Modelling dropouts allows for unbiased identification of marker genes in scRNASeq experiments

Tallulah S. Andrews[1] and Martin Hemberg[1]
[1]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK

Abstract

Single-cell RNASeq (scRNASeq) differs from bulk RNASeq in that a large number of genes have zero reads in some cells, but relatively high expression in the remaining cells. We propose that these zeros, or dropouts, are due to failure of the reverse transcription, and we model the process using the Michaelis-Menten (MM) equation. We show that the MM equation provides an equivalent or superior fit to existing scRNASeq datasets compared to other models. In addition, identifying genes significantly to the right of the MM curve is a fast and accurate method to distinguish differentially expressed genes without prior identification of subpopulations of cells. We applied our method to a mouse preimplantation dataset and demonstrate that clustering the selected genes identifies biologically meaningful clusters. Furthermore, this feature selection makes it possible to overcome batch effects and cluster cells from five different datasets by their biological groups rather than by experimental origin.

Main

Single-cell RNASeq (scRNASeq) has made it possible to analyze the transcriptome from individual cells. However, single-cell data exhibits high technical variability and low detection rates compared to bulk data. This leads to what are known as “dropouts”, where a gene is not detected at all in some cells despite relatively high expression in others. Although some dropouts represent genes that were not expressed due to the underlying biological variability, a key step when analyzing scRNASeq data is to characterize the impact of the technical noise on the dropout events.

When analyzing gene expression data, one is often interested in feature selection, i.e. identifying the most relevant subset of genes. Two common examples of feature selection are identification of highly variable genes and identification of differentially expressed (DE) genes between two or more conditions. Highly variable genes vary more across cells than expected by a null model[1,2]. Several DE methods have been developed for bulk[3–5] and scRNA-seq[6,7], however the traditional assumption of well-defined experimental conditions may not be fulfilled in scRNA-seq experiments. Unlike bulk RNASeq, in scRNASeq experiment the groups of cells may not be known *a priori*. We present Michaelis-Menten Modelling of Dropouts (M3Drop) which uses the relationship between expression level and dropout rate (**Figure 1**) to select DE genes without making any assumptions regarding the underlying group structure. We demonstrate that this unbiased approach allows us to identify genes that are consistently differentially expressed between different cell-types, from experiments across different labs and protocols.

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[8], 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[9,10], 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[11], to model the relationship 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**).

To evaluate the MM model, we fit it to eight recent scRNASeq datasets (**Table S1**), and we find that it performs better than two previous models (**Figure 1 A,B**). Pierson and Yau[12] 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. This similarity is not surprising since the MM models is a special case of the logistic model with a coefficient of one (**Supplementary Note 1**). However, due to the high noise levels the 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**). This is evident from the Zeisel and Klein datasets where the use of unique molecular identifiers eliminates noise due to amplification which greatly reduces the total noise in the datasets and the fitted logistic and MM models are nearly indistinguishable (**Figure S3**).

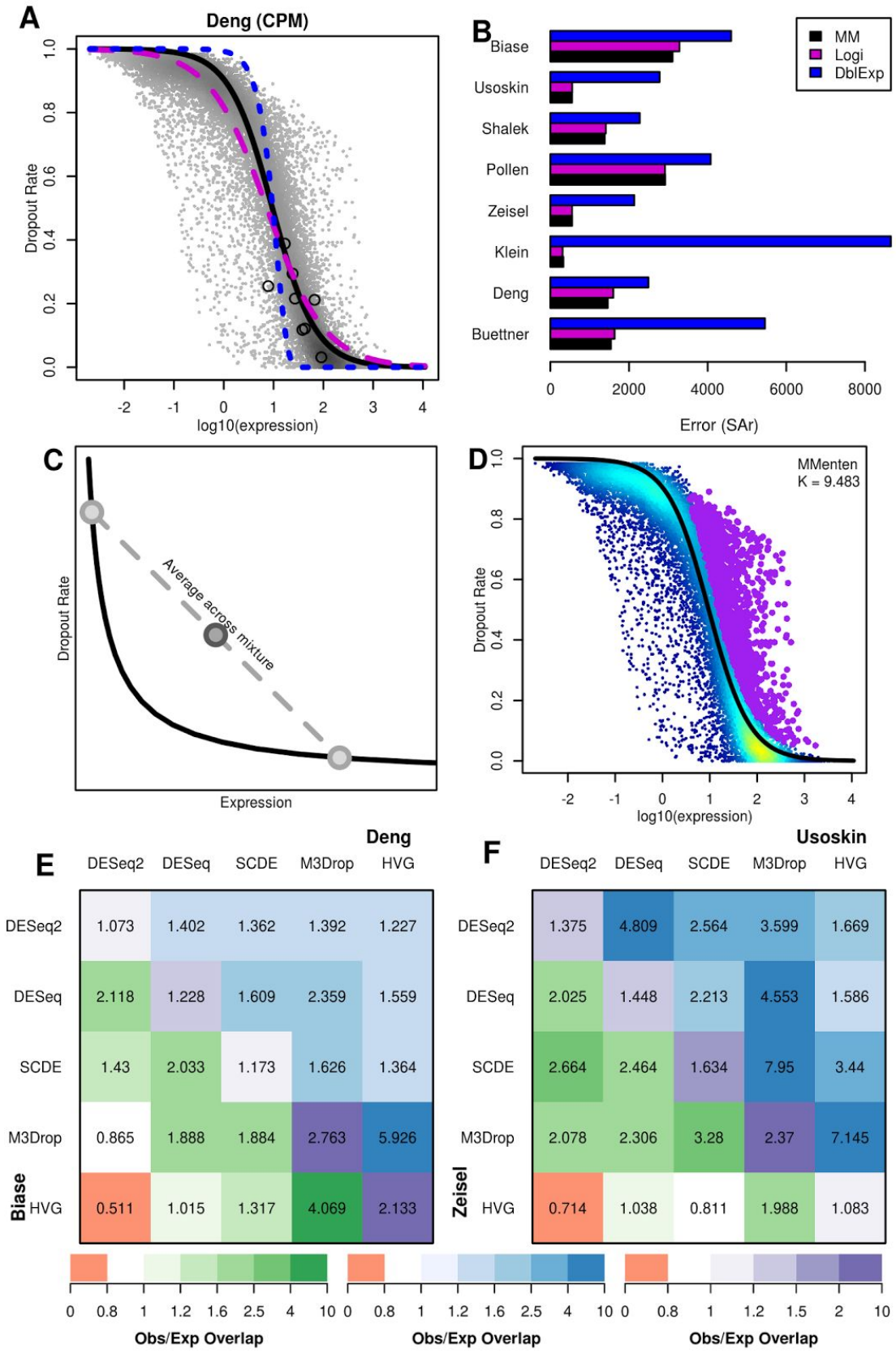

Figure 1: Michaelis-Menten identifies reproducibly differentially expressed genes. *(A) The Michaelis-Menten (solid black), logistic (dashed purple), and double exponential (dotted blue) models are fit to Deng dataset[13]. Expression (counts per million) was averaged across all cells for each gene (points) 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. (E & F) 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 (E green lower triangle), Deng (E blue upper triangle), Zeisel (F green lower triangle), and Usoskin (F 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 (E) and two datasets collected from mouse neuronal tissue (F). Cells in white are not significantly different from chance (obs/exp = 1) based on Fisher's exact test.*

An assumption of the MM model is that each gene has a uniform expression across cells. A shift to the right of the curve indicates that observed dropouts results from real differences in expression rather than reaching the detection limit (**Figure 1C, D**). Thus, the fitted curve can serve as a null model, and we can test for genes that are significantly to the right of the curve (**Methods**). We refer to the method as M3Drop and it allows us to identify genes that are DE between subsets of the cells. We compared the performance of M3Drop to three existing DE methods: DESeq[4], DESeq2[3], and SCDE[7], in addition to a method for identifying highly variable genes (HVG)[14]. 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 S4B**). In contrast, DESeq2 and SCDE took several hours. DESeq, DESeq2 and SCDE have a tendency to report ~50% of genes as DE, many of which are known housekeeping genes or external spike-ins, whereas M3Drop and HVG tend to be more conservative and report ~10% of the genes (**Figure S4D**). Consequently, M3Drop and HVG have higher enrichments of true positive genes and lower false positive rates for six of the datasets (**Figure S4A,C**). The 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 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 much more consistent across datasets which examined the same biological system than the other methods (**Figure 1E, F**). We found that M3Drop, DESeq2, DESeq, and SCDE are all significantly consistent with each other within datasets ($p < 10^{-50}$). In contrast, 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 ($p < 10^{-30}$). Importantly, M3Drop was significantly more reproducible across datasets than the other methods with [2.37, 3.29]-fold enrichment (95% CI) between Deng and Biase datasets and [1.99, 2.91]-fold enrichment between Usoskin and Zeisel datasets. In contrast, DESeq, DESeq2, and SCDE were only 1.069 to 1.634 fold enriched between datasets. HVG was inconsistent between the two pairs of dataset with 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.

The main advantage of using M3Drop to identify DE genes is that no assumptions about the underlying groups is required. Thus, we asked if the DE genes identified by M3Drop were informative for identifying cell subpopulations. We applied M3Drop to a set of 255 cells across early mouse development[13]. There were 1,478 significantly differentially expressed genes identified at 1% FDR (**Figure 1D**). Clustering the cells using Ward's hierarchical clustering[15] recapitulates the known sampling times of the cells (**Figure 2A**). 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) and trophectoderm (TE). To test this hypothesis, we compared the expression levels of six known marker genes (**Figure 2 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[16–18]. This enrichment was not observed when blastocyst cells were grouped by time-point (**Figure S5**).

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[19]. 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 which would be important for implantation of the embryo and invasion of the uterine wall (**Figure 2H, Table S4**). In contrast, the putative ICM cells expresses genes involved in DNA methylation/demethylation, stem cell maintenance, and gastrulation (**Figure 2I, Table S4**), supporting existing evidence that DNA methylation is reset during the blastocyst stage[20].

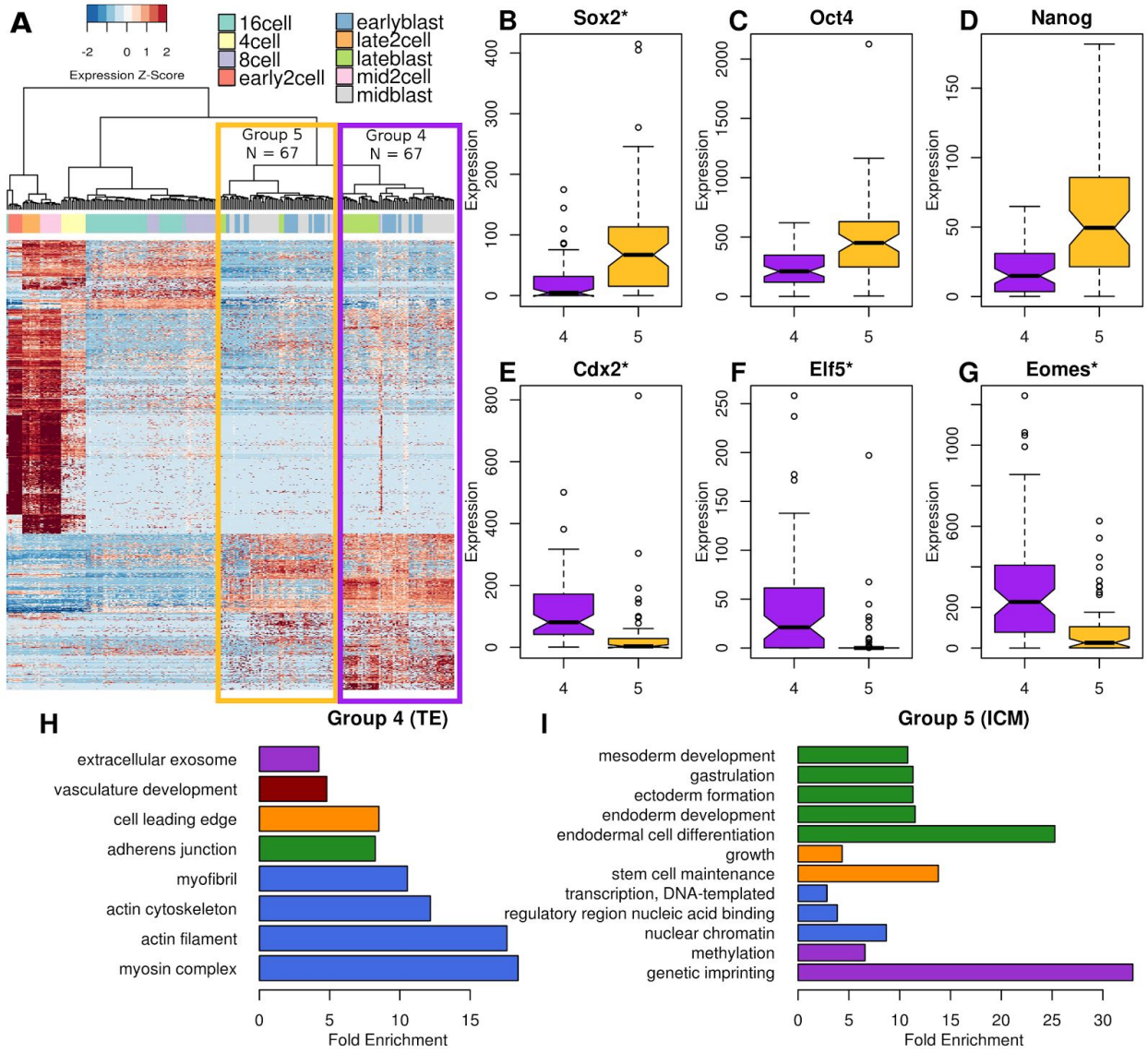

Figure 2: Identification of biologically relevant clusters from a development time-course. *(A) Expression of the 1,478 DE genes identified by M3Drop (1% FDR) in the Deng dataset reveals two clusters of cells across all three blastocyst timepoints using Ward's hierarchical clustering. (B-D) Expression 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.*

As an additional application of M3Drop feature selection, we consider the problem of merging datasets from different labs. We obtained three additional mouse preimplantation datasets[21–23] and we pooled those cells with the ones from 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 3**). By contrast using the genes identified by HVG or the full set of genes results in a clustering where the cells are grouped by batch instead (**Figure S7-9**).

Amongst these 316 genes were: Zscan4d, which is known to have highly specific expression in 2-cell embryos [24], H1foo, an oocyte-specific histone, Tdgf1 which is required for gastrulation[25], the oocyte-specific growth factor Bmp15[26] and Bhmt which is involved in mouse embryonic methylation[27]. Four of the six Obox gene family members which have been previously associated with oogenesis[28] were amongst this set of genes, Obox1, Obox3, Obox5, and Obox6. However, only Obox1 and Obox5 were preferentially expressed among zygotes and oocytes (**Table S5**). 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.

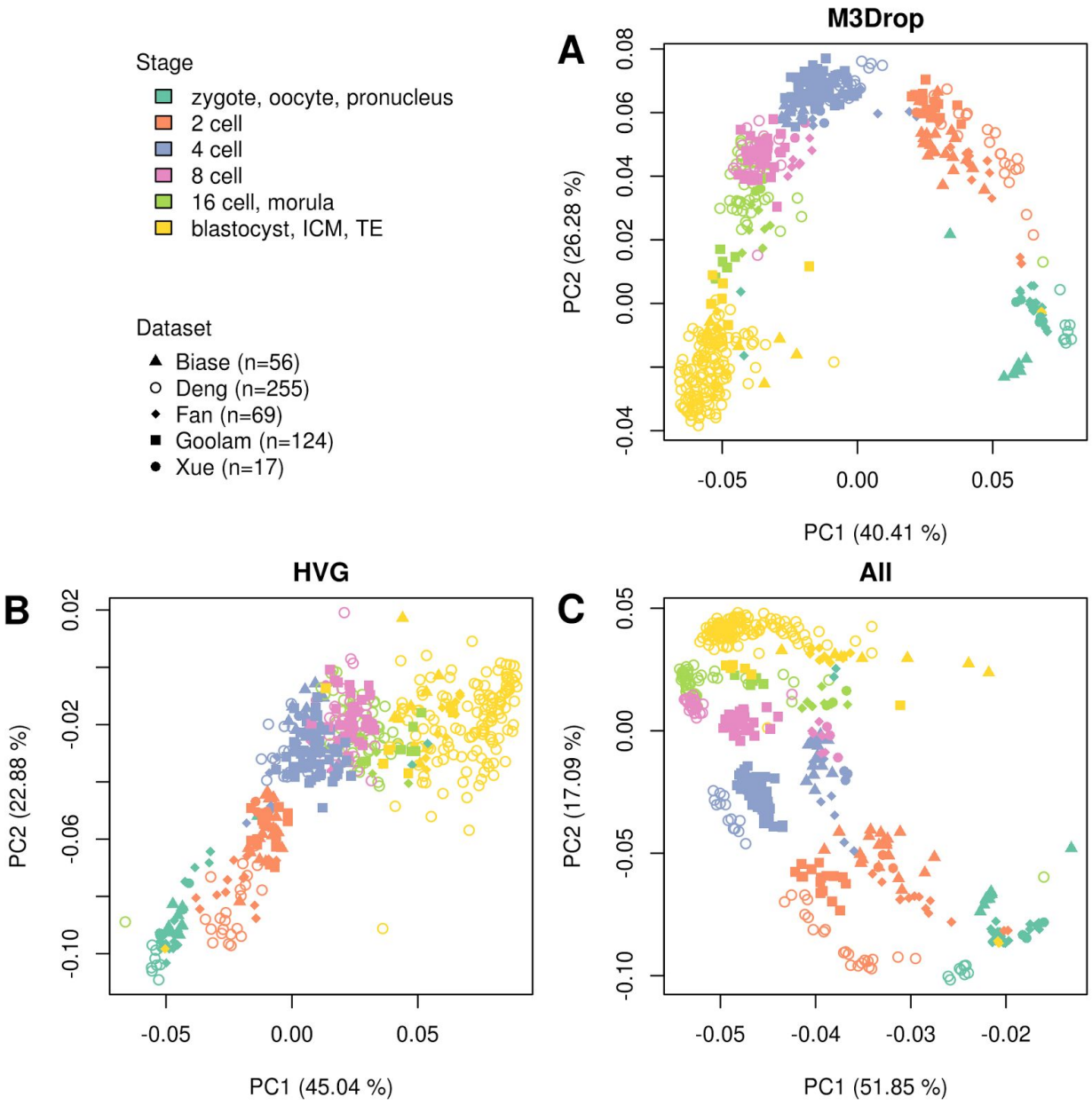

Figure 3: M3Drop marker genes makes it possible to merge datasets from different experiments. *Principle component analysis based on the genes detected by M3Drop (A) or HVG (B) in both Deng and Biase datasets or all detected genes (C).*

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. For each dataset cells with fewer than 2000 detected genes (after excluding processed pseudogenes) were removed. Genes detected in fewer than 4 cells or with average normalized expression < 10^{-5} were excluded from consideration. Finally all genes annotated as processed pseudogenes in Ensembl (version 80) were excluded. For the Deng data only single mouse embryo cells analyzed using the SmartSeq protocol were considered to avoid technical artefacts. For the Shalek dataset to facilitate the identification of true positive DE genes only the two replicates of Unstimulated and after 4h LPS stimulation were considered; in addition technical artefacts as described by the authors were removed[29].

Table 1: Eight publicly available datasets.

Dataset	Cell-types	Original Labels	Method	N	Source
Buettner	Mouse ESC	Cell-cycle stage	Smartseq Counts -> (CPM)	279	[30] E-MTAB-2805
Deng	Mouse embryos	Developmental timepoint	Smartseq Counts -> (CPM)	255	[13] GSE45719
Usoskin	Mouse neurons	PCA-based clustering	5' Seq Counts -> (CPM)	530	[31] GSE59739
Klein	Mouse ESC	Differentiation timepoint	CEL-Seq UMIs -> (CPM)	2448	[32] GSE65525
Zeisel	Mouse brain	BackSPIN clustering	5' Seq UMIs -> (CPM)	2542	[33] GSE60361
Shalek	Mouse bone marrow cells	Stimulated & unstimulated	Smartseq FPKMs	173	[29] GSE48968
Pollen	Human cell lines & tissues	Cell line identity	Smartseq FPKMs	301	[34] SRP041736**
Biase	Mouse embryos	Developmental stage	Smartseq FPKMs	56	[35] GSE57249

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

*** Processed data was provided by the authors*

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. The double exponential model[12]:

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

was fit by log transforming the equation then using the lm R function to fit the coefficient to the resulting quadratic model:

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

Differential Expression (DE)

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

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

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

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

The K_j's were log-normally distributed thus we tested each one against the fitted K_M using a one-sided Z-test for the log-transformed Ks (Eqn 5). The error for K_M was estimated as the standard error of the residuals and added to the error on each K_j (Eqn 3,4).

$$\sigma_{log(K_M)} = \frac{sd(log(K_j) - log(K_M))}{\sqrt{N}} \quad (3)$$

$$\sigma_{log(K_j)} = log(K_j) - log(K_j - \sigma_{K_j}) \quad (4)$$

$$Z = \frac{(log(K_j) - log(K_M)}{\sqrt{\sigma^2_{log(K_j)} + \sigma^2_{log(K_M)}}} \quad (5)$$

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 1000 cells were downsampled to 1000 cells. 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 using parameters max.pairs to 500 and min.pairs.per.cell 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 a 1% FDR.

Highly Variable Genes (HVG) were calculated using the published method[14]. However, rather than fitting the model to the spike-ins, of which there were less 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)[30] 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[36].

De novo identification of cell types was performed using Ward’s hierarchical clustering based on the expression of the DE/HVG genes. The resulting dendrogram was cut to produce the same number of groups as described in the original paper (+/- 3) and compared to the groups described in the original publication using the adjusted Rand index (ARI)[37]. ARI is calculated as the proportion of all possible pairs of cells which are consistently in the same group or in different groups 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.

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 proportion of all detected genes that were identified as DE in each case. Significance was evaluated using Fisher's exact test.

Preimplantation Mouse Development

M3Drop was applied to the 255 cells of the Deng dataset. Cells were clustered using Ward's hierarchical clustering[15], 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 which of the blastocyst groups each cell belonged to [19]. 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[21–23]. Each dataset was filtered for quality using identical criteria and the Deng and Goolam data was normalized using CPM, all other datasets were already FPKM/RPKM normalized. Filtering criteria were the same as before (see above). 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.

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)[37].

Code Availability

M3Drop is freely available on github : https://github.com/tallulandrews/M3Drop
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).

References

1. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1095 (2013).
2. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
3. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
4. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** (2010).
5. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **8,** 1765–1786 (2013).
6. Sengupta, D. *et al. Fast, scalable and accurate differential expression analysis for single cells*. (2016). doi:10.1101/049734
7. Kharchenko, P. V., Lev, S. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–742 (2014).
8. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17,** 63 (2016).
9. Reiter, M. *et al.* Quantification noise in single cell experiments. *Nucleic Acids Res.* **39,** e124 (2011).
10. 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.* **9,** 63 (2008).
11. Michaelis, L. & Menten, M. L. Die Kinetik der Invertinwirkung. *Biochem. Z.* **49,** 333–369

(1913).

12. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16,** 241 (2015).
13. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193–196 (2014).
14. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1095 (2013).
15. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58,** 236–244 (1963).
16. Hamatani, T., Toshio, H., Carter, M. G., Sharov, A. A. & Ko, M. S. H. Dynamics of Global Gene Expression Changes during Mouse Preimplantation Development. *Dev. Cell* **6,** 117–131 (2004).
17. Marikawa, Y. & Alarcón, V. B. Establishment of trophectoderm and inner cell mass lineages in the mouse embryo. *Mol. Reprod. Dev.* **76,** 1019–1032 (2009).
18. Chen, Y., Wang, K., Gong, Y. G., Khoo, S. K. & Leach, R. Roles of CDX2 and EOMES in human induced trophoblast progenitor cells. *Biochem. Biophys. Res. Commun.* **431,** 197–202 (2013).
19. Kiselev, V. Y. *et al. SC3 - consensus clustering of single-cell RNA-Seq data*. (2016). doi:10.1101/036558
20. Messerschmidt, D. M., Knowles, B. B. & Solter, D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev.* **28,** 812–828 (2014).
21. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593–597 (2013).

22. Fan, X. *et al.* Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16,** 148 (2015).
23. Goolam, M. *et al.* Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165,** 61–74 (2016).
24. Falco, G. *et al.* Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Dev. Biol.* **307,** 539–550 (2007).
25. Jin, J.-Z. & Ding, J. Cripto is required for mesoderm and endoderm cell allocation during mouse gastrulation. *Dev. Biol.* **381,** 170–178 (2013).
26. Mottershead, D. G. *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.* **290,** 24007–24020 (2015).
27. Lee, M. B. *et al.* Betaine Homocysteine Methyltransferase Is Active in the Mouse Blastocyst and Promotes Inner Cell Mass Development. *J. Biol. Chem.* **287,** 33094–33103 (2012).
28. Rajkovic, A., Yan, C., Yan, W., Klysik, M. & Matzuk, M. M. Obox, a family of homeobox genes preferentially expressed in germ cells. *Genomics* **79,** 711–717 (2002).
29. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369 (2014).
30. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33,** 155–160 (2015).
31. Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18,** 145–153 (2014).
32. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).

33. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–1142 (2015).

34. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053–1058 (2014).

35. Biase, F. H., Cao, X. & Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **24,** 1787–1796 (2014).

36. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29,** 569–574 (2013).

37. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66,** 846 (1971).

Supplementary Material

Supplementary Note 1 - Alternatives to Michaelis-Menten

The Michaelis-Menten function is a special case of the logistic function where the coefficient is set to 1 and the intercept is the natural logarithm of K_M. Thus the Hill equation, $P_{dropout} = 1-S^n/(K_d+S^n)$, which is a variant on the Michaelis-Menten used to model cooperative binding is exactly equivalent to the logistic function on the log expression where the coefficient is n and the intercept is $\ln(K_d)$. We also considered a double Michaelis-Menten, $P_{dropout} = 1-S^2/[(K_1+S)(K_2+S)]$, where both $K_1 > 0$ and $K_2 > 0$; however the fitted second K was < 0.001 for all but the Zeisel data (**Table S1**) and K_1 was almost the same as for the MM model.

Table S1: Fitting the single (K_M) & double Michaelis-Menten (K_1, K_2)

	Buettner	Deng	Usoskin	Klein	Zeisel	Shalek	Pollen	Biase
K_M	10.3	9.5	49.7	39.3	75.8	22.9	11.3	1.9
K_1	9.7	7.3	48.5	39.4	72.4	18.7	11.3	2.3
K_2	$5x10^{-4}$	$9x10^{-4}$	$4x10^{-4}$	$9x10^{-4}$	1.09	$9x10^{-4}$	$3x10^{-4}$	$8x10^{-4}$

Supplementary Note 2 - Calculating Average Expression:

We average expression overall cells including those with zero reads for a particular gene whereas in [12] average expression was calculated by averaging only over non-zero expression

values. If there is a strong relationship between dropout rate and expression level then it implies that the presences of dropouts is indicative of low expression. Thus, excluding zeros when calculating average expression would lead to overestimation of the expression level for rarely detected genes. We attempted to fit the three different models, Michaelis-Menten, Logistic, and double exponential, to the dropout rate vs average expression when calculated excluding zeros (**Figure S1**). Fitting quality was highly variable across datasets; for the UMI datasets only the logistic function fit well, whereas for Shalek, Pollen and Biase datasets the double exponential function was the best fit, followed closely by the Michaelis-Menten. For the Deng and Buettner datasets all three models fit reasonably well.

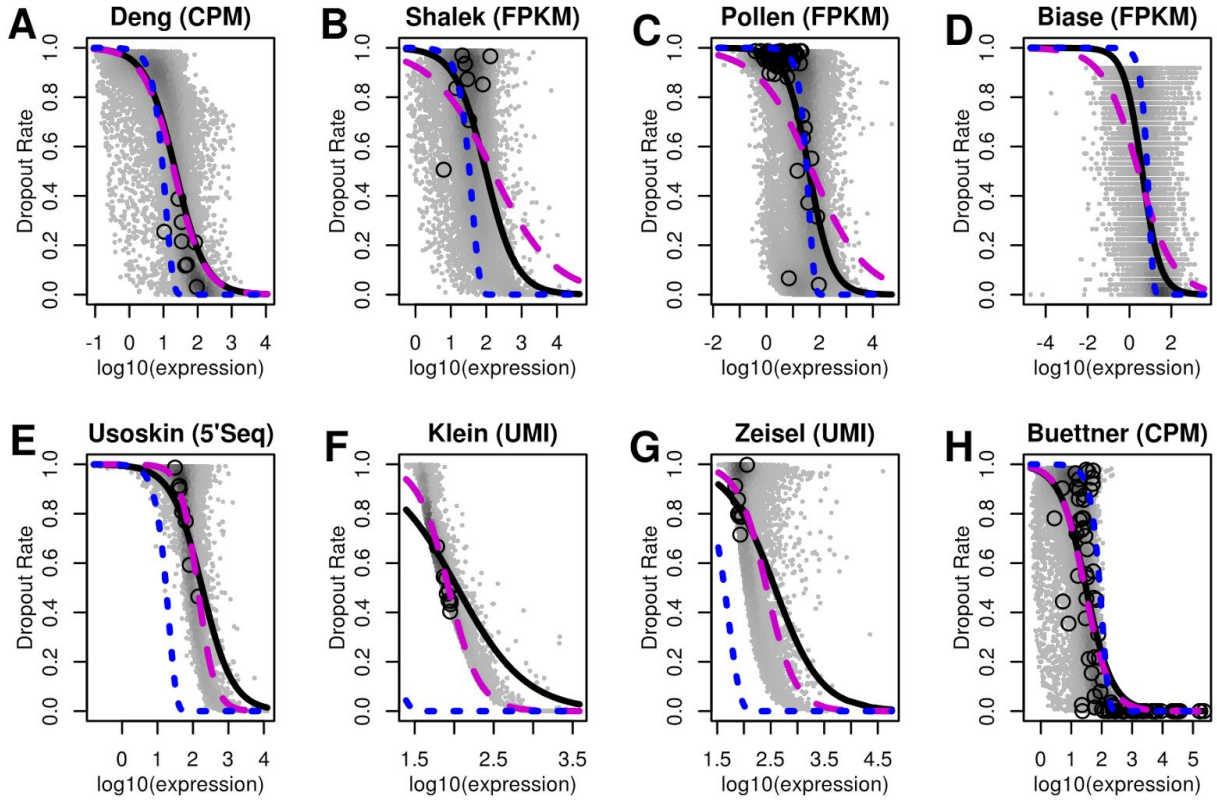

Figure S1 *The Michaelis-Menten (solid black), logistic (dashed purple), and double exponential (dotted blue) models fit to data when average expression is calculated using only expression values > 0. Black circles indicate spike-in RNAs.*

Supplementary Note 3 - Interpretation of K_M

The fitted Michaelis constant (K_M) is the mean level of expression for a gene to be detected in 50% of cells on average. This is related to the detection rate, a higher K_M indicates lower detection rate, which in turn is related to sequencing depth (**Figure S2**). However, it is also related to the single-cell RNASeq method employed. Full transcript SMART-seq datasets had

lower a K_M than the datasets which sequenced a single end of each transcript; and datasets employing UMIs had high K_M relative to the number of detected genes.

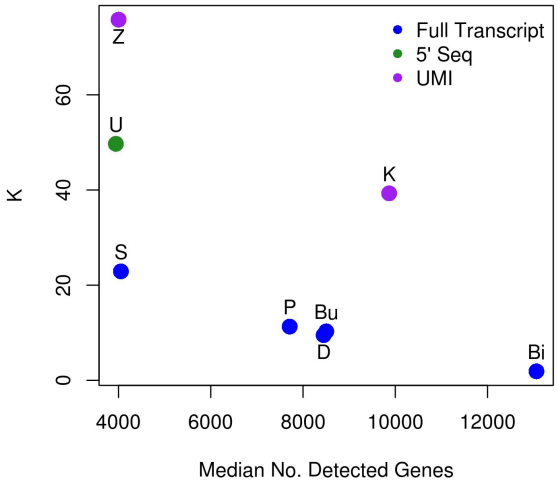

***Figure S2**: K_M is related to detection rate and sequencing method. Datasets are labelled by the first letter(s) of the author's name.*

Supplementary Note 4 - Differential Expression in publicly available scRNASeq datasets. In order 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.

Table S2: *Groups used for pairwise differential expression analysis*

Buettner	**Shalek**	**Deng**	**Usoskin***	**Klein****	**Zeisel**	**Pollen**	**Biase**
G1	Unstimulated	2-4 cell	NP	Day 0	Neuron	Skin	zygote
S	4h LPS	8-16 cell	TH	Day 2	Glia	Blood	2 cell
G2/M		blast	NF	Day 4	Microglia	hiPSC	4 cell
			PEP	Day 7	Non-Neuro	Neuronal	ICM
							TE

**NP = nonpeptidergic nociceptors, TH = tyrosine hydroxylase expressing, NF = neurofilament expressing, PEP = peptidergic nociceptors*

***ESCs following LIF withdrawal*

Supplementary Figures

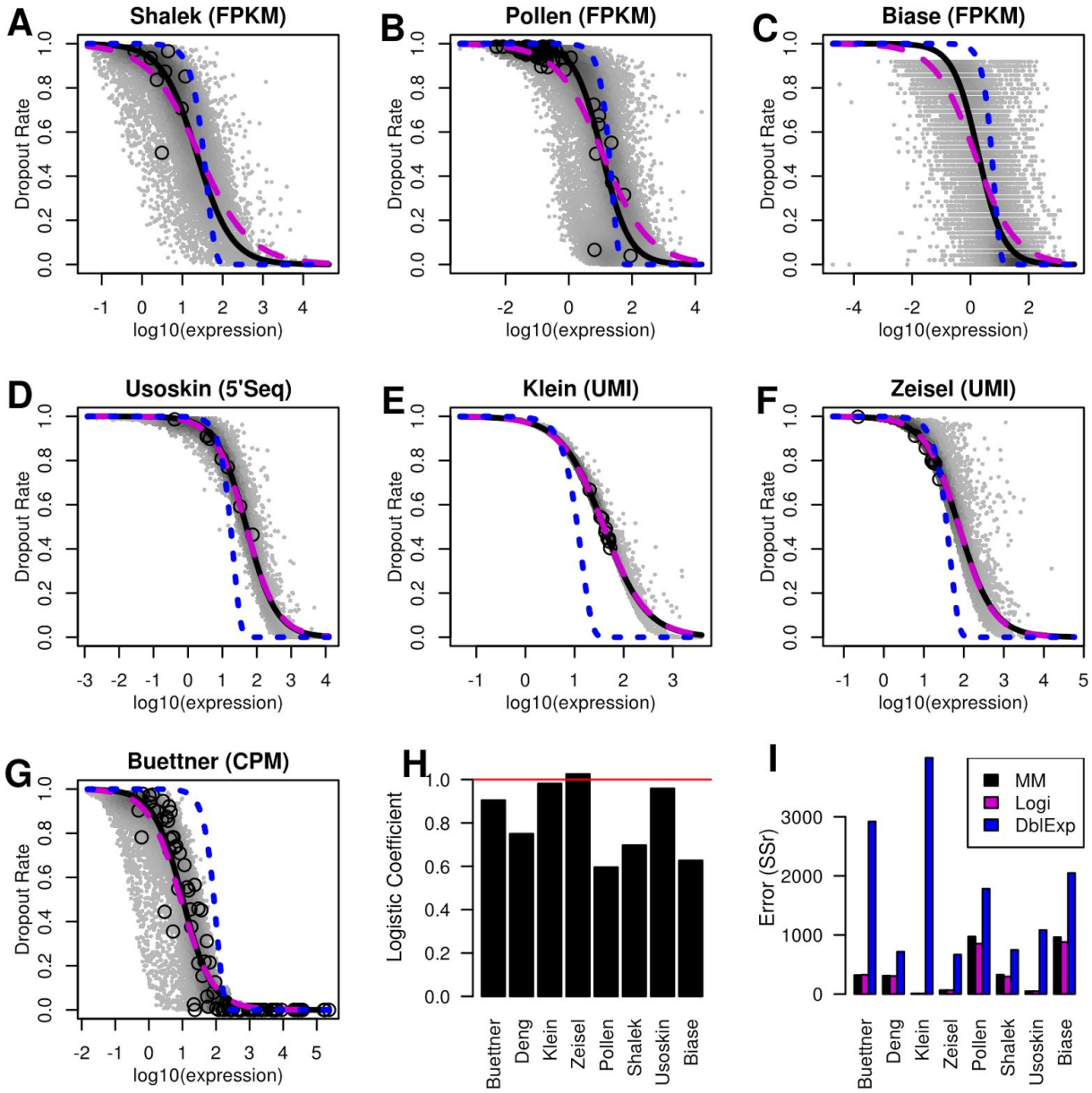

Figure S3 (A-G)*The Michaelis-Menten (solid black), logistic (dashed purple), and double exponential (dotted blue) models are fit to the other five published datasets. Black circles indicate spike-in RNAs. (H) Due to noise in the data the logistic regression fits a coefficient < 1 for most datasets giving a flatter curve compared to Michaelis-Menten. (I) Logistic regression had the smallest sum of squared residuals (SSr) across all eight datasets considered.*

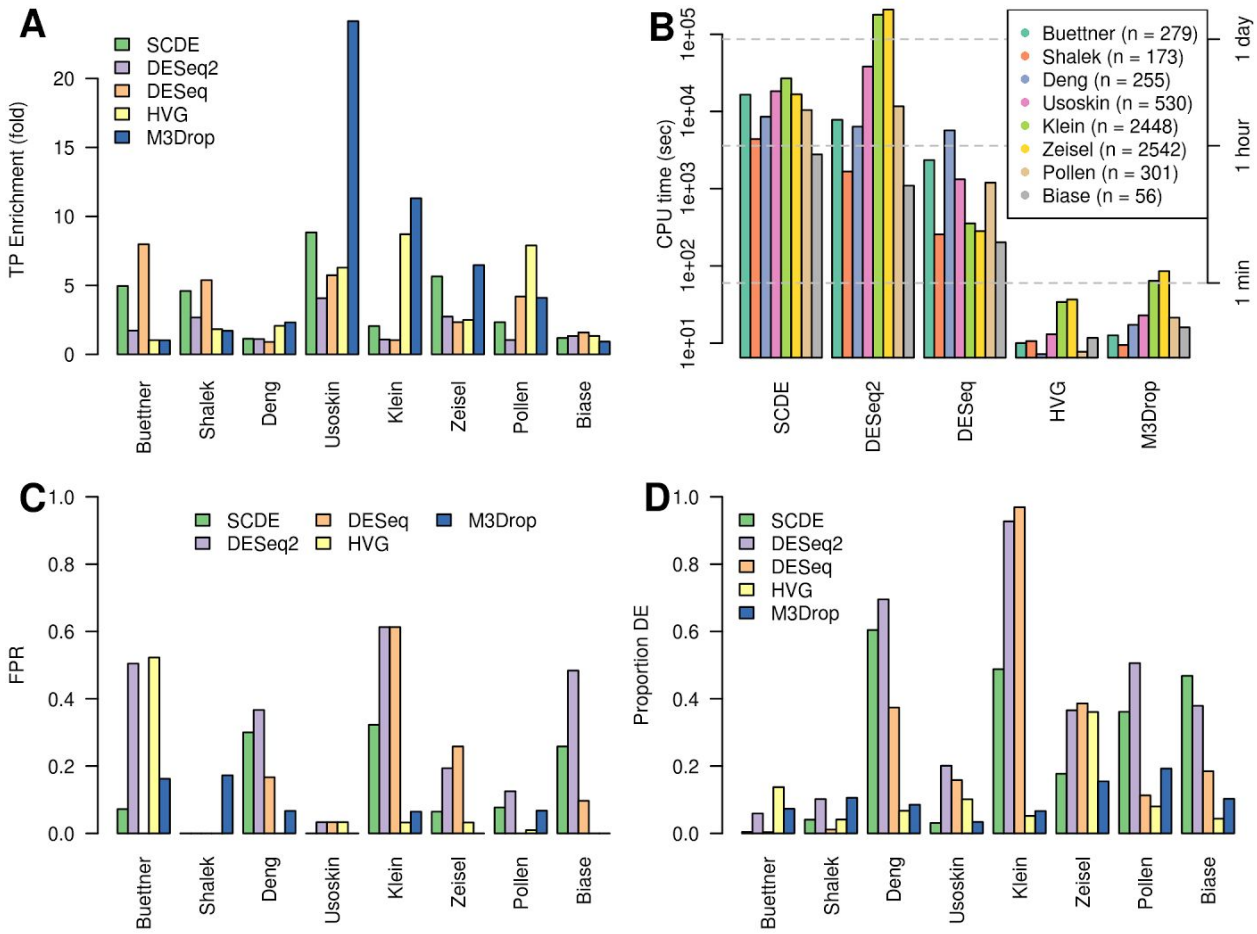

Figure S4 *(A) Fold enrichment of true positive (cell-type markers or members of relevant pathways) among DE genes for each dataset. (B) HVG and M3Drop are the most computationally efficient methods. SCDE was limited to 500 cells and DESeq2 was limited to 1000 cells both were run in parallel on 10 processors. DESeq HVG and M3Drop were run on the full datasets on a single processor. (C) Proportion of housekeeping genes and spike-ins called as DE (false positive rate) for each method. (D) Proportion of all detected genes that were called as DE (1% FDR)*

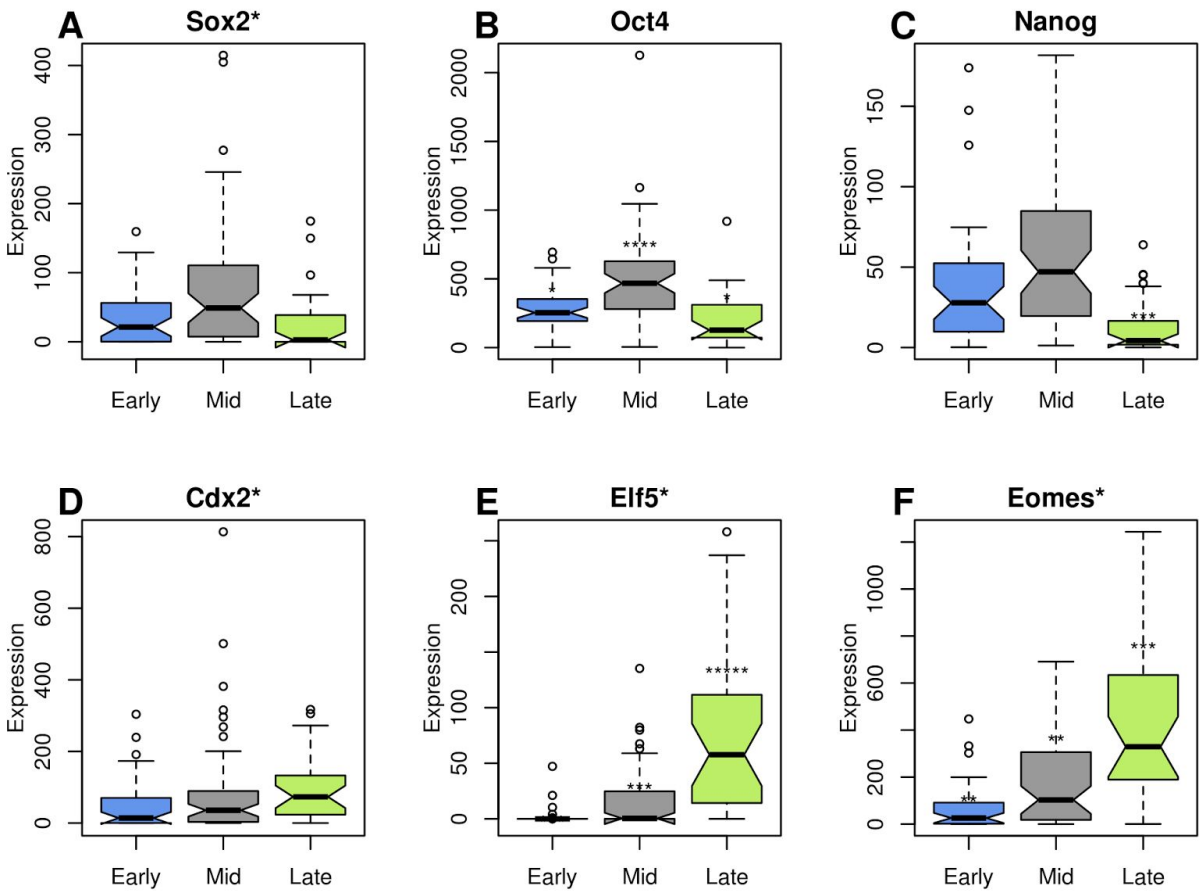

Figure S5 *(A-C) Expression of markers of the inner cell mass (ICM) for the three blastocyst timepoints. (D-F) Expression of markers of the trophectoderm (TE), Only the late blastocyst timepoint is significantly different from the other timepoints for multiple markers. This is due to a higher proportion of TE cells (**Figure S6**). Stars indicate order of magnitude of significant differences from other groups from a Wilcox rank-sum test (one star indication p < 0.01).*

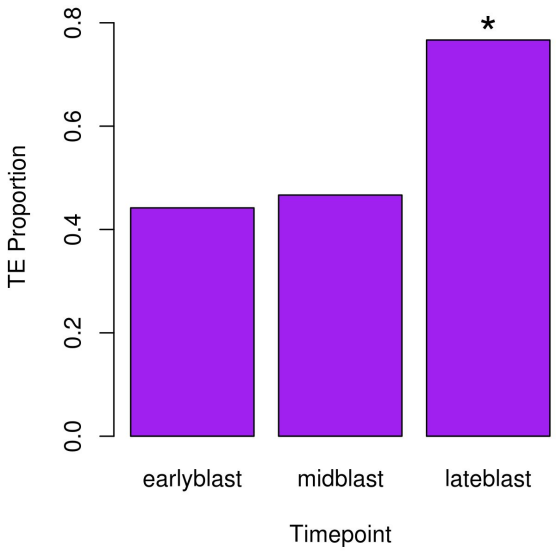

Figure S6 *Late blastocysts were significantly enriched in TE (Group 4) cells compared to early and mid stage blastocysts. Star indicates p < 0.01 using Fisher's exact test.*

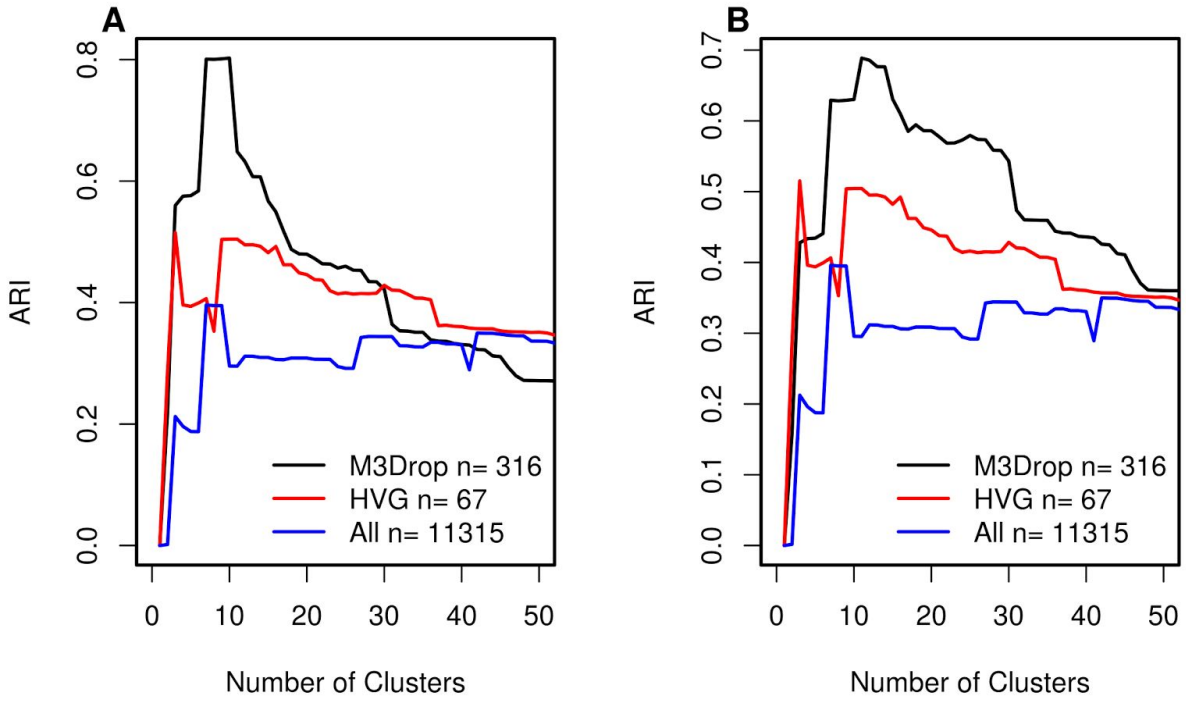

Figure S7: *Clustering 521 cells from four different datasets using genes selected by M3Drop (black), HVG (red) or all detected genes (blue). Clusters were compared to the true developmental stages: zygote, 2-cell, 4-cell, 8-cell, 16-cell, and blastocyst (A) or TE & ICM (B)*

using the Adjusted Rand Index across a range of k. Random chance: ARI = 0, identical clustering: ARI = 1.

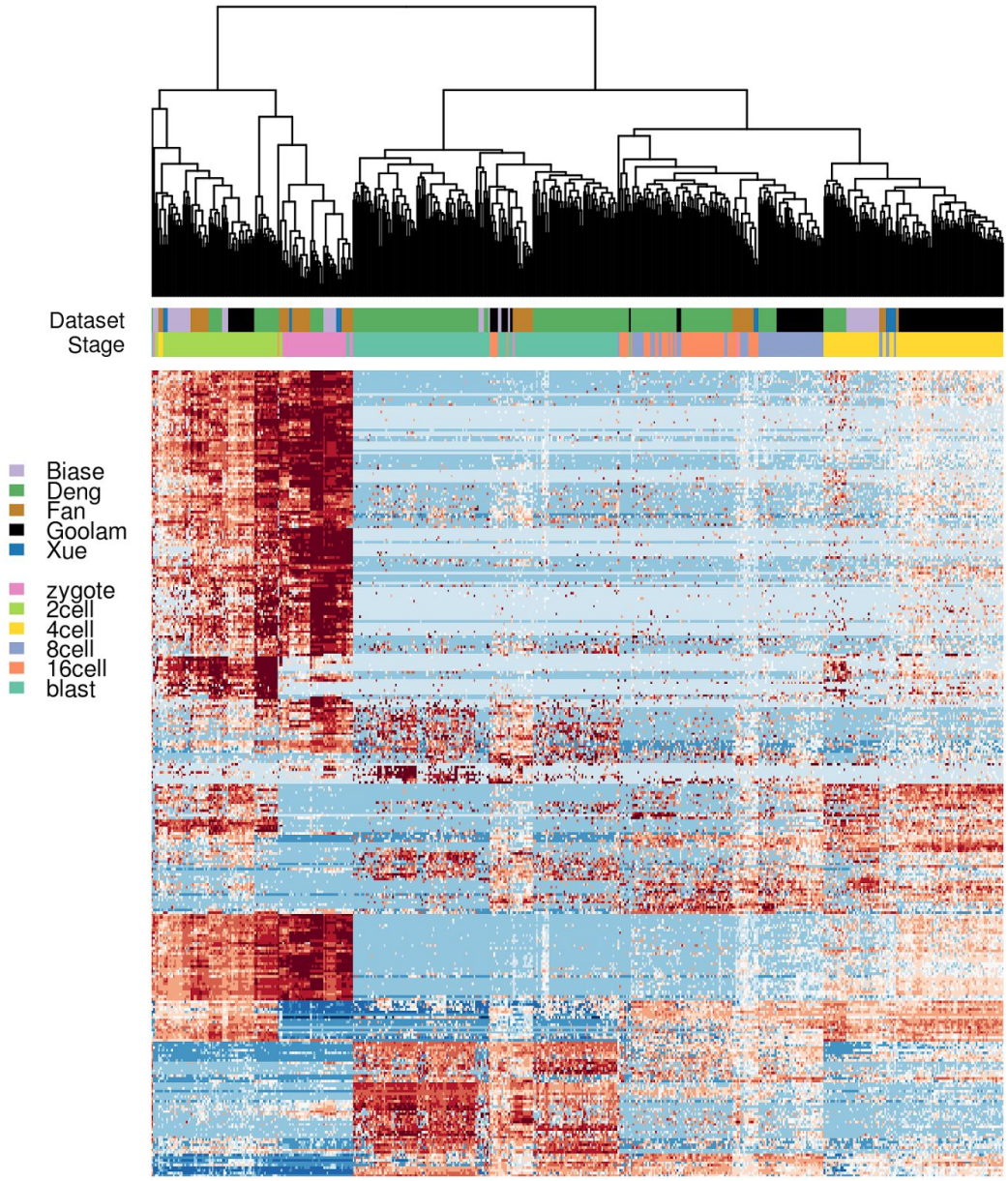

Figure S8: M3Drop marker genes makes it possible to merge datasets from different experiments. *Expression of 316 differentially expressed using M3Drop genes seen in both Deng and Biase datasets across five mouse preimplantation datasets.*

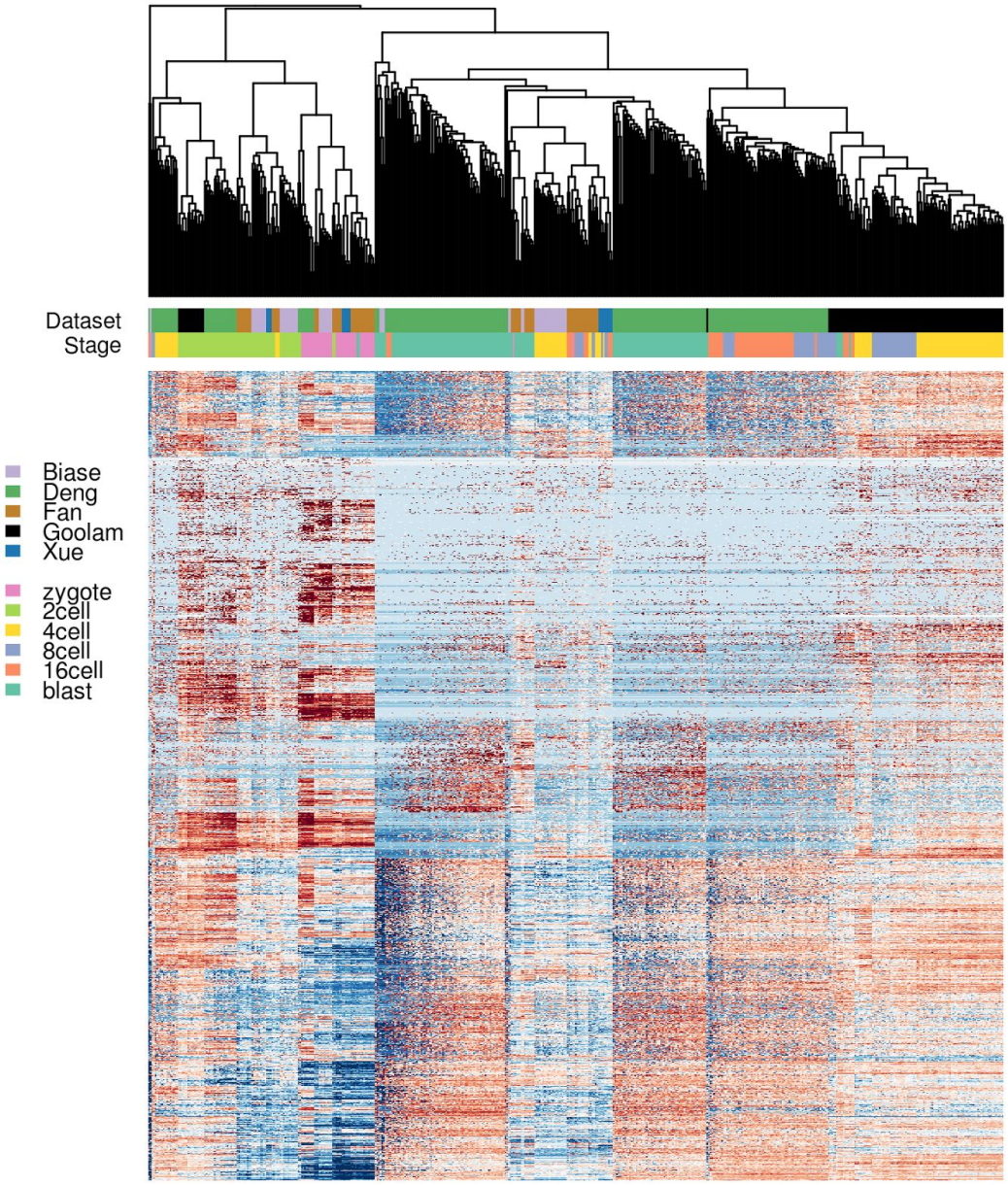

Figure S9: *Expression of all 11,440 genes detected in all datasets across five mouse preimplantation datasets.*