Supplementary Tables

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 Figures

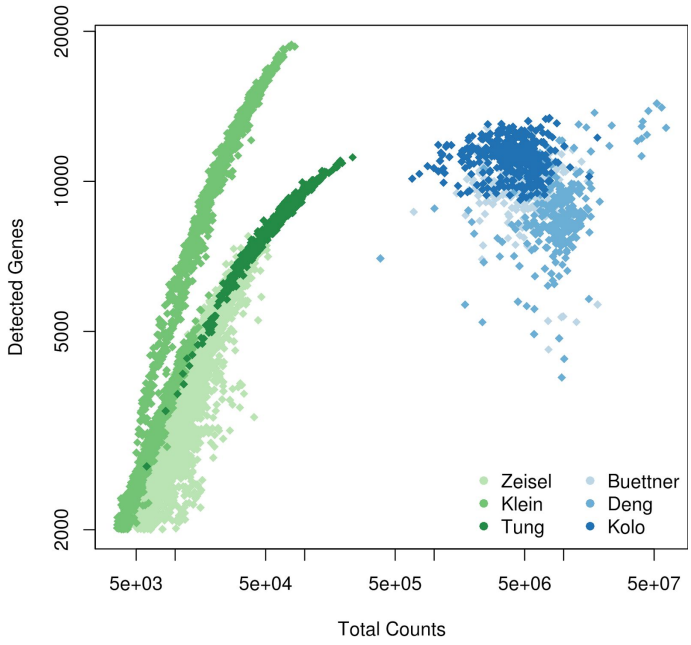

Figure S1: Importance of sequencing depth/tagging efficiency in UMI-tagged (green) vs full-transcript (blue) datasets.

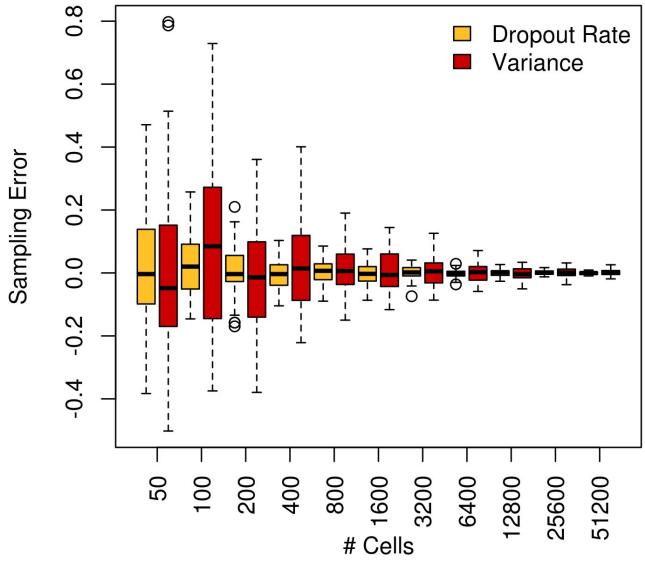

Figure S2 Sampling errors affect variance more than dropout rate. Expression was simulated using the DANB model for a gene with a mean expression of 1 molecule per cell for 1 million cells. 50 samples of different sizes were selected and error in the sample variance and dropout rate was calculated relative to that observed for all 1 million cells.

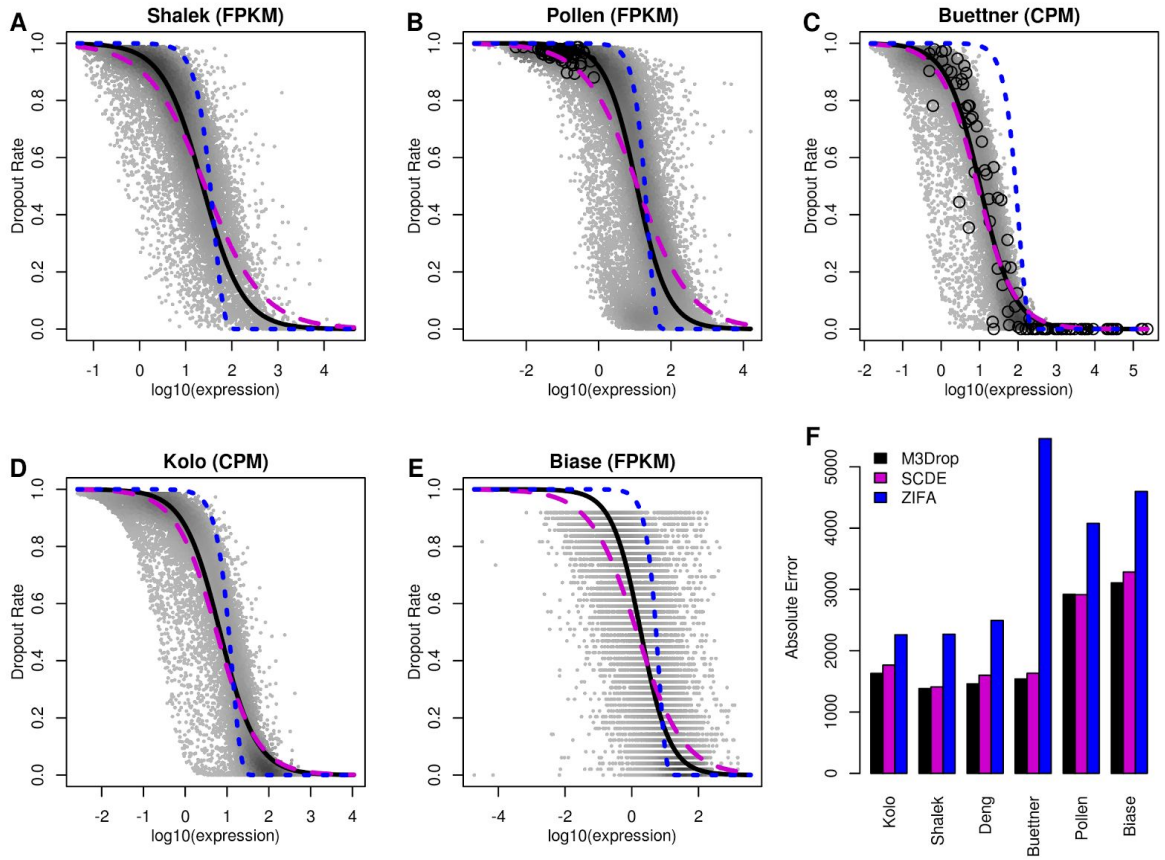

Figure S3 (A-E) The Michaelis-Menten (M3Drop, solid black), logistic (SCDE, dashed purple), and double exponential (ZIFA, dotted blue) models are fit to the other five published datasets. Black circles indicate spike-in RNAs. (F) M3Drop fits full-transcript data better than the alternatives.

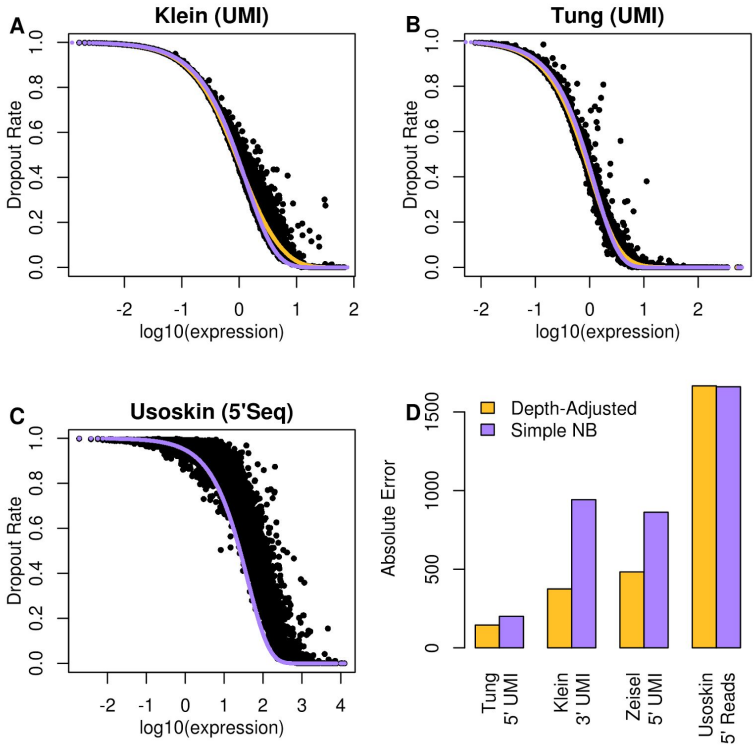

Figure S4 The depth-adjusted negative binomial (DANB) fits UMI-tagged datasets better than a simple negative binomial model. (A-C) Expected relationship as fit by each model (purple & yellow) compared to observed relationships using the globally fit relationship between mean & dispersion. (D) Sum of absolute error between observed and expected dropout rates.

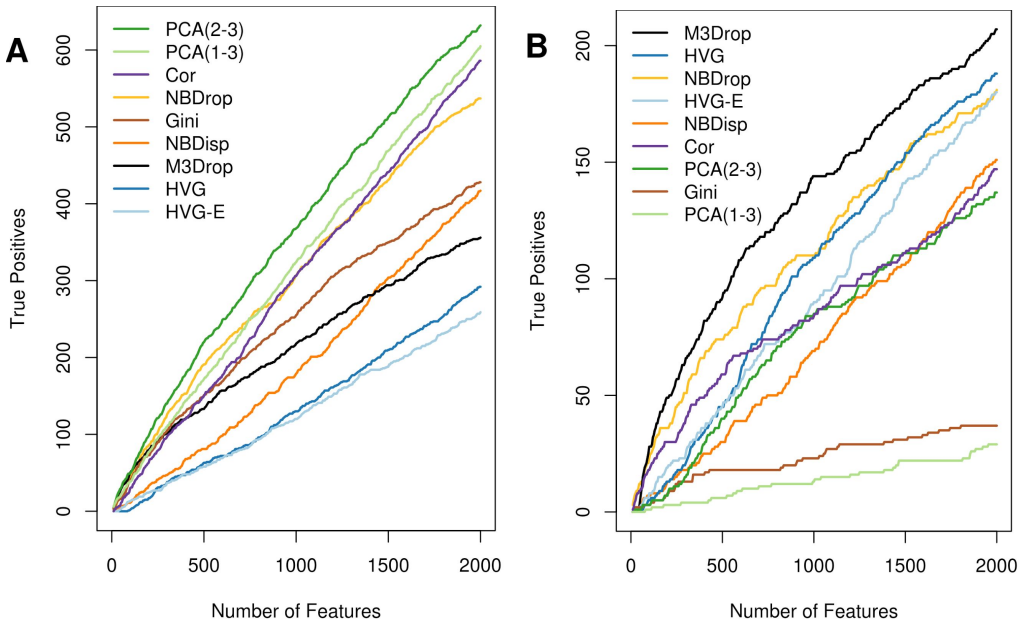

Figure S5 Performance of feature selection methods on UMI-tagged (A) and full-transcript (B) scRNASeq data when considering only the top features.

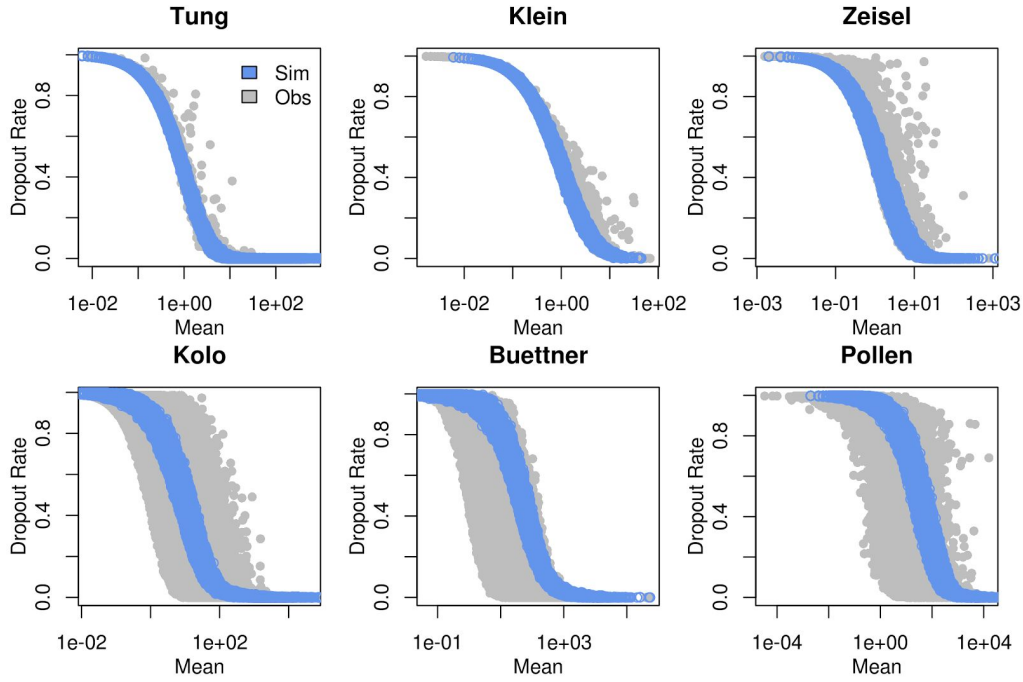

Figure S6 Simulations recapitulate observed relationship between gene expression and dropout rate. Only genes with log fold changes smaller than one (ground truth negatives) are plotted for the simulated data (blue).

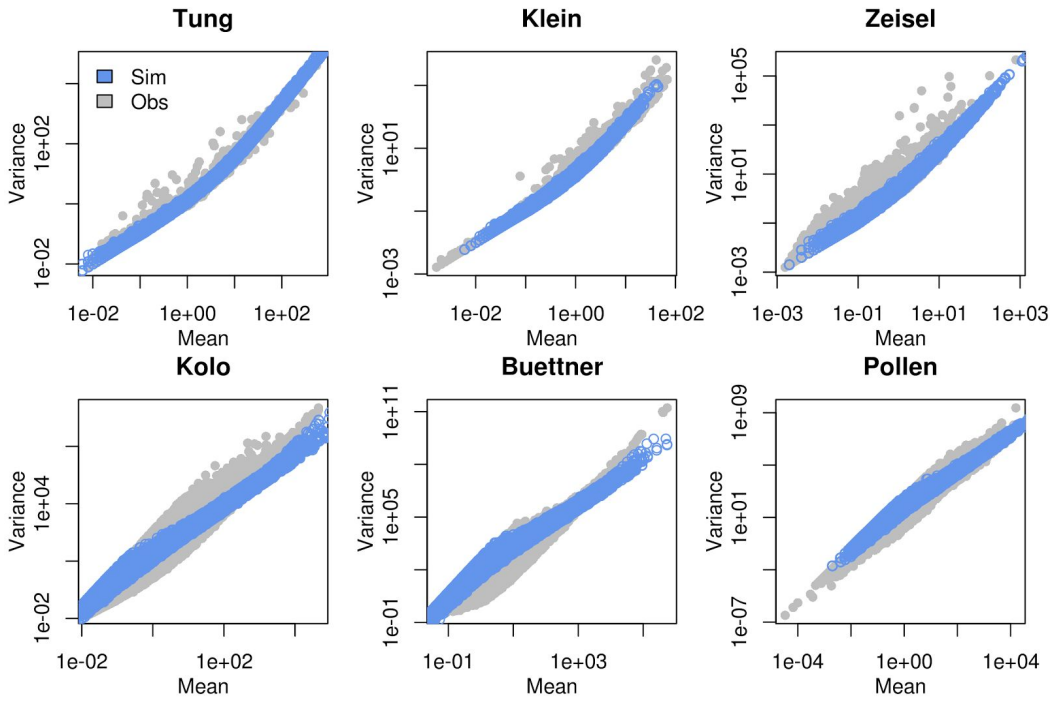

Figure S7 Simulations recapitulate observed relationship between gene expression and variance. Only genes with log fold changes smaller than one (ground truth negatives) are plotted for the simulated data (blue).

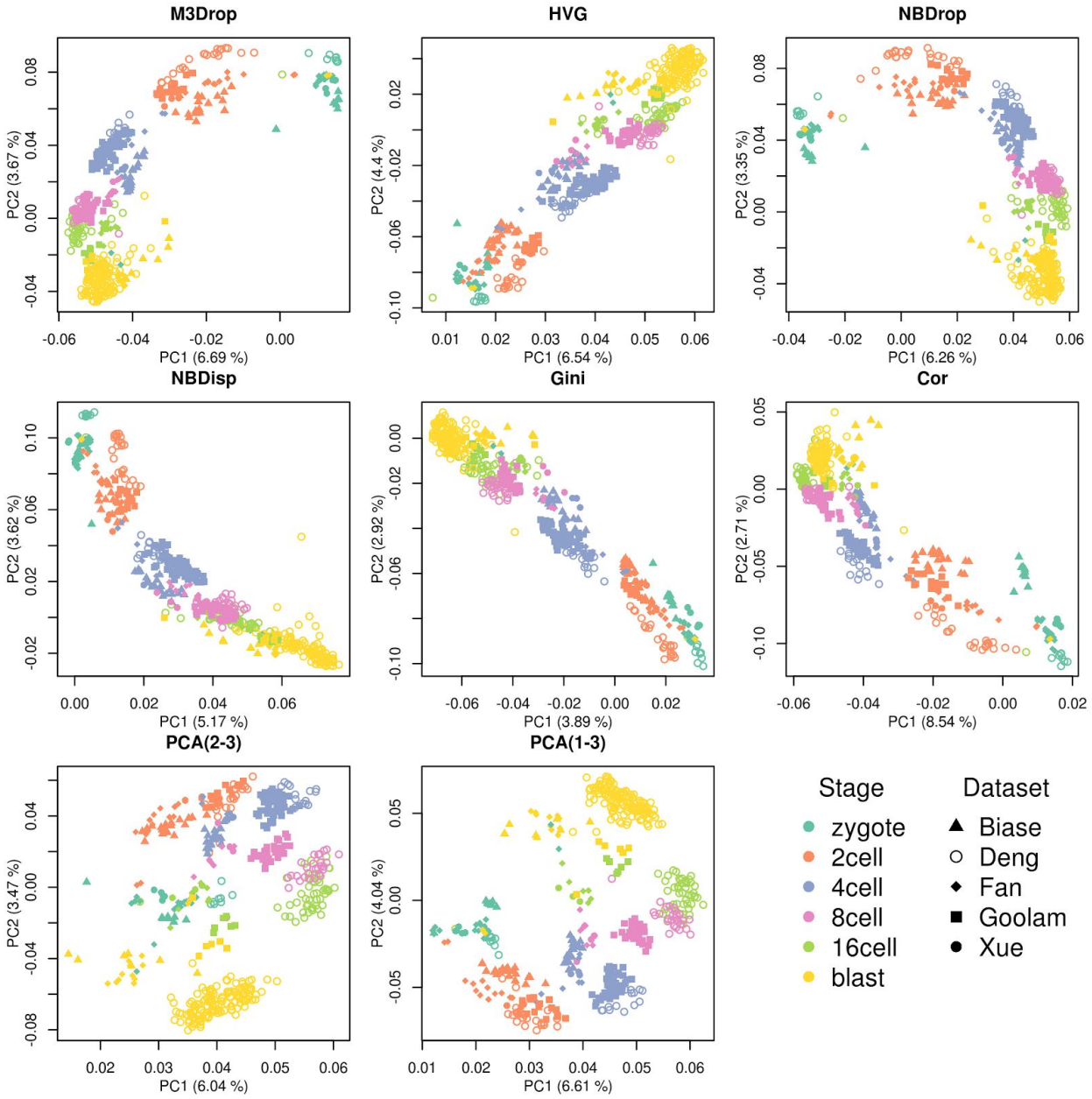

Figure S8 PCA plots after combining five mouse embryo datasets after different methods of feature selection. All except PCA-based feature selection preserve the developmental trajectory from zygote to blastocyst while reducing batch effects.

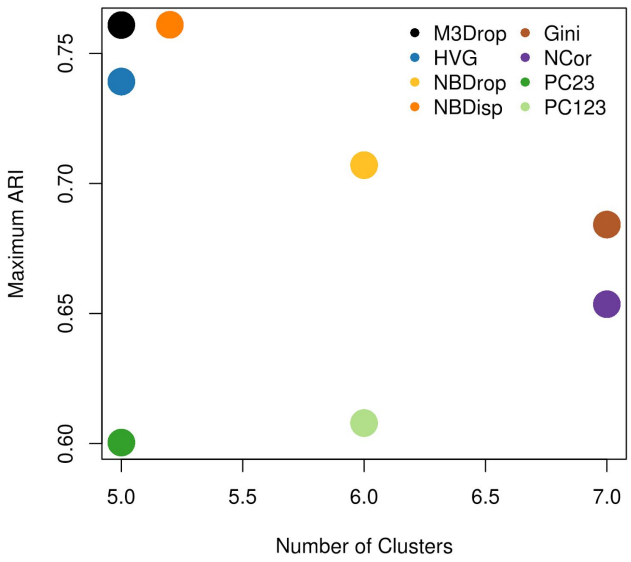

Figure S9 Batch effect removal quality across five different developmental datasets. After feature selection cells were clustered using Ward's hierarchical clustering and cut at all possible heights, the cut which had the highest adjusted rand index (ARI) with the true stages of the cells for each feature selection method is plotted. M3Drop and NBDisp (high variance based on the depth-adjusted negative binomial) where the most efficient at removing batch effects.

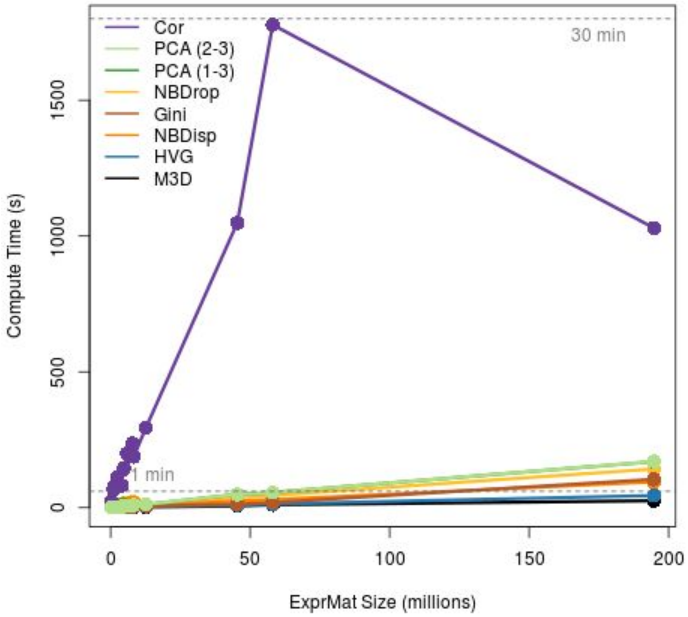

Figure S10 Only gene-gene correlations do not scale to large datasets. Run-times for each feature selection method were calculated for 13 scRNASeq datasets in **Table 1**, in addition to the Macosko (2015) [1] dataset of 45,000 mouse retinal cells (rightmost point). The x-axis reports the total number of entries in the gene expression matrix (#genes * #cells). Gene-gene correlations are more dependent on the number of genes rather than the number of cells; thus

run time decreases for the Macosko dataset since only 4,000 genes were detected across the 45,000 cells, whereas all other datasets detected > 10,000 genes.

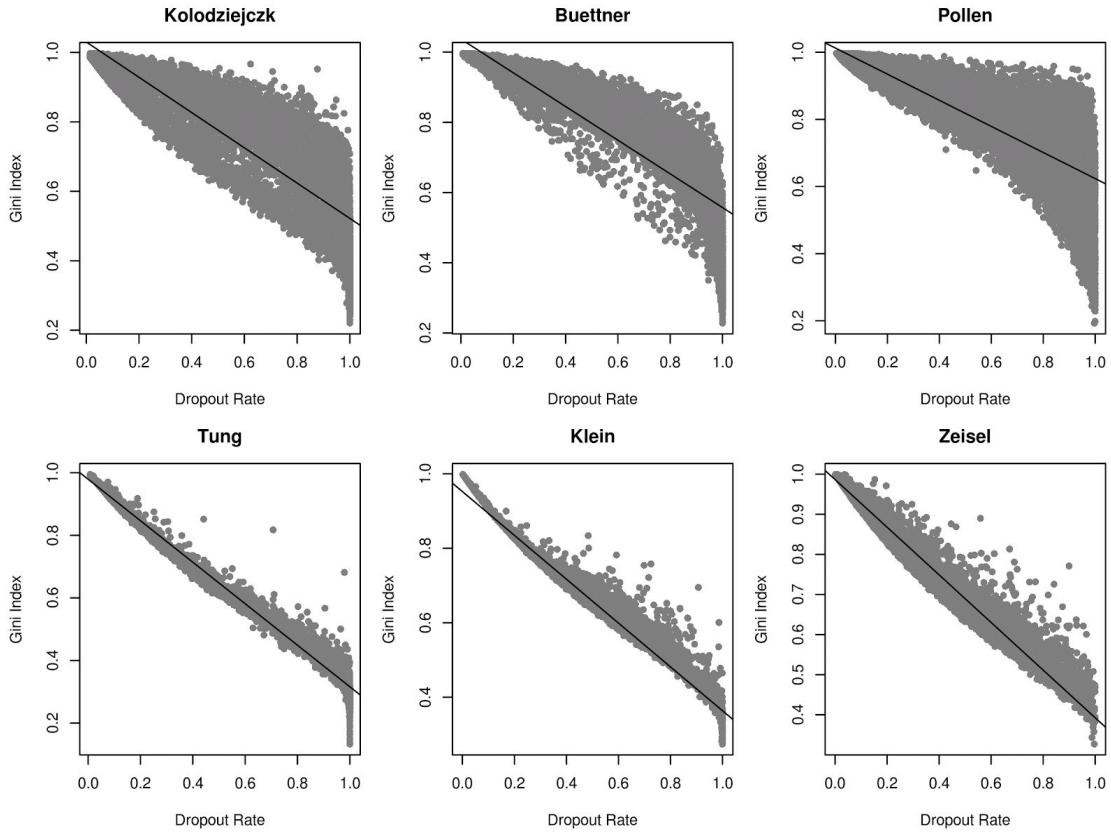

Figure S11 Due to the relationship between dropout rate (x-axis) and Gini index (y-axis), genes were rank by the residual from a linear regression (black line) between them.

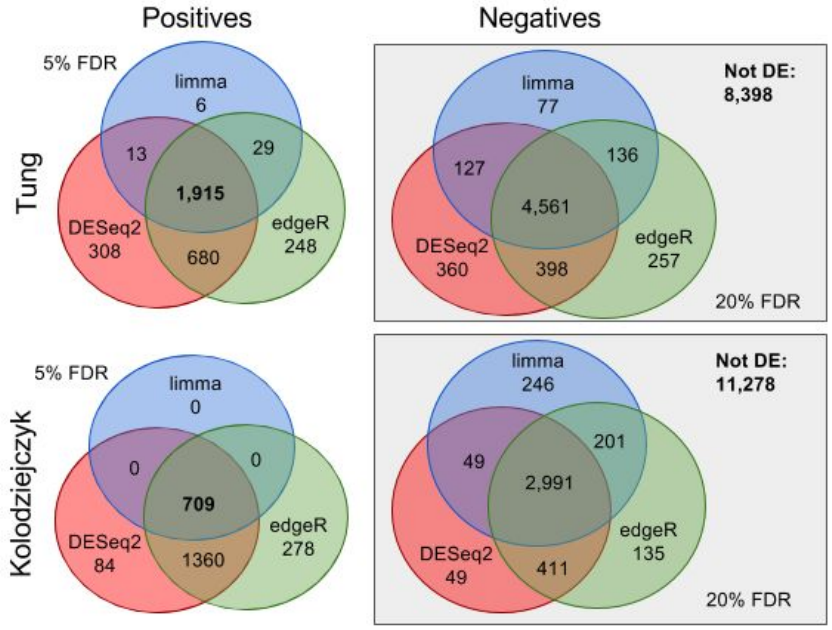

Figure S12 Ground truth DE genes were defined using the intersection (positives) and the complement of the union (negatives) of three standard differential expression method on the respective bulk RNASeq data of the Tung and Kolodziejzck datasets.

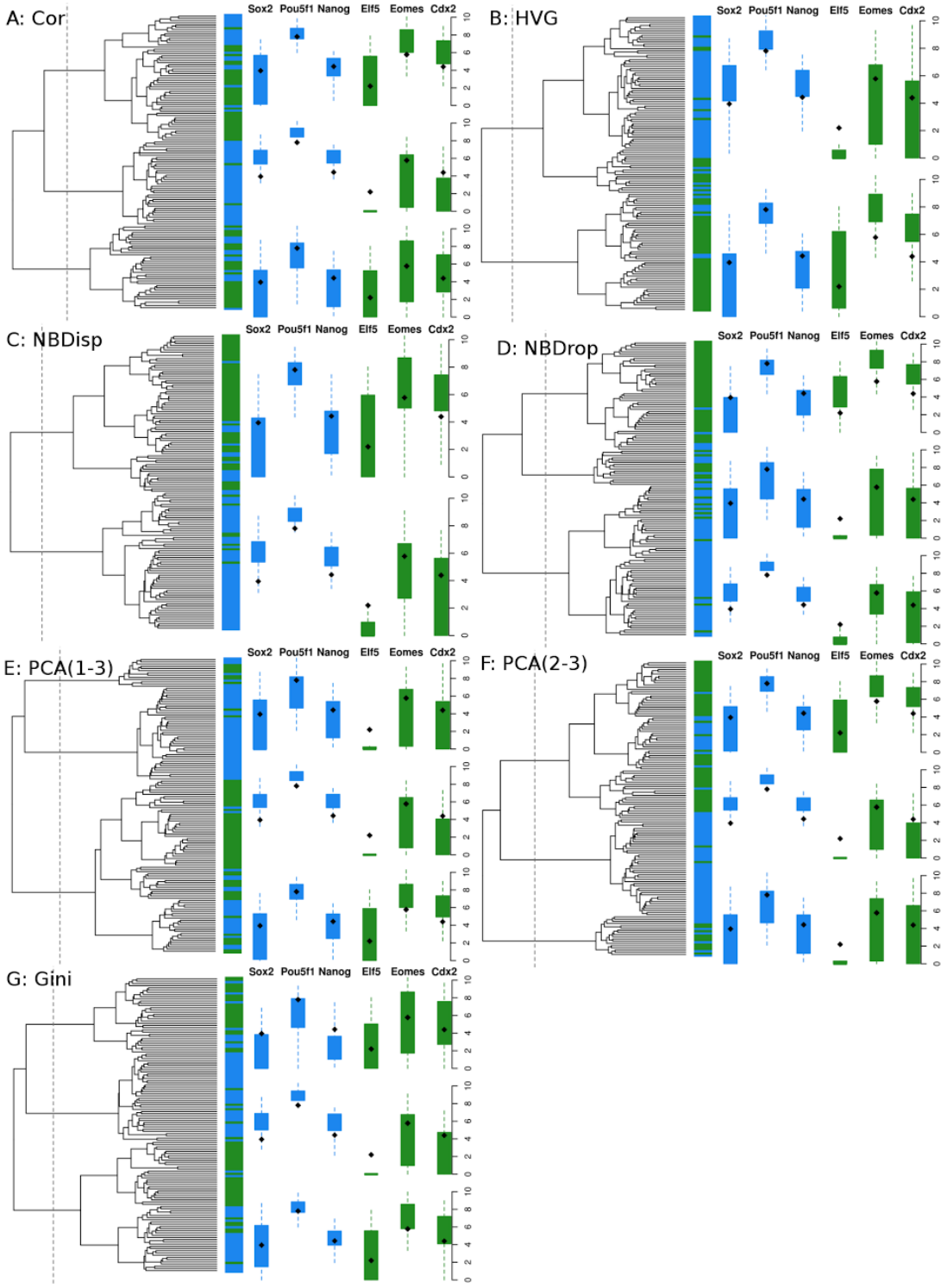

Figure S13 Identifying ICM and TE with other feature selection methods.

Bibliography

1. 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.