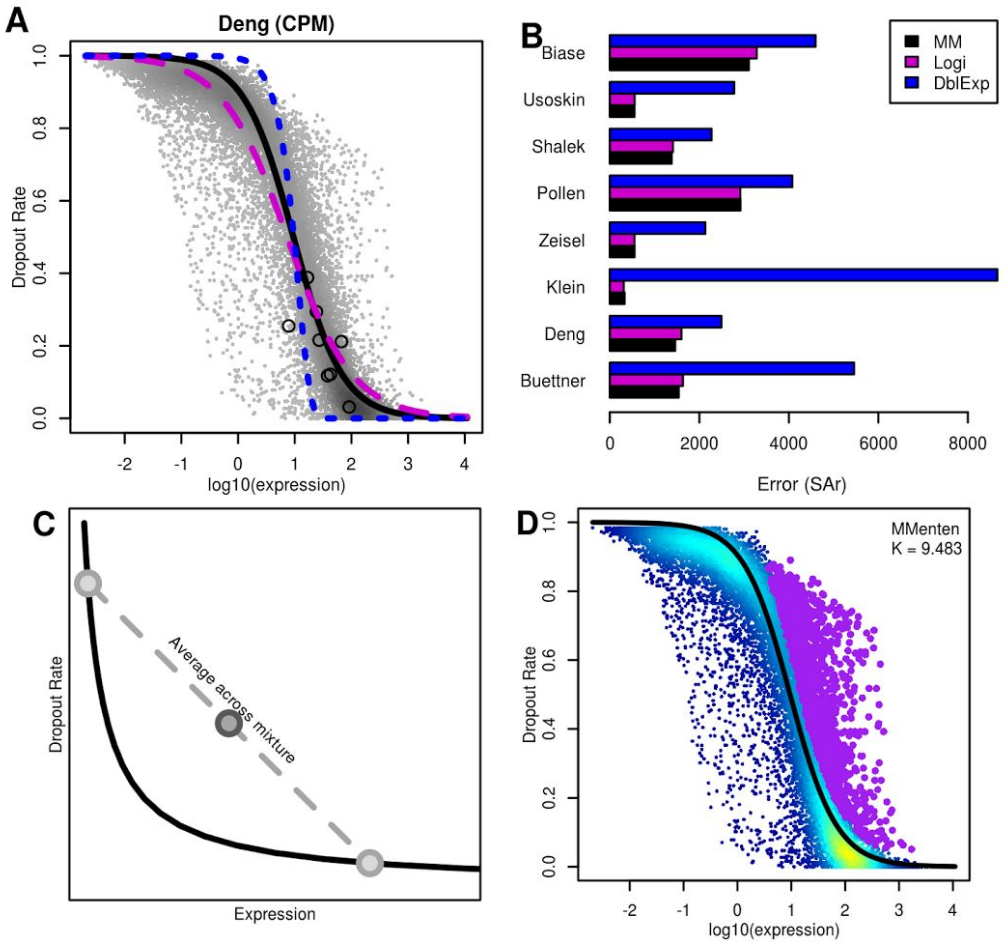

Figure 1

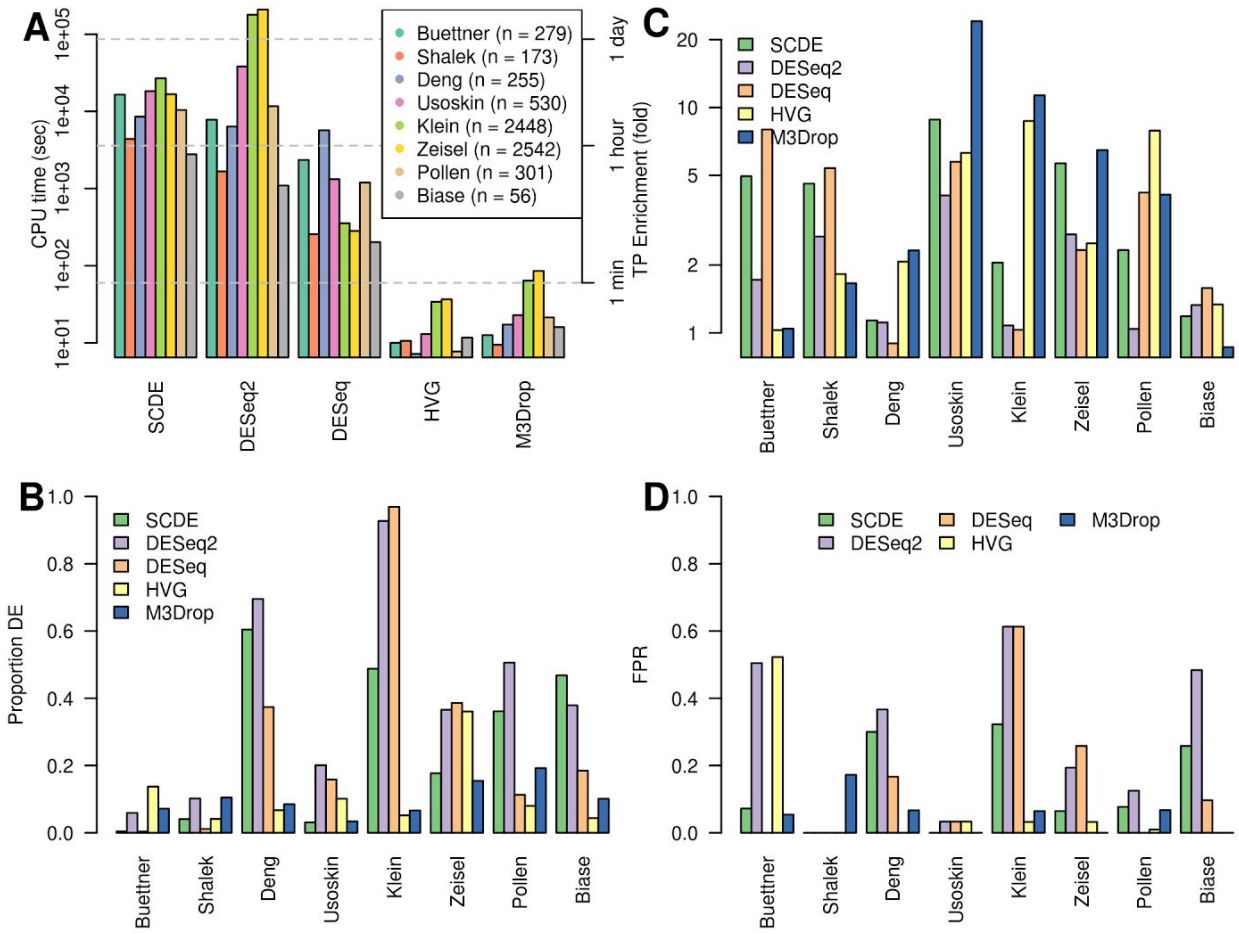

Figure 2

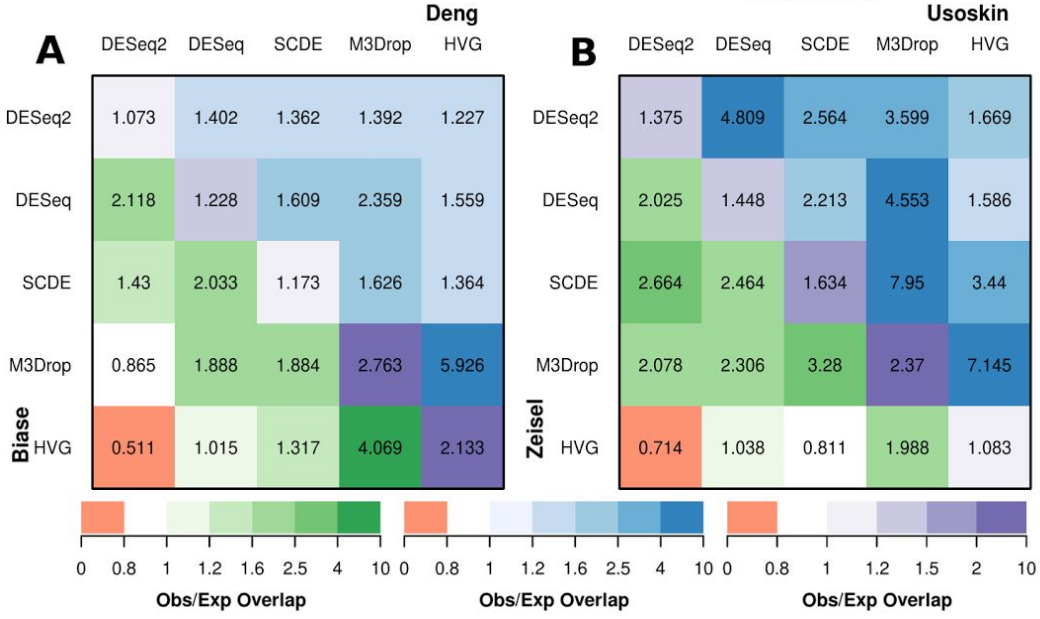

Figure 3

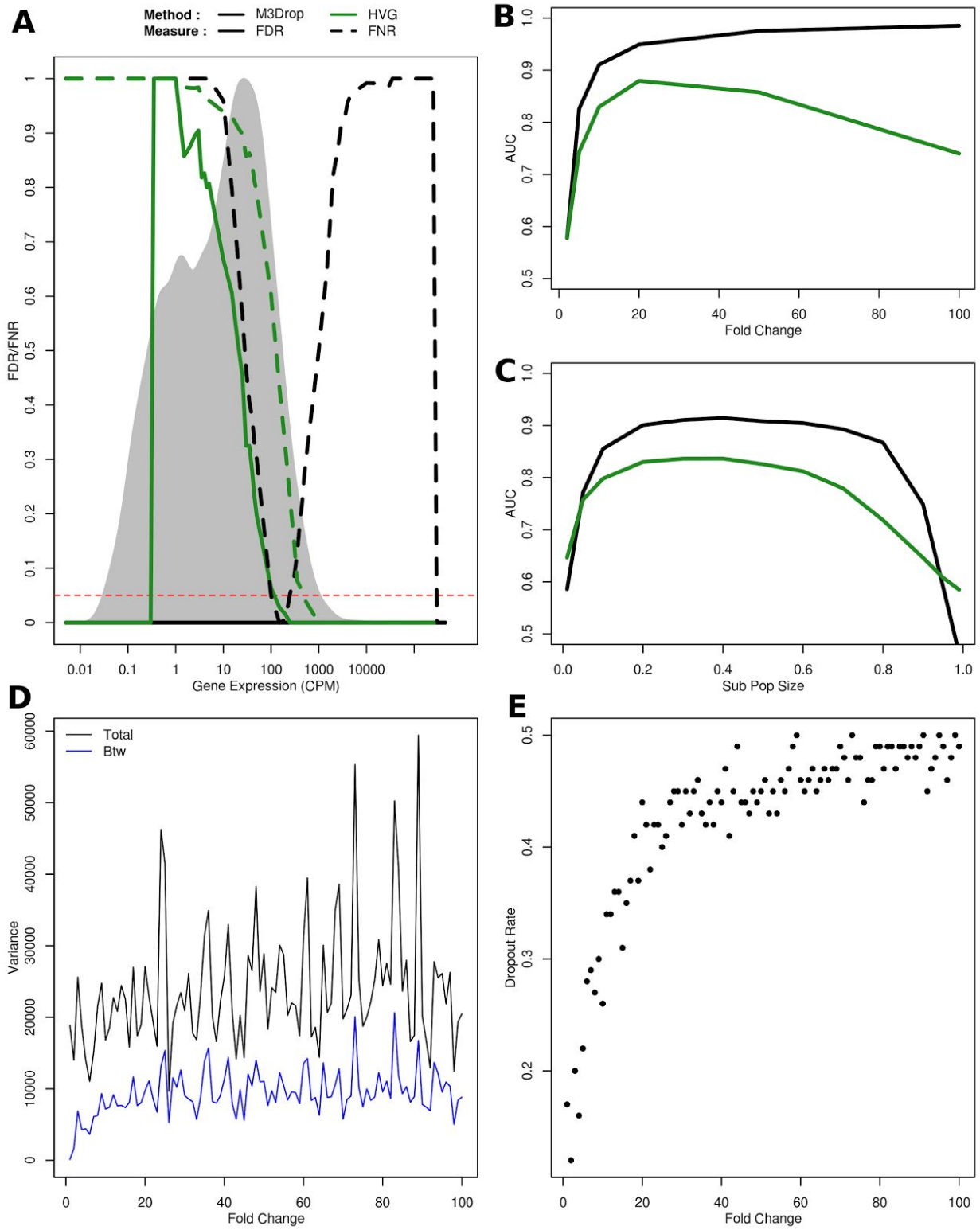

Figure 4

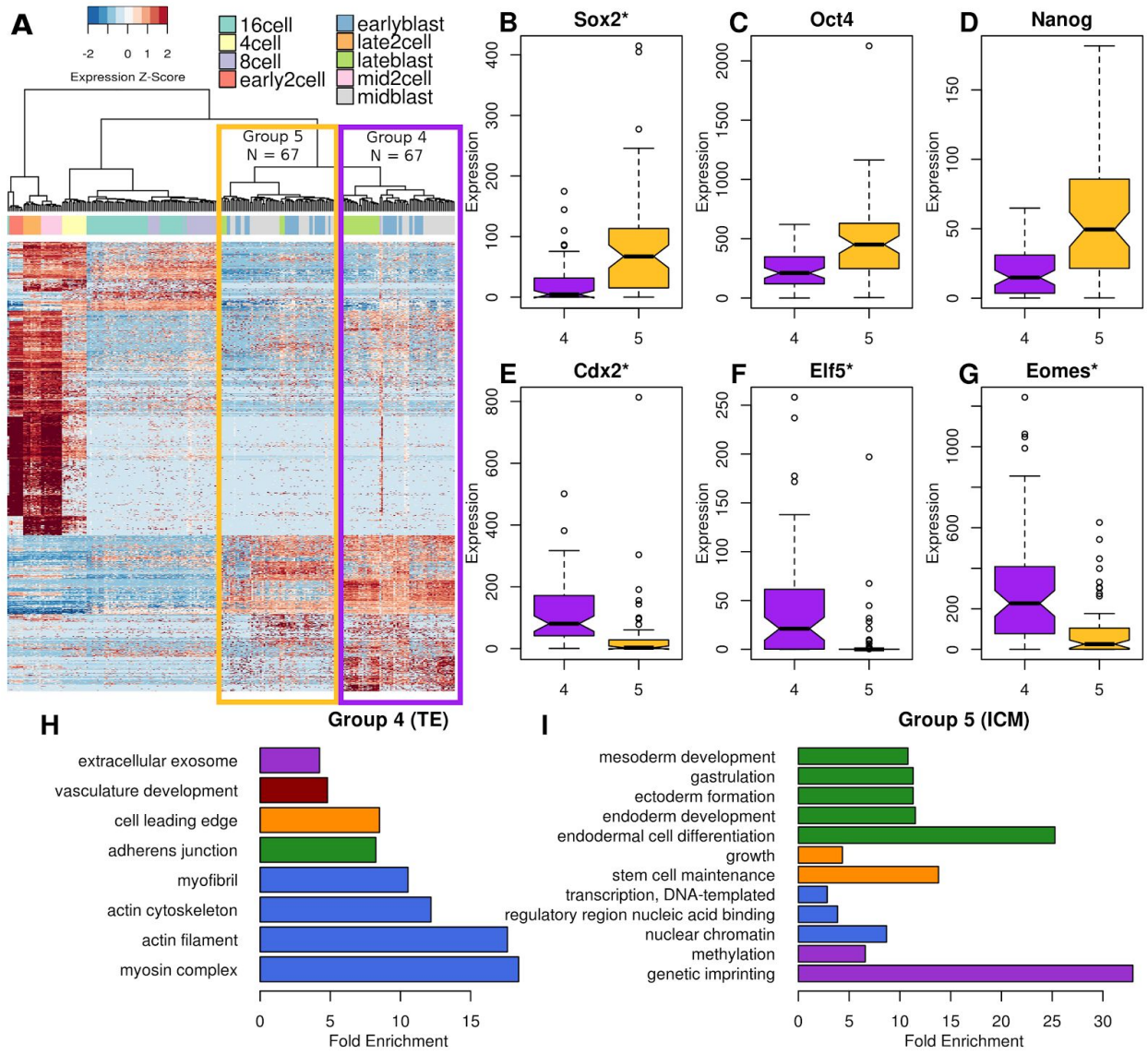

Figure 5

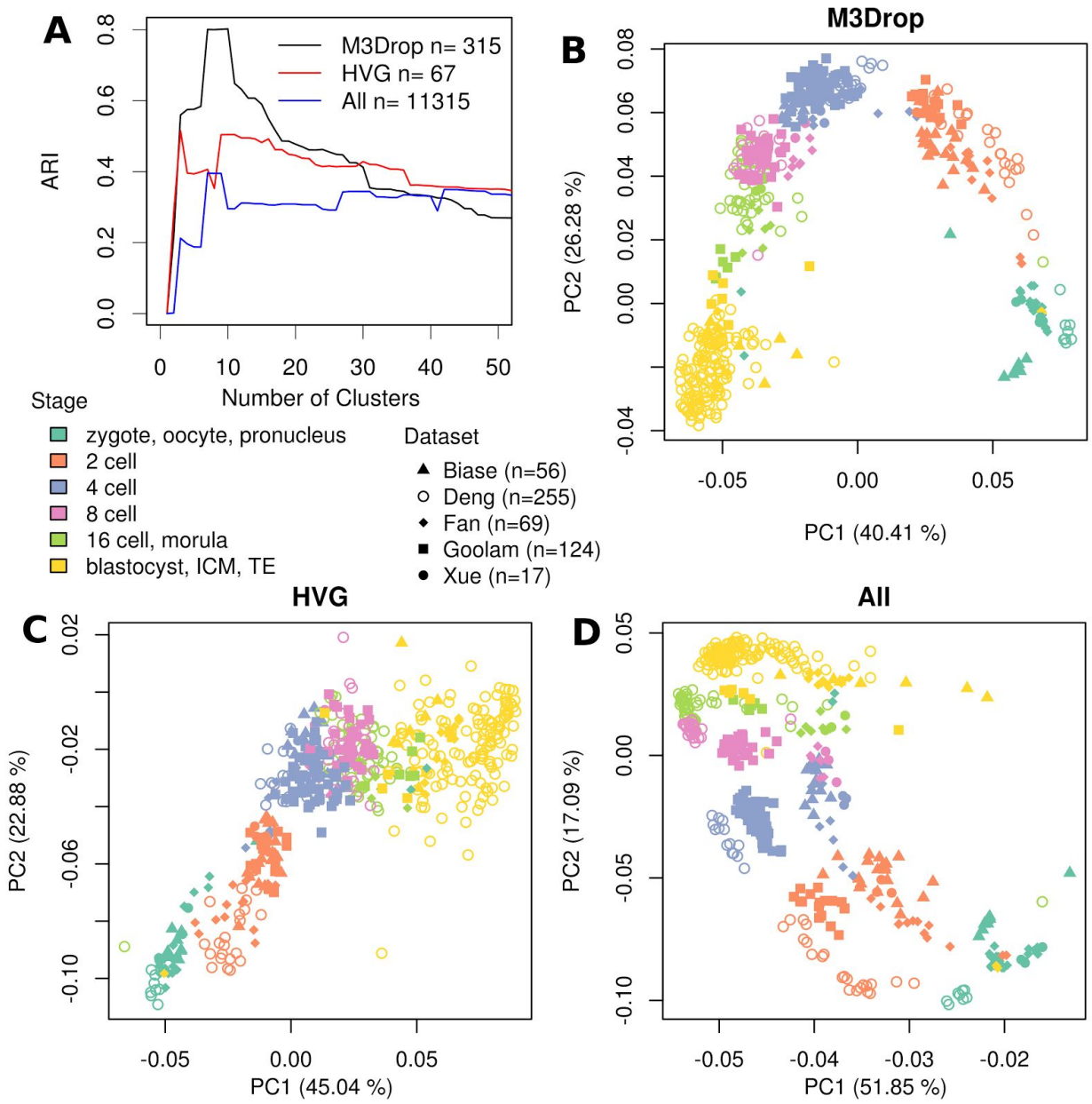

Figure 6