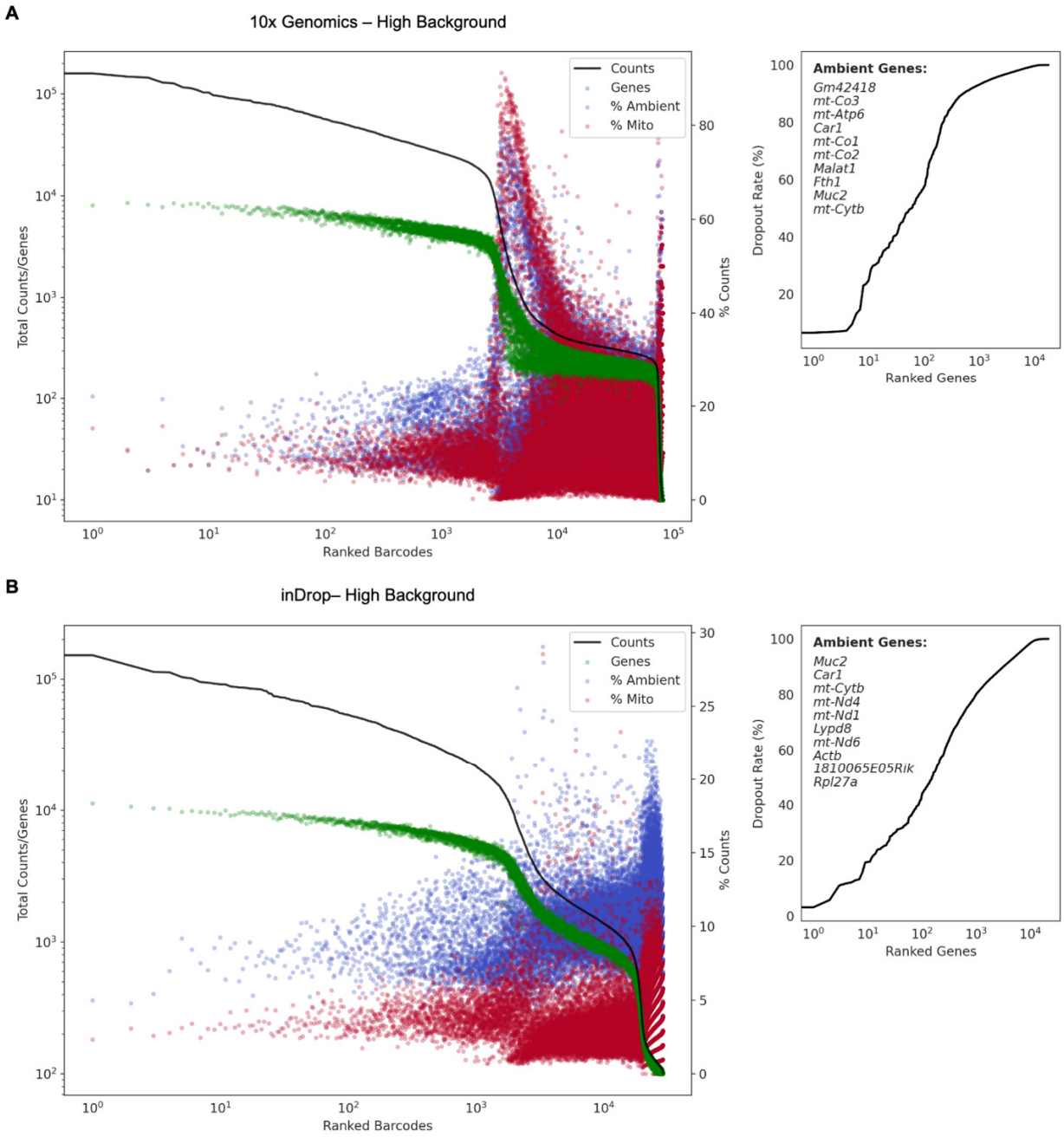
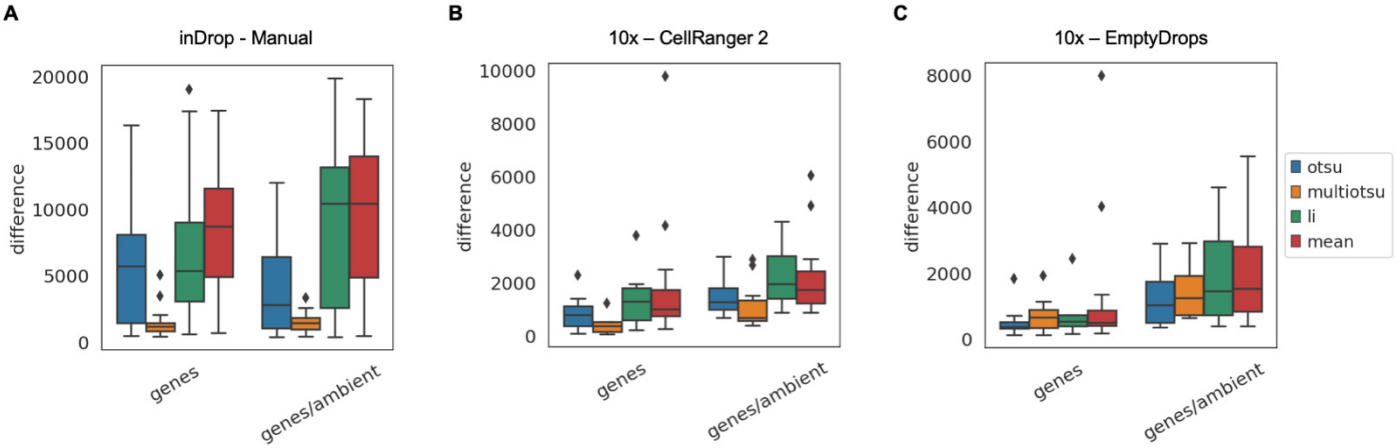
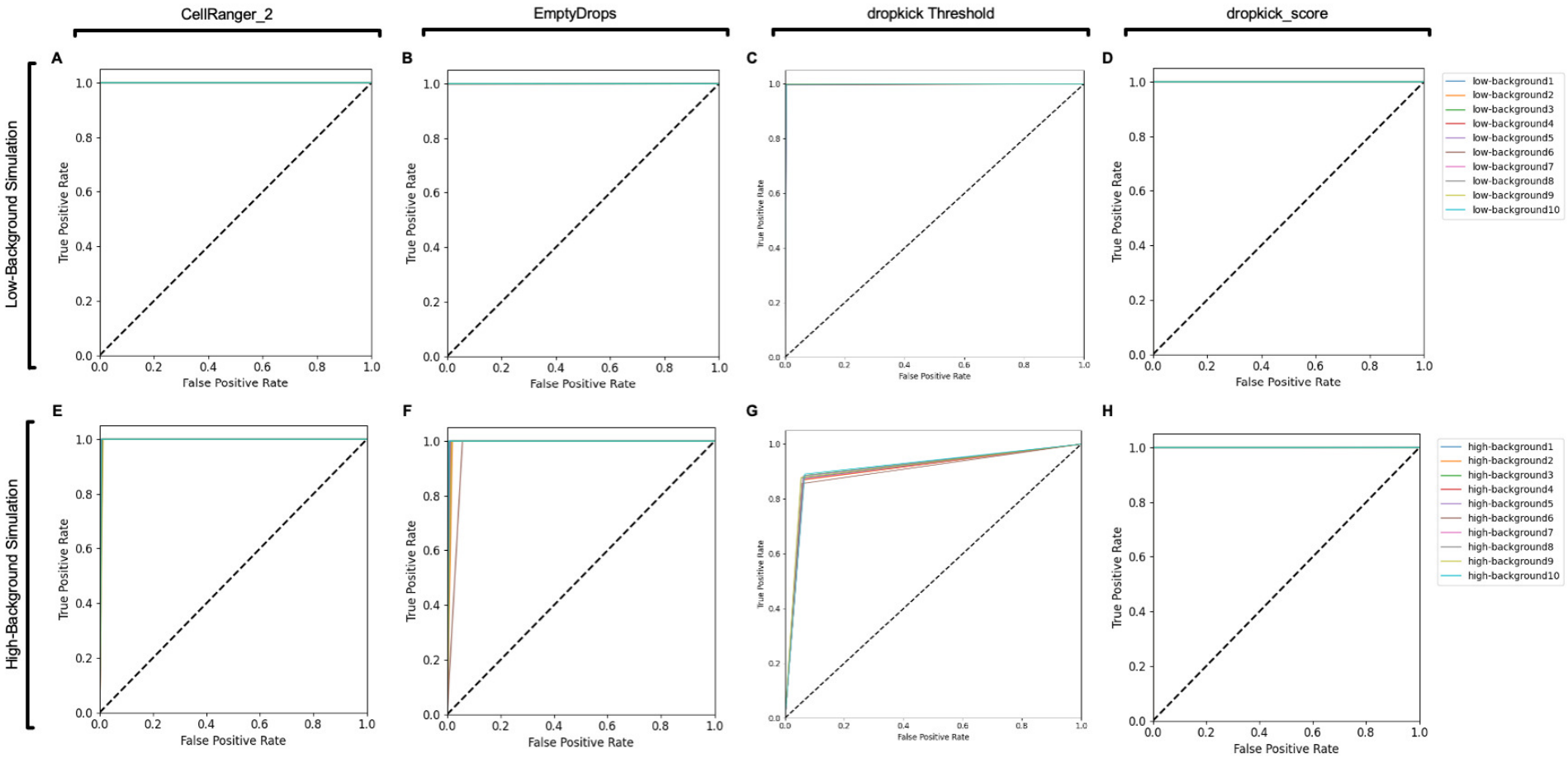
SUPPLEMENTARY FIGURES, TABLES AND LEGENDS



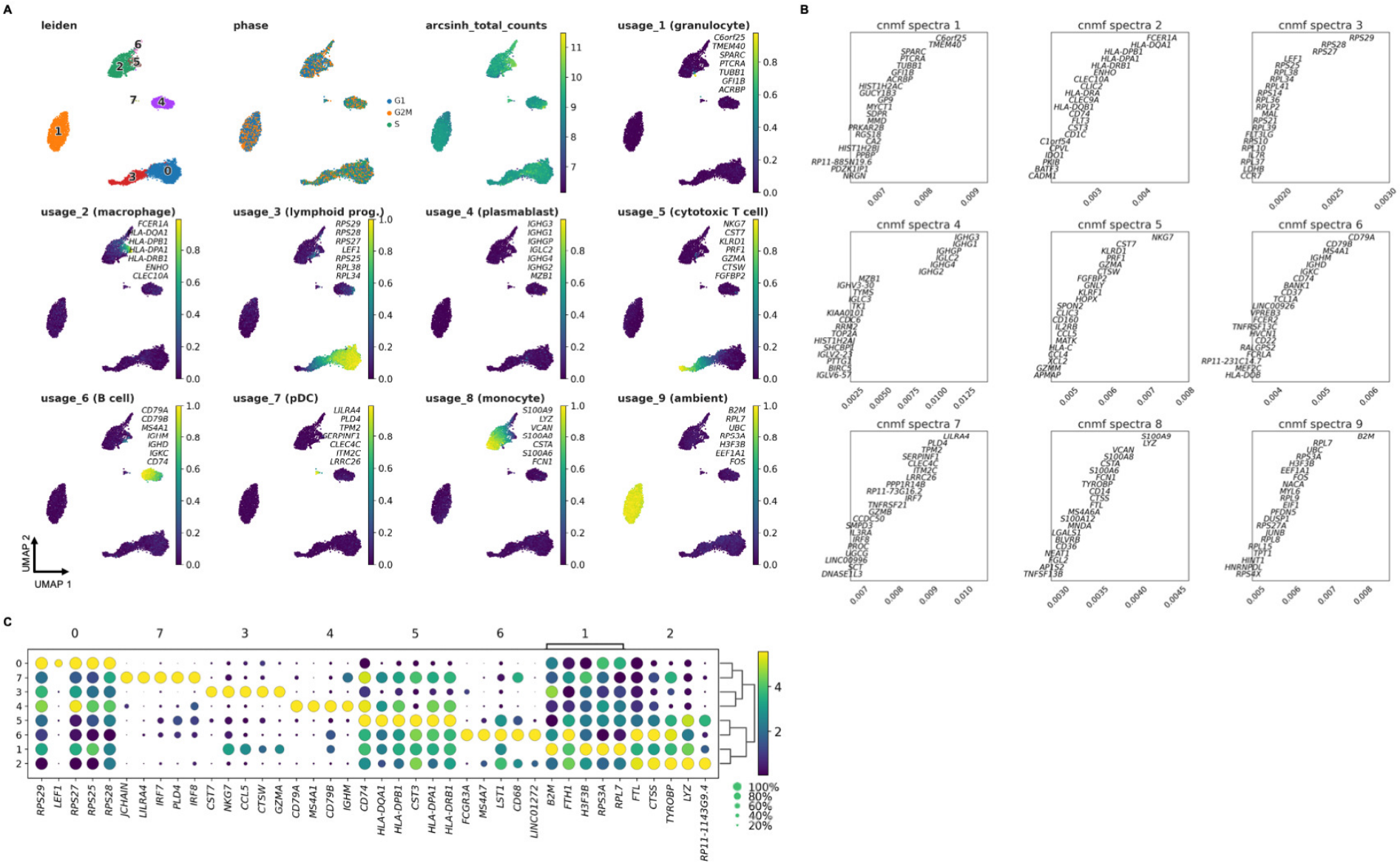
Supplementary Figure 1. dropkick QC reports for a mouse colonic epithelium sample analyzed by both 10x Genomics (A) and inDrop (B) scRNA-seq. In contrast to Figure 1, this is considered a high-background sample due to the height (increased total counts) of the second plateau (empty droplets) and presence of epithelial marker genes (*Car1*, *Muc2*) in the ambient profile.



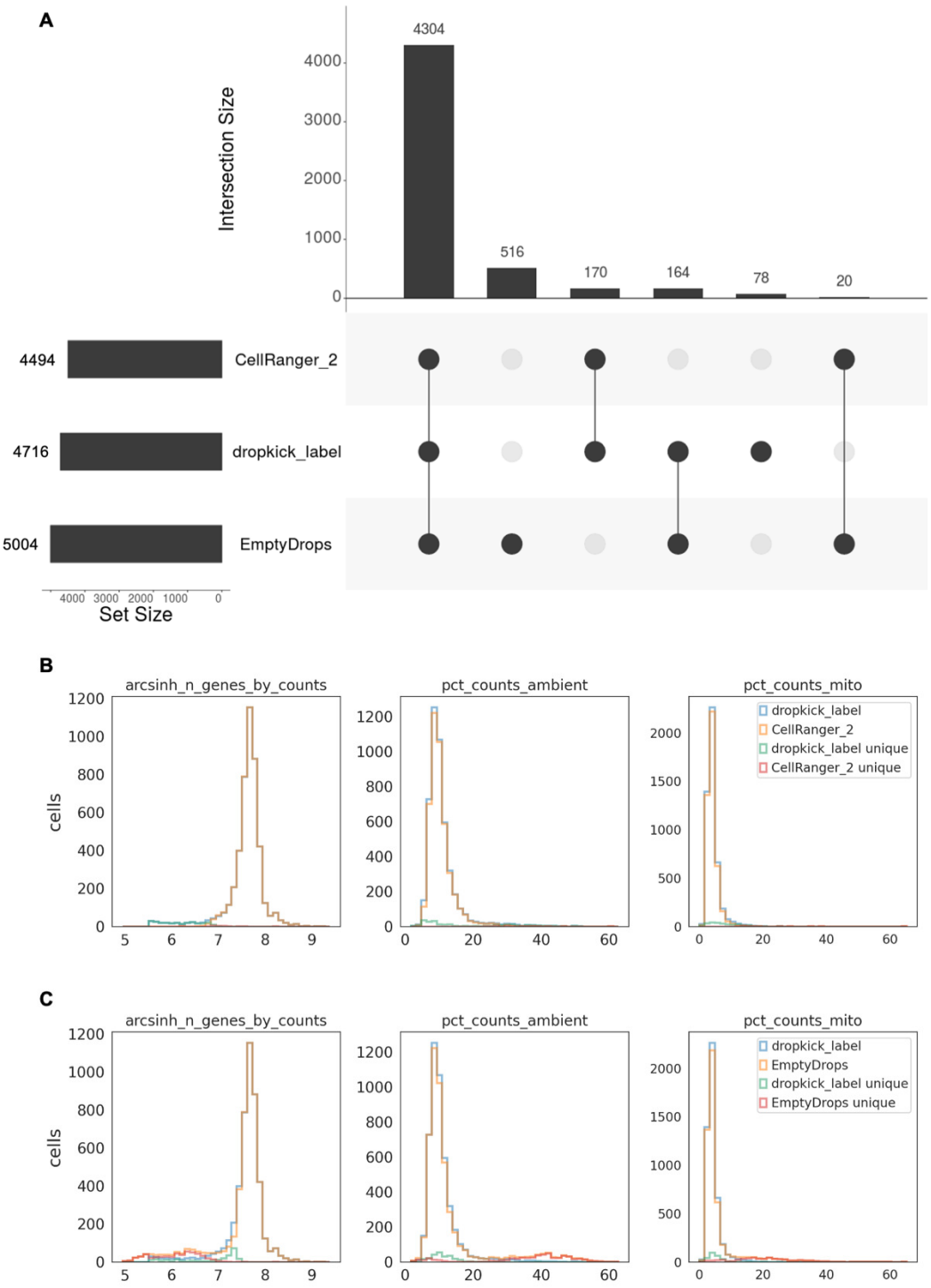
Supplementary Figure 2. Optimal heuristics and thresholding for determination of dropkick training set. A) Barcode set differences between initial dropkick thresholding and manual filtering of 33 inDrop scRNA-seq datasets. Four automated thresholding techniques were used to label cells based on the distribution of arcsinh-transformed genes detected alone (genes), or the combination of genes and percent ambient counts as calculated by the dropkick QC module (genes/ambient; see Methods: Quality control and ambient RNA quantification with dropkick QC module). B) Same as in A for 13 10x Genomics scRNA-seq datasets, with set differences compared to CellRanger_2. C) Same as in B, with set differences compared to EmptyDrops.



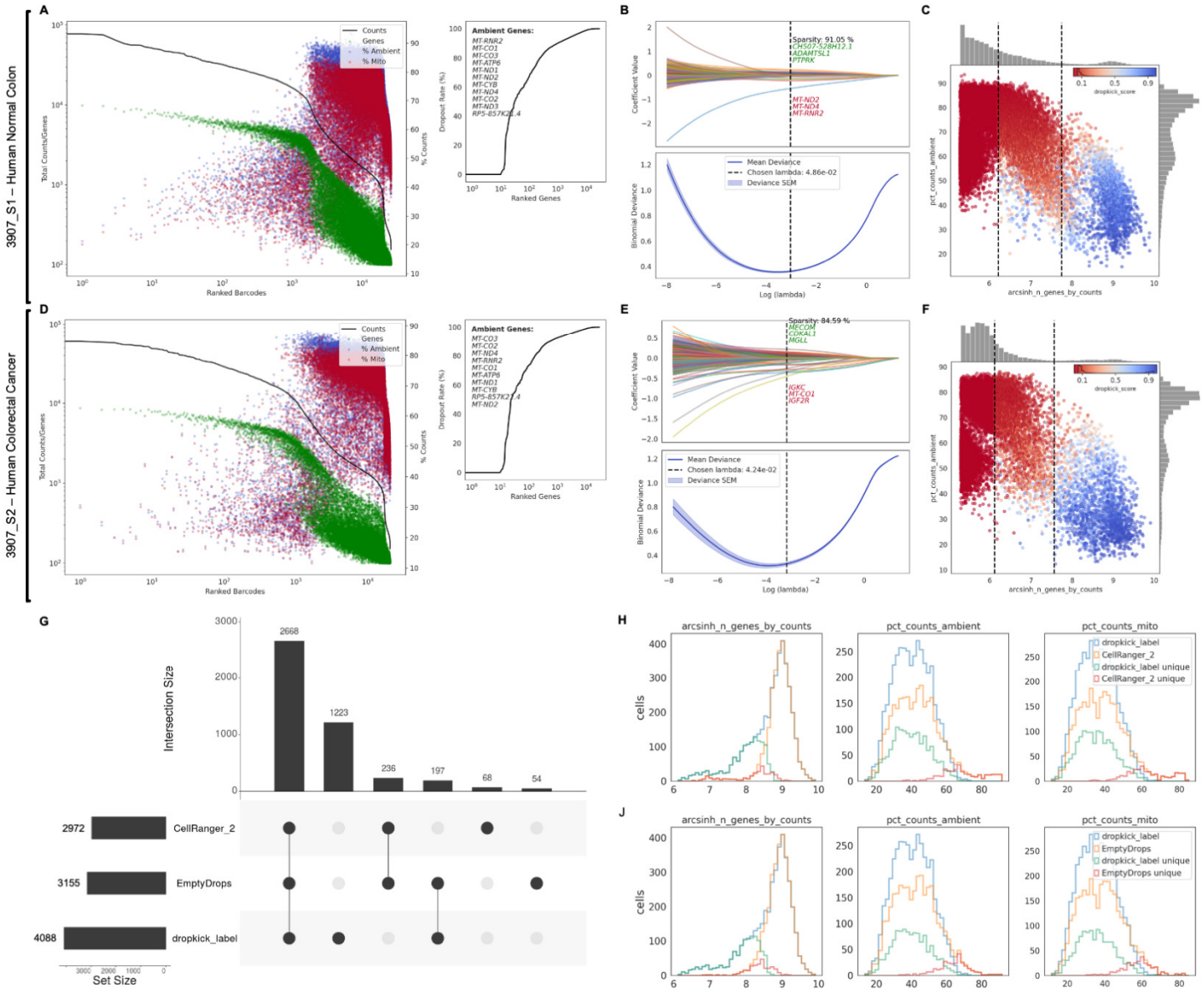
Supplementary Figure 3. Evaluating dropkick filtering performance with synthetic data. A) Receiver operating characteristic (ROC) curves for CellRanger_2 vs. ground truth in ten low-background simulations. B) ROC curves for EmptyDrops vs. ground truth in ten low-background simulations. C) ROC curves for dropkick training labels (threshold) vs. ground truth in ten low-background simulations. D) ROC curves for final dropkick score vs. ground truth in ten low-background simulations. E-H) Same as in A-D, for ten high-background simulations.



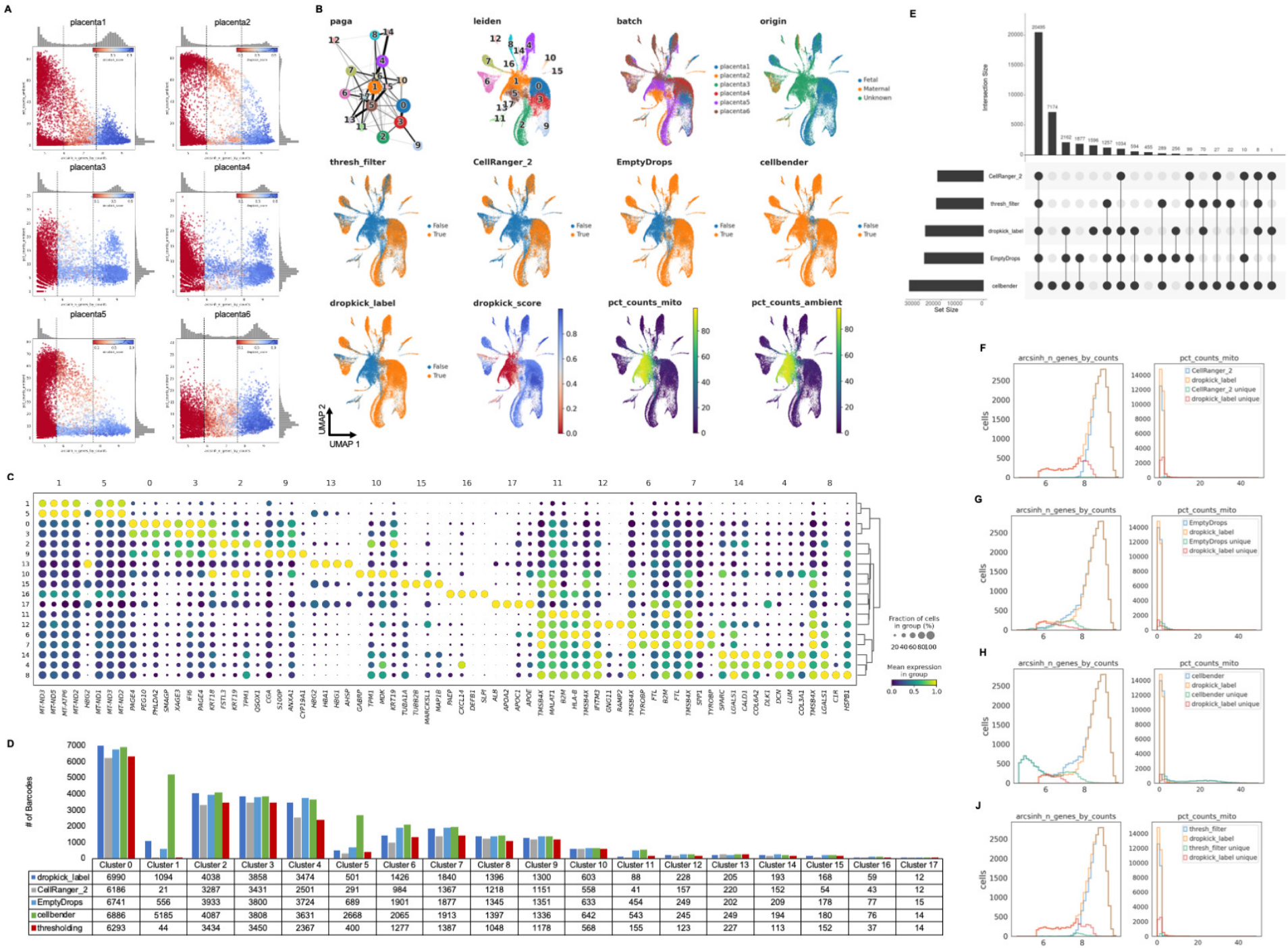
Supplementary Figure 4. Benchmarking dropkick performance on simulated high-background data. A) UMAP embedding of all barcodes kept by dropkick, CellRanger 2, and EmptyDrops. NMF results were used to generate leiden clusters; usage scores shown with a description of each cell type they represent and top 7 gene loadings for each. B) Top 20 gene loadings for each NMF metagene. C) Top 5 differentially expressed genes in the 8 leiden clusters.



Supplementary Figure 5. Barcode set differences for 4k pan-T cell dataset. A) UpSet plot showing global set differences between dropkick_label (dropkick score ≥ 0.5), CellRanger_2 and EmptyDrops. B) Histograms showing global distribution of heuristics (arcsinh-transformed genes, left, percent ambient counts, middle, and percent mitochondrial counts, right) in barcodes kept by dropkick_label and CellRanger_2. Distribution of barcodes unique to each label set also overlaid to show difference. C) Same as in B, for dropkick_label compared to EmptyDrops.

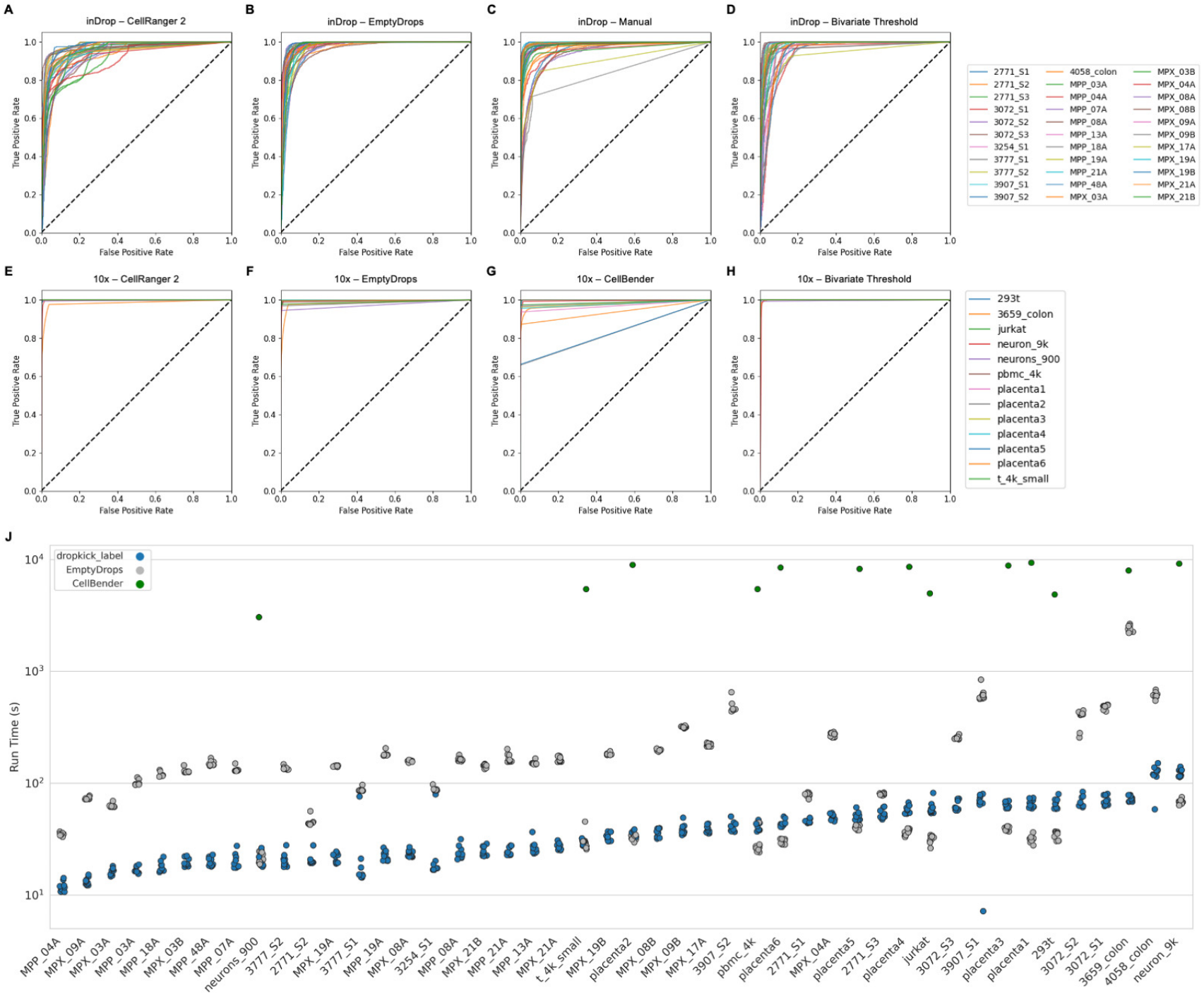


Supplementary Figure 6. dropkick plots and barcode set differences for human colorectal carcinoma (CRC) inDrop samples. A) dropkick QC report for human normal colonic mucosa, 3907_S1 and CRC, 3907_S2. B) dropkick coefficient plots, showing coefficient values (top) and binomial deviance (bottom) along the tested lambda regularization path. Dashed line indicates chosen lambda value of trained model. Top and bottom three genes by coefficient value and total model sparsity noted in top plot. C) dropkick score plot showing scatter of percent counts ambient versus arcsinh-transformed total genes detected per barcode. Dashed lines indicate location of automated dropkick thresholds used for model training. Points colored by final dropkick score. D-F) Same as in A-C, but for adjacent human normal colonic mucosa sample, 3907_S2. G) UpSet plot showing global set differences between dropkick_label (dropkick score ≥ 0.5), CellRanger_2 and EmptyDrops. H) Histograms showing global distribution of heuristics (arcsinh-transformed genes, left, percent ambient counts, middle, and percent mitochondrial counts, right) in barcodes kept by dropkick_label and CellRanger_2. Distribution of barcodes unique to each label set also overlaid to show difference. J) Same as in H, for dropkick_label compared to EmptyDrops.

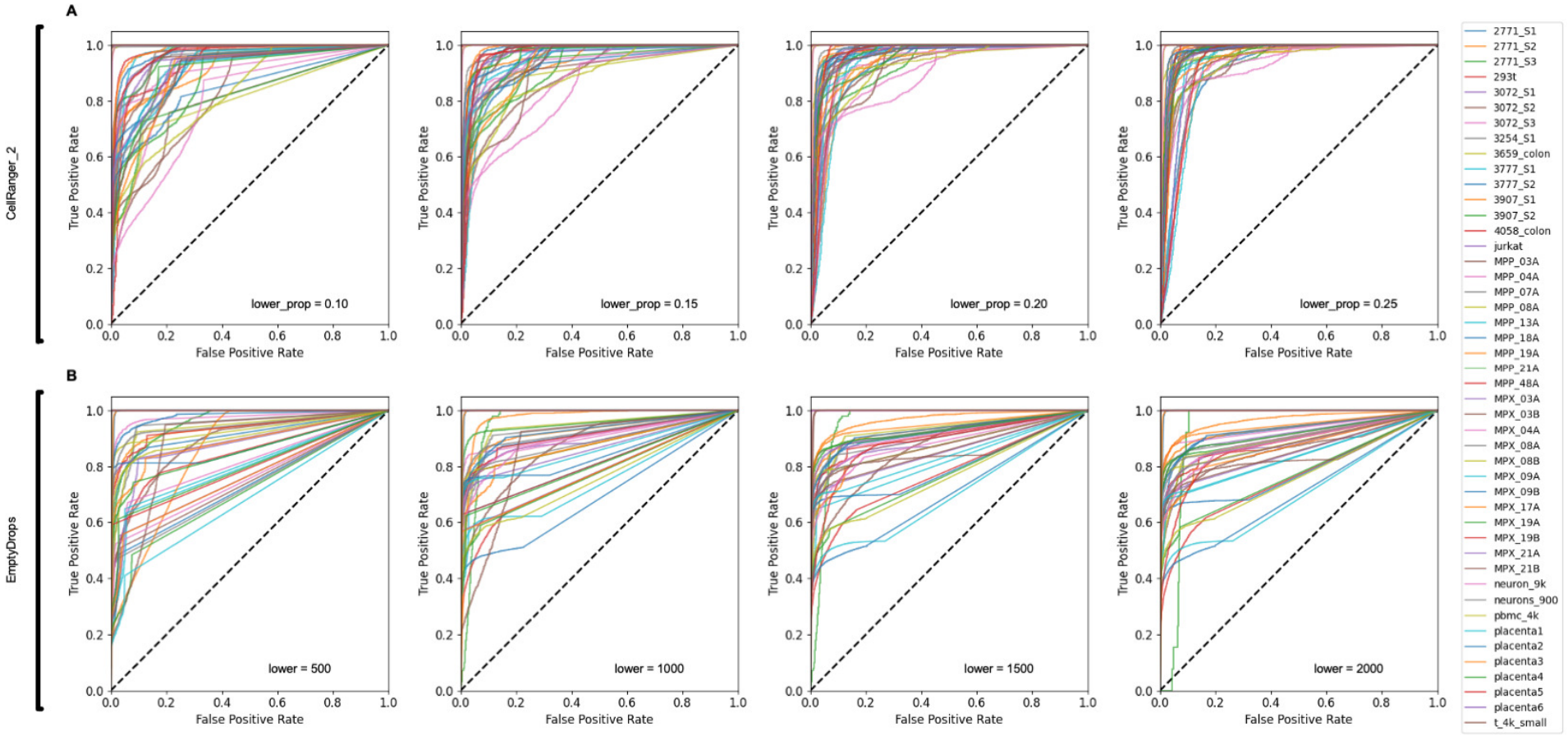


	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Cluster 14	Cluster 15	Cluster 16	Cluster 17
dropkick_label	6990	1094	4038	3858	3474	501	1426	1840	1396	1300	603	88	228	205	193	168	59	12
CellRanger_2	6186	21	3287	3431	2501	291	984	1367	1218	1151	558	41	157	220	152	54	43	12
EmptyDrops	6741	556	3933	3800	3724	689	1901	1877	1345	1351	633	454	249	202	209	178	77	15
cellbender	6886	5185	4087	3808	3631	2668	2065	1913	1397	1336	642	543	245	249	194	180	76	14
thresholding	6293	44	3434	3450	2367	400	1277	1387	1048	1178	568	155	123	227	113	152	37	14

Supplementary Figure 7. dropkick filters reproducibly across scRNA-seq batches. A) dropkick score plots for six placenta replicates. B) PAGA graph and UMAP embedding of all barcodes kept by dropkick_label (dropkick score ≥ 0.5), CellRanger_2, EmptyDrops, CellBender, and bivariate thresholding for the aggregate placenta dataset. Points colored by each of the five filtering labels as well as original batch, tissue of origin, Leiden clusters, dropkick_score (cell probability), percent counts ambient, and percent counts mitochondrial. C) Dot plot showing top five differentially expressed genes for each cluster. The size of each dot indicates the percentage of cells in the population with nonzero expression for the given gene, while the color indicates the average expression value in that population. D) Table and bar graph enumerating the total number of barcodes detected by each algorithm in all clusters. E) UpSet plot showing global set differences between dropkick_label, CellRanger_2, EmptyDrops, CellBender, and bivariate thresholding. F) Histograms showing global distribution of heuristics (arcsinh-transformed genes, left, and percent mitochondrial counts, right) in barcodes kept by dropkick_label and CellRanger_2. Distribution of barcodes unique to each label set also overlaid to show difference. G) Same as in F, for dropkick_label compared to EmptyDrops. H) Same as in F, for dropkick_label compared to CellBender. J) Same as in F, for dropkick_label compared to bivariate thresholding.


A
inDrop – CellRanger 2
B
inDrop – EmptyDrops
C
inDrop – Manual
D
inDrop – Bivariate Threshold
True Positive Rate
False Positive Rate
2771_S1
2771_S2
2771_S3
3072_S1
3072_S2
3072_S3
3254_S1
3777_S1
3777_S2
3907_S1
3907_S2
4058_colon
MPP_03A
MPP_04A
MPP_07A
MPP_08A
MPP_13A
MPP_18A
MPP_19A
MPP_21A
MPP_48A
MPX_03A
MPX_03B
MPX_04A
MPX_08A
MPX_08B
MPX_09A
MPX_09B
MPX_17A
MPX_19A
MPX_19B
MPX_21A
MPX_21B
E
10x – CellRanger 2
F
10x – EmptyDrops
G
10x – CellBender
H
10x – Bivariate Threshold
293t
3659_colon
jurkat
neuron_9k
neurons_900
pbmc_4k
placenta1
placenta2
placenta3
placenta4
placenta5
placenta6
t_4k_small
J
dropkick_label
EmptyDrops
CellBender
Run Time (s)
MPP_04A
MPX_09A
MPX_03A
MPP_03A
MPP_18A
MPX_03B
MPP_48A
MPP_07A
neurons_900
3777_S2
2771_S2
MPX_19A
3777_S1
MPP_19A
MPX_08A
3254_S1
MPP_08A
MPX_21B
MPP_21A
MPP_13A
MPX_21A
t_4k_small
MPX_19B
placenta2
MPX_08B
MPX_09B
MPX_17A
3907_S2
pbmc_4k
placenta6
2771_S1
MPX_04A
placenta5
2771_S3
placenta4
jurkat
3072_S3
3907_S1
placenta3
placenta1
293t
3072_S2
3072_S1
3659_colon
4058_colon
neuron_9k

Supplementary Figure 8. Comparing dropkick probability scores to five alternative cell labels using receiver operating characteristic (ROC) curves. AUC = area under the ROC curve. A) ROC curves for 33 inDrop scRNA-seq datasets, using CellRanger_2 as reference. B) Same as in A, with EmptyDrops as reference. C) Same as in A, with manually curated cell labels as reference (see methods: CellRanger 2, EmptyDrops, CellBender, and manual filtering of real-world scRNA-seq datasets). D) Same as in A, with bivariate thresholding as a reference (see methods: CellRanger 2, EmptyDrops, CellBender, and manual filtering of real-world scRNA-seq datasets). E) ROC curves for 13 10x Genomics scRNA-seq datasets, using CellRanger_2 as reference. F) Same as in E, with EmptyDrops labels as reference. G) Same as in E, with CellBender remove-background labels as a reference. H) Same as in E, with bivariate thresholding as a reference (see methods: CellRanger 2, EmptyDrops, CellBender, and manual filtering of real-world scRNA-seq datasets). J) Total run time, in seconds, for EmptyDrops, CellBender remove-background, and dropkick. EmptyDrops and dropkick were run ten times on all datasets; CellBender was run once on 10x Genomics samples. Points represent single replicates.



Supplementary Figure 9. Comparing dropkick probability scores to titrations of CellRanger_2 (A) and EmptyDrops (B) parameters using receiver operating characteristic (ROC) curves. AUC = area under the ROC curve. Results shown for 46 scRNA-seq datasets from 10x Genomics (n = 13) and inDrop (n = 33) encapsulation platforms. The “lower_prop” parameter in CellRanger version 2 (A) is used to exclude a bottom fraction of total barcodes prior to calculating the knee point of the log-rank total counts curve. The “lower” parameter in EmptyDrops is used to determine the total UMI cutoff below which all droplets are considered empty for model building.

Supplementary Table 1. Global comparison statistics between automated thresholding (dropkick training labels) and trained dropkick model vs. ground-truth cell labels for low and high-background simulations.

Simulation	Rep.	Threshold total cells	dropkick label total cells	True label total cells	Threshold sensitivity	Threshold Specificity	Threshold AUC	dropkick Sensitivity	dropkick Specificity	dropkick AUC
Low Background	1	3028	3001	3000	0.9983	0.9963	0.9973	1.0000	0.9999	1.0000
	2	3022	3003	3000	0.9990	0.9972	0.9981	1.0000	0.9997	1.0000
	3	3017	3003	3000	0.9990	0.9978	0.9984	1.0000	0.9997	1.0000
	4	3022	3002	3000	0.9967	0.9965	0.9966	1.0000	0.9998	1.0000
	5	3026	3001	3000	0.9990	0.9968	0.9979	1.0000	0.9999	1.0000
	6	3023	3005	3000	0.9987	0.9970	0.9978	1.0000	0.9994	1.0000
	7	3028	3003	3000	0.9983	0.9963	0.9973	1.0000	0.9997	1.0000
	8	3012	3000	3000	0.9983	0.9981	0.9982	1.0000	1.0000	1.0000
	9	3020	3000	3000	0.9990	0.9975	0.9982	1.0000	1.0000	1.0000
	10	3030	3002	3000	0.9993	0.9965	0.9979	1.0000	0.9998	1.0000
High Background	1	3269	3005	3000	0.8910	0.9379	0.9124	1.0000	0.9994	1.0000
	2	3178	2998	3000	0.8693	0.9404	0.9030	0.9990	0.9999	0.9995
	3	3246	3002	3000	0.8860	0.9387	0.9103	1.0000	0.9998	1.0000
	4	3167	3001	3000	0.8710	0.9420	0.9047	0.9997	0.9998	0.9998
	5	3206	3002	3000	0.8833	0.9418	0.9108	1.0000	0.9998	1.0000
	6	3095	2999	3000	0.8563	0.9448	0.8989	0.9997	1.0000	0.9998
	7	3200	2999	3000	0.8737	0.9396	0.9047	0.9993	0.9999	0.9997
	8	3127	2998	3000	0.8760	0.9475	0.9103	0.9993	1.0000	0.9997
	9	3118	2998	3000	0.8780	0.9490	0.9121	0.9993	1.0000	0.9997
	10	3199	2998	3000	0.8773	0.9407	0.9072	0.9990	0.9999	0.9995

Supplementary Table 2. Global comparison statistics between CellRanger_2 and EmptyDrops versus ground-truth cell labels for low and high-background simulations.

Simulation	Rep.	CellRanger 2 total cells	EmptyDrops total cells	True label total cells	CellRanger 2 sensitivity	CellRanger 2 Specificity	CellRanger 2 AUC	EmptyDrops Sensitivity	EmptyDrops Specificity	EmptyDrops AUC
Low Background	1	3000	3001	3000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	2	3000	3003	3000	1.0000	1.0000	1.0000	0.9997	1.0000	0.9998
	3	3000	3003	3000	1.0000	1.0000	1.0000	0.9997	1.0000	0.9998
	4	3000	3002	3000	1.0000	1.0000	1.0000	0.9997	1.0000	0.9998
	5	3000	3001	3000	1.0000	1.0000	1.0000	0.9993	1.0000	0.9997
	6	3000	3005	3000	1.0000	1.0000	1.0000	0.9997	1.0000	0.9998
	7	3000	3003	3000	1.0000	1.0000	1.0000	0.9997	1.0000	0.9998
	8	3000	3000	3000	1.0000	1.0000	1.0000	0.9997	1.0000	0.9998
	9	3000	3000	3000	1.0000	1.0000	1.0000	0.9997	1.0000	0.9998
	10	3000	3002	3000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
High Background	1	3070	3005	3000	1.0000	0.9923	0.9961	1.0000	0.9881	0.9940
	2	3061	2998	3000	1.0000	0.9933	0.9966	1.0000	0.9938	0.9969
	3	3091	3002	3000	1.0000	0.9900	0.9949	1.0000	0.9807	0.9902
	4	3079	3001	3000	1.0000	0.9913	0.9956	1.0000	0.9921	0.9960
	5	3077	3002	3000	1.0000	0.9915	0.9957	1.0000	0.9848	0.9923
	6	3077	2999	3000	1.0000	0.9915	0.9957	1.0000	0.9914	0.9957
	7	3055	2999	3000	1.0000	0.9939	0.9969	1.0000	0.9461	0.9715
	8	3103	2998	3000	1.0000	0.9887	0.9943	1.0000	0.9851	0.9924
	9	3093	2998	3000	1.0000	0.9898	0.9948	1.0000	0.9925	0.9962
	10	3111	2998	3000	1.0000	0.9878	0.9938	1.0000	0.9838	0.9918

Supplementary Table 3. Global comparison statistics between dropkick, CellRanger_2, and EmptyDrops for 13 10x Genomics scRNA-seq datasets.

	dropkick label total cells	CellRanger 2 total cells	CellRanger 2 sensitivity	CellRanger 2 specificity	CellRanger 2 AUC	EmptyDrops total cells	EmptyDrops sensitivity	EmptyDrops specificity	EmptyDrops AUC
293t	3224	2897	0.9717	0.9986	0.9994	3028	0.9508	0.9988	0.9991
jurkat	4020	3259	0.9963	0.9974	0.9989	3452	0.9655	0.9977	0.9985
neuron_9k	11050	9094	0.8745	0.9958	0.9964	11027	0.8422	0.9976	0.9950
neurons_900	1297	792	0.9369	0.9992	0.9963	1940	0.6649	1.0000	0.9721
pbmc_4k	4830	4335	0.9993	0.9993	0.9999	4231	0.9804	0.9991	0.9935
t_4k_small	4716	4494	0.9955	0.9993	0.9998	5004	0.8929	0.9993	0.9884
placenta1	6285	5636	0.9963	0.9991	0.9999	7149	0.8741	1.0000	0.9983
placenta2	2771	2101	0.9700	0.9990	0.9997	3530	0.7697	0.9999	0.9844
placenta3	5373	4019	0.9988	0.9982	0.9997	4600	0.9833	0.9988	0.9987
placenta4	5038	3906	0.9964	0.9984	0.9996	4532	0.9709	0.9991	0.9981
placenta5	4064	2723	0.9927	0.9982	0.9994	3885	0.8976	0.9992	0.9972
placenta6	3942	3289	0.9960	0.9991	0.9998	4238	0.9033	0.9998	0.9971
3659_colon	3376	5649	0.5936	1.0000	0.9877	5655	0.5929	1.0000	0.9878

Supplementary Table 4. Global comparison statistics between dropkick, CellRanger_2 and EmptyDrops for 33 inDrop scRNA-seq datasets.

	dropkick label total cells	CellRanger 2 total cells	CellRanger 2 sensitivity	CellRanger 2 specificity	CellRanger 2 AUC	EmptyDrops total cells	EmptyDrops sensitivity	EmptyDrops specificity	EmptyDrops AUC
2771_S1	2222	2401	0.7380	0.9264	0.9263	2519	0.7594	0.9472	0.9590
2771_S2	1248	925	0.8465	0.9031	0.9237	1219	0.8515	0.9506	0.9634
2771_S3	1382	1422	0.6392	0.9643	0.9154	983	0.8688	0.9616	0.9771
3072_S1	962	1536	0.5137	0.9899	0.9619	1049	0.6616	0.9848	0.9788
3072_S2	2370	2239	0.6807	0.9434	0.9196	1571	0.8250	0.9322	0.9531
3072_S3	3112	2873	0.8190	0.9102	0.9331	3210	0.8059	0.9334	0.9485
3254_S1	1285	1356	0.8201	0.9536	0.9501	1235	0.8907	0.9521	0.9783
3777_S1	1066	828	0.9541	0.9082	0.9669	908	0.9361	0.9247	0.9735
3777_S2	927	1205	0.6929	0.9802	0.9435	1699	0.5221	0.9902	0.9712
3907_S1	2260	2016	0.8661	0.9788	0.9782	1948	0.9004	0.9792	0.9892
3907_S2	1828	2057	0.7321	0.9828	0.9652	1207	0.9205	0.9641	0.9840
4058_colon	2186	2155	0.7940	0.9827	0.9788	1973	0.8246	0.9798	0.9790
MPP_03A	945	997	0.6861	0.9534	0.9224	914	0.8271	0.9663	0.9771
MPP_04A	702	766	0.8055	0.9365	0.9281	847	0.8040	0.9824	0.9842
MPP_07A	1384	1352	0.9098	0.9761	0.9837	1256	0.9594	0.9728	0.9895
MPP_08A	1473	1601	0.7077	0.9597	0.9290	1433	0.8130	0.9641	0.9692
MPP_13A	1815	1634	0.8782	0.9501	0.9563	1744	0.8893	0.9643	0.9744
MPP_18A	875	753	0.8539	0.9655	0.9558	661	0.9213	0.9612	0.9799
MPP_19A	1503	1477	0.8538	0.9732	0.9809	1416	0.8905	0.9734	0.9876
MPP_21A	2149	2096	0.8545	0.9522	0.9574	1518	0.9697	0.9193	0.9663
MPP_48A	811	287	0.9477	0.9389	0.9652	934	0.8084	0.9927	0.9931
MPX_03A	1228	1197	0.8797	0.9370	0.9471	1217	0.9260	0.9624	0.9872
MPX_03B	407	782	0.3964	0.9861	0.9343	493	0.5558	0.9818	0.9779
MPX_04A	744	975	0.6338	0.9900	0.9136	730	0.8274	0.9891	0.9878
MPX_08A	1277	1100	0.8236	0.9563	0.9599	906	0.9183	0.9492	0.9767
MPX_08B	947	962	0.8753	0.9889	0.9751	1050	0.8333	0.9923	0.9915
MPX_09A	549	507	0.7929	0.9655	0.9622	609	0.8144	0.9870	0.9857
MPX_09B	1462	1235	0.8899	0.9767	0.9785	1143	0.9283	0.9745	0.9850
MPX_17A	889	813	0.7282	0.9751	0.9638	1347	0.5457	0.9863	0.9709
MPX_19A	857	798	0.7406	0.9673	0.9630	865	0.7260	0.9715	0.9737
MPX_19B	2830	2753	0.8794	0.9493	0.9709	2769	0.9029	0.9586	0.9804
MPX_21A	1647	1629	0.8422	0.9660	0.9713	1622	0.8662	0.9700	0.9812
MPX_21B	1377	1287	0.8143	0.9541	0.9608	1020	0.9255	0.9426	0.9748

Supplementary Table 5. Global comparison statistics between dropkick and manual cell labelling for 13 10x Genomics scRNA-seq datasets. “Thresholding” refers to bivariate thresholding on total counts and mitochondrial percentage (see Methods: CellRanger 2, EmptyDrops, CellBender, and manual filtering of real-world scRNA-seq datasets).

	dropkick label total cells	CellBender total cells	CellBender sensitivity	CellBender specificity	CellBender AUC	Thresholding total cells	Thresholding sensitivity	Thresholding specificity	Thresholding AUC
293t	3224	3007	0.9488	0.9987	0.9993	2454	0.9833	0.9972	0.9987
jurkat	4020	3310	0.9909	0.9975	0.9990	2820	0.9996	0.9959	0.9982
neuron_9k	11050	10937	0.8207	0.9972	0.9939	6278	0.8767	0.9925	0.9946
neurons_900	1297	1483	0.8213	0.9999	0.9837	815	0.9362	0.9993	0.9958
pbmc_4k	4830	4759	0.9950	0.9999	0.9998	3339	1.0000	0.9980	0.9994
t_4k_small	4716	5153	0.8921	0.9997	0.9867	3487	0.9997	0.9966	0.9992
placenta1	6285	7859	0.7969	1.0000	0.9681	5628	0.9963	0.9991	0.9999
placenta2	2771	5773	0.4745	1.0000	0.8289	2819	0.8915	0.9996	0.9997
placenta3	5373	5055	0.9318	0.9991	0.9816	4343	0.9965	0.9986	0.9998
placenta4	5038	5076	0.9088	0.9994	0.9778	4242	0.9873	0.9988	0.9997
placenta5	4064	6336	0.5418	0.9991	0.8308	2977	0.9886	0.9985	0.9995
placenta6	3942	5020	0.7695	0.9999	0.9358	2258	0.9969	0.9977	0.9991
3659_colon	3376	4261	0.7728	0.9989	0.9750	2873	0.9715	0.9927	0.9979

Supplementary Table 6. Global comparison statistics between dropkick and manual cell labelling for 33 inDrop scRNA-seq datasets. “Manual label” refers to the manual cluster gating, while “Thresholding” refers to bivariate thresholding on total counts and mitochondrial percentage (see Methods: CellRanger 2, EmptyDrops, CellBender, and manual filtering of real-world scRNA-seq datasets).

	dropkick label total cells	Manual label total cells	Manual label sensitivity	Manual label specificity	Manual label AUC	Thresholding total cells	Thresholding sensitivity	Thresholding specificity	Thresholding AUC
2771_S1	2222	2715	0.7193	0.9521	0.9515	2643	0.7465	0.9561	0.9598
2771_S2	1248	1536	0.7637	0.9802	0.9635	1305	0.8705	0.9724	0.9841
2771_S3	1382	1303	0.8365	0.9778	0.9854	1288	0.8168	0.9751	0.9834
3072_S1	962	2639	0.3297	0.9942	0.9360	2837	0.3123	0.9952	0.9424
3072_S2	2370	2725	0.6712	0.9618	0.9382	3500	0.6291	0.9871	0.9752
3072_S3	3112	2990	0.8084	0.9160	0.9375	8809	0.3494	0.9810	0.9415
3254_S1	1285	1442	0.8336	0.9766	0.9829	920	0.8848	0.8945	0.9544
3777_S1	1066	2253	0.4589	0.9761	0.8325	1239	0.7627	0.9505	0.9446
3777_S2	927	2326	0.3947	0.9974	0.8971	1729	0.5095	0.9887	0.9302
3907_S1	2260	2288	0.8488	0.9866	0.9909	1271	0.8812	0.9555	0.9847
3907_S2	1828	2003	0.8522	0.9935	0.9920	1284	0.9829	0.9713	0.9943
4058_colon	2186	2182	0.7988	0.9838	0.9796	1930	0.8187	0.9782	0.9787
MPP_03A	945	993	0.8449	0.9805	0.9793	203	1.0000	0.8920	0.9740
MPP_04A	702	854	0.7787	0.9692	0.9684	349	1.0000	0.8256	0.9502
MPP_07A	1384	1397	0.9170	0.9838	0.9916	856	0.9988	0.9278	0.9855
MPP_08A	1473	1548	0.8637	0.9836	0.9834	687	0.9840	0.9188	0.9778
MPP_13A	1815	1794	0.9498	0.9845	0.9942	969	0.9876	0.9020	0.9587
MPP_18A	875	834	0.9329	0.9851	0.9925	444	0.9955	0.9402	0.9878
MPP_19A	1503	1719	0.8214	0.9895	0.9896	1116	0.9839	0.9577	0.9880
MPP_21A	2149	2149	0.9614	0.9884	0.9966	682	1.0000	0.8535	0.9726
MPP_48A	811	785	0.8994	0.9867	0.9948	410	0.9902	0.9527	0.9787
MPX_03A	1228	1242	0.8712	0.9460	0.9643	689	0.9942	0.8514	0.9510
MPX_03B	407	1313	0.3024	0.9984	0.9583	365	0.7178	0.9806	0.9840
MPX_04A	744	818	0.8264	0.9947	0.9883	346	0.9855	0.9702	0.9937
MPX_08A	1277	1200	0.9225	0.9792	0.9917	597	0.9832	0.9259	0.9831
MPX_08B	947	872	0.9507	0.9876	0.9960	648	0.9954	0.9697	0.9974
MPX_09A	549	590	0.8678	0.9909	0.9896	306	0.9935	0.9463	0.9801
MPX_09B	1462	1455	0.9162	0.9915	0.9949	962	0.9844	0.9679	0.9890
MPX_17A	889	691	0.9450	0.9803	0.9950	663	0.8869	0.9751	0.9899
MPX_19A	857	880	0.8045	0.9812	0.9874	535	0.9832	0.9609	0.9908
MPX_19B	2830	2807	0.9184	0.9679	0.9853	1280	0.9898	0.8538	0.9469
MPX_21A	1647	1615	0.9467	0.9851	0.9923	889	0.9888	0.9176	0.9866
MPX_21B	1377	1369	0.9065	0.9803	0.9867	666	0.9895	0.9122	0.9834

Supplementary Table 7. Parameters used for filtering techniques on all 46 samples. “Chosen alpha” describes the optimal alpha chosen by the dropkick filtering model when cross validated using alphas from [0.1, 0.25, 0.5, 0.75, 0.9]. All other results discussed in this manuscript have alpha set to 0.1 by default.

		EmptyDrops	CellRanger	Bivariate Threshold		CellBender	dropkick
Sample	Exp. Cells	Lower	Lower	Min. UMI	Max. % Mito.	Total Droplets Included	Chosen alpha
2771_S1	2400	Inflection pt.	0.17	1000	40	NA	0.1
2771_S2	1000	Inflection pt.	0.15	1000	40	NA	0.1
2771_S3	1300	Inflection pt.	0.1	1000	40	NA	0.1
3072_S1	1500	Inflection pt.	0.19	1000	40	NA	0.1
3072_S2	2000	Inflection pt.	0.27	1000	40	NA	0.1
3072_S3	2000	Inflection pt.	0.09	1000	40	NA	0.1
3254_S1	1300	Inflection pt.	0.09	1000	40	NA	0.1
3777_S1	1000	Inflection pt.	0.08	1000	40	NA	0.5
3777_S2	1200	Inflection pt.	0.17	1000	40	NA	0.1
3907_S1	2000	Inflection pt.	0.1	3000	40	NA	0.9
3907_S2	2000	Inflection pt.	0.09	3000	40	NA	0.1
4058_colon	2000	Inflection pt.	0.2	10000	40	NA	0.1
MPP_03A	900	Inflection pt.	0.14	5000	40	NA	0.1
MPP_04A	700	Inflection pt.	0.16	5000	40	NA	0.1
MPP_07A	1300	Inflection pt.	0.21	5000	40	NA	0.1
MPP_08A	1500	Inflection pt.	0.17	5000	40	NA	0.1
MPP_13A	1600	Inflection pt.	0.18	5000	40	NA	0.1
MPP_18A	700	Inflection pt.	0.14	5000	40	NA	0.1
MPP_19A	1400	Inflection pt.	0.22	5000	40	NA	0.1
MPP_21A	2000	Inflection pt.	0.15	5000	40	NA	0.1
MPP_48A	500	Inflection pt.	0.1	5000	40	NA	0.1
MPX_03A	1100	Inflection pt.	0.15	5000	40	NA	0.1
MPX_03B	800	Inflection pt.	0.22	5000	40	NA	0.5
MPX_04A	900	Inflection pt.	0.19	5000	40	NA	0.1
MPX_08A	1100	Inflection pt.	0.24	5000	40	NA	0.1
MPX_08B	900	Inflection pt.	0.2	5000	40	NA	0.1
MPX_09A	500	Inflection pt.	0.2	5000	40	NA	0.1
MPX_09B	1200	Inflection pt.	0.23	5000	40	NA	0.1
MPX_17A	800	Inflection pt.	0.18	5000	40	NA	0.1
MPX_19A	800	Inflection pt.	0.16	5000	40	NA	0.1
MPX_19B	2700	Inflection pt.	0.17	5000	40	NA	0.1
MPX_21A	1600	Inflection pt.	0.17	5000	40	NA	0.1
MPX_21B	1200	Inflection pt.	0.01	5000	40	NA	0.1
293t	2800	Inflection pt.	0.1	10000	40	15000	0.1
jurkat	3200	Inflection pt.	0.1	10000	40	15000	0.9
neuron_9k	9000	Inflection pt.	0.1	5000	40	20000	0.1
neurons_900	900	Inflection pt.	0.1	5000	40	10000	0.1
pbmc_4k	4000	Inflection pt.	0.1	3000	40	15000	0.9
t_4k_small	4000	Inflection pt.	0.1	3000	40	15000	0.5
placenta1	5000	Inflection pt.	0.1	3000	40	20000	0.1
placenta2	5000	Inflection pt.	0.1	3000	40	20000	0.1
placenta3	5000	Inflection pt.	0.1	3000	40	20000	0.9
placenta4	5000	Inflection pt.	0.1	3000	40	20000	0.9
placenta5	5000	Inflection pt.	0.1	5000	40	20000	0.75
placenta6	5000	Inflection pt.	0.1	10000	40	20000	0.1
3659_colon	5000	Inflection pt.	0.1	10000	40	20000	0.1

Supplementary Table 8. AUC values for dropkick_label versus CellRanger_2 for 13 10x Genomics scRNA-seq datasets with “lower_prop” parameter titrated from 10 to 25 % of total barcodes.

	dropkick label total cells	Lower = 0.1 total cells	Lower = 0.1 AUC	Lower = 0.15 total cells	Lower = 0.15 AUC	Lower = 0.2 total cells	Lower = 0.2 AUC	Lower = 0.25 total cells	Lower = 0.25 AUC
293t	3224	2897	0.9994	2849	0.9993	2769	0.9992	2635	0.9990
jurkat	4020	3259	0.9989	3196	0.9988	3080	0.9986	2907	0.9984
neuron_9k	11050	9094	0.9964	8043	0.9960	6843	0.9952	5482	0.9940
neurons_900	1297	792	0.9963	531	0.9992	340	0.9990	253	0.9989
pbmc_4k	4830	4335	0.9999	4018	0.9998	3560	0.9995	2622	0.9989
t_4k_small	4716	4494	0.9998	4362	0.9998	4224	0.9997	3994	0.9996
placenta1	6285	5636	0.9999	4735	0.9995	3752	0.9991	3015	0.9988
placenta2	2771	2101	0.9997	1757	0.9996	1486	0.9996	1289	0.9995
placenta3	5373	4019	0.9997	3600	0.9994	3075	0.9990	2461	0.9985
placenta4	5038	3906	0.9996	3506	0.9994	3040	0.9992	2542	0.9990
placenta5	4064	2723	0.9994	2362	0.9993	2041	0.9990	1795	0.9988
placenta6	3942	3289	0.9998	2768	0.9994	2258	0.9991	1727	0.9988
3659_colon	3376	3063	0.9984	2836	0.9978	2600	0.9969	2203	0.9942

Supplementary Table 9. AUC values for dropkick_label versus CellRanger_2 for 33 inDrop scRNA-seq datasets with “lower_prop” parameter titrated from 10 to 25 % of total barcodes.

	dropkick label total cells	Lower = 0.1 total cells	Lower = 0.1 AUC	Lower = 0.15 total cells	Lower = 0.15 AUC	Lower = 0.2 total cells	Lower = 0.2 AUC	Lower = 0.25 total cells	Lower = 0.25 AUC
2771_S1	2222	3442	0.9103	2613	0.9241	2072	0.9272	1648	0.9165
2771_S2	1248	1418	0.9192	925	0.9237	656	0.9231	457	0.9139
2771_S3	1382	2243	0.8792	1433	0.9150	1072	0.9270	835	0.9407
3072_S1	962	1523	0.9619	1037	0.9763	846	0.9787	728	0.9795
3072_S2	2370	5960	0.8171	3232	0.8808	2120	0.9235	1614	0.9431
3072_S3	3112	7970	0.9002	5963	0.8649	4139	0.9013	3131	0.9275
3254_S1	1285	1273	0.9516	979	0.9552	776	0.9413	613	0.9312
3777_S1	1066	800	0.9655	629	0.9502	478	0.9318	354	0.9165
3777_S2	927	989	0.9677	731	0.9783	582	0.9735	460	0.9681
3907_S1	2260	2888	0.9379	2147	0.9726	1874	0.9831	1658	0.9852
3907_S2	1828	2057	0.9652	1506	0.9742	1194	0.9784	994	0.9771
4058_colon	2186	2042	0.9787	1724	0.9769	1381	0.9730	1034	0.9664
MPP_03A	945	2058	0.7972	1352	0.8741	997	0.9224	770	0.9409
MPP_04A	702	912	0.9151	734	0.9316	593	0.9232	483	0.9120
MPP_07A	1384	1729	0.9601	1363	0.9832	1218	0.9852	1102	0.9819
MPP_08A	1473	4403	0.7994	2571	0.8820	1742	0.9236	1286	0.9422
MPP_13A	1815	2869	0.9159	1837	0.9532	1383	0.9545	1076	0.9453
MPP_18A	875	1563	0.8420	940	0.9237	717	0.9654	616	0.9739
MPP_19A	1503	1483	0.9811	1193	0.9806	970	0.9737	811	0.9671
MPP_21A	2149	5868	0.8171	3287	0.9153	2330	0.9508	1793	0.9596
MPP_48A	811	456	0.9317	303	0.9637	227	0.9705	165	0.9721
MPX_03A	1228	1199	0.9468	1075	0.9573	964	0.9535	846	0.9433
MPX_03B	407	761	0.9343	568	0.9657	506	0.9727	447	0.9743
MPX_04A	744	4089	0.7892	1820	0.8327	1114	0.8929	800	0.9413
MPX_08A	1277	2103	0.9105	1333	0.9472	1016	0.9648	821	0.9680
MPX_08B	947	2654	0.8298	1484	0.9019	1118	0.9540	962	0.9751
MPX_09A	549	981	0.9169	675	0.9494	511	0.9620	411	0.9638
MPX_09B	1462	2187	0.9199	1489	0.9674	1243	0.9779	1092	0.9814
MPX_17A	889	2189	0.8975	1386	0.9292	988	0.9536	775	0.9663
MPX_19A	857	1562	0.9031	1000	0.9470	718	0.9702	572	0.9785
MPX_19B	2830	3251	0.9670	2573	0.9684	2045	0.9481	1570	0.9286
MPX_21A	1647	2743	0.9204	1978	0.9564	1651	0.9697	1407	0.9733
MPX_21B	1377	2433	0.8975	1582	0.9413	1207	0.9643	1016	0.9678

Supplementary Table 10. AUC values for dropkick_label versus EmptyDrops for 13 10x Genomics scRNA-seq datasets with “lower” parameter titrated between 500 and 2,000 total UMI counts.

	dropkick label total cells	Lower = 500 total cells	Lower = 500 AUC	Lower = 1000 total cells	Lower = 1000 AUC	Lower = 1500 total cells	Lower = 1500 AUC	Lower = 2000 total cells	Lower = 2000 AUC
293t	3224	2299	0.9994	2299	0.9994	2233	0.9994	2233	0.9994
jurkat	4020	2565	0.9991	2565	0.9991	2565	0.9991	2565	0.9991
neuron_9k	11050	6048	0.9944	6053	0.9944	6068	0.9944	6137	0.9944
neurons_900	1297	581	0.9984	581	0.9984	585	0.9984	586	0.9984
pbmc_4k	4830	3378	0.9994	3325	0.9993	3304	0.9993	3269	0.9993
t_4k_small	4716	3398	0.9996	3398	0.9996	3398	0.9996	3387	0.9996
placenta1	6285	4349	0.9994	4347	0.9994	4334	0.9994	4296	0.9993
placenta2	2771	601	0.9994	601	0.9994	601	0.9994	601	0.9994
placenta3	5373	3232	0.9991	3232	0.9991	3232	0.9991	3232	0.9991
placenta4	5038	3214	0.9993	3214	0.9993	3204	0.9993	3126	0.9993
placenta5	4064	1019	0.9982	1019	0.9982	1019	0.9982	1019	0.9982
placenta6	3942	2115	0.9990	2115	0.9990	2115	0.9990	2114	0.9990
3659_colon	3376	2214	0.9943	2214	0.9943	2214	0.9943	2214	0.9943

Supplementary Table 11. AUC values for dropkick_label versus EmptyDrops for 33 inDrop scRNA-seq datasets with “lower” parameter titrated between 500 and 2,000 total UMI counts.

	dropkick label total cells	Lower = 500 total cells	Lower = 500 AUC	Lower = 1000 total cells	Lower = 1000 AUC	Lower = 1500 total cells	Lower = 1500 AUC	Lower = 2000 total cells	Lower = 2000 AUC
2771_S1	2222	4419	0.9137	4322	0.9135	4093	0.9166	3866	0.9274
2771_S2	1248	2507	0.9382	2424	0.9311	2327	0.9323	2198	0.9369
2771_S3	1382	1527	0.9554	151	0.9537	151	0.9537	13	0.9286
3072_S1	962	4537	0.8438	6105	0.7668	5959	0.7715	6359	0.7560
3072_S2	2370	11818	0.8814	6726	0.8945	7006	0.8858	7000	0.8844
3072_S3	3112	8334	0.8842	7303	0.8734	7291	0.8624	5370	0.8548
3254_S1	1285	1667	0.9762	2295	0.9174	2238	0.9224	2092	0.9351
3777_S1	1066	1598	0.9462	1613	0.9280	1641	0.9148	1757	0.9045
3777_S2	927	2140	0.9366	2027	0.9411	2038	0.9258	2441	0.8935
3907_S1	2260	14527	0.7998	7491	0.7505	11595	0.6960	11905	0.6981
3907_S2	1828	12705	0.7290	6153	0.6936	7132	0.6999	7303	0.7007
4058_colon	2186	11692	0.8506	1298	0.9719	2940	0.9683	2998	0.9648
MPP_03A	945	4841	0.7077	1709	0.9557	1944	0.9281	2073	0.9095
MPP_04A	702	1049	0.9366	1111	0.9123	1116	0.9093	1155	0.8918
MPP_07A	1384	1965	0.8981	2679	0.8647	2827	0.8538	2887	0.8507
MPP_08A	1473	7006	0.7388	2923	0.8823	3537	0.8565	3693	0.8530
MPP_13A	1815	6779	0.8102	3600	0.9046	3796	0.9003	3717	0.8997
MPP_18A	875	5645	0.7244	2266	0.9227	2446	0.9060	2531	0.9009
MPP_19A	1503	2507	0.9332	3340	0.8843	3349	0.8833	3271	0.8883
MPP_21A	2149	7098	0.7913	7146	0.8064	7053	0.8152	6708	0.8326
MPP_48A	811	1289	0.8747	1557	0.8454	1840	0.7984	1938	0.7852
MPX_03A	1228	1820	0.8879	1901	0.8739	1909	0.8706	1942	0.8663
MPX_03B	407	1874	0.8532	2539	0.7739	2613	0.7679	2569	0.7705
MPX_04A	744	11374	0.7947	10713	0.8128	3598	0.8835	3704	0.8636
MPX_08A	1277	6765	0.7626	6654	0.7826	3117	0.9263	3215	0.9190
MPX_08B	947	6718	0.8796	6841	0.8665	2290	0.8498	2534	0.8345
MPX_09A	549	3543	0.7547	1485	0.8908	1543	0.8783	1489	0.8855
MPX_09B	1462	2901	0.9463	4964	0.8871	5104	0.8783	5061	0.8815
MPX_17A	889	1775	0.9054	4291	0.7568	4746	0.7598	4812	0.7601
MPX_19A	857	6200	0.6804	2189	0.8592	2863	0.8379	2971	0.8388
MPX_19B	2830	3468	0.9792	4560	0.9296	4584	0.9292	4468	0.9333
MPX_21A	1647	7168	0.7603	7258	0.7821	3504	0.9489	3576	0.9465
MPX_21B	1377	5983	0.7933	5908	0.8038	2525	0.9298	2755	0.9167

Supplementary Table 12. Code and data resources.

Resource	Source	Identifier
Deposited Data		
293T Cells	Zheng, et al. 2017	support.10xgenomics.com/single-cell-gene-expression/datasets
Jurkat Cells	Zheng, et al. 2017	
9k Neurons	10x Genomics	
900 Neurons	10x Genomics	
4k PBMCs	10x Genomics	
4k Pan-T Cells	10x Genomics	
Placenta	10x Genomics	
3907_S1 & 3907_S2	this manuscript	GSE158636
Software and Algorithms		
Python version 3.8.2	Python Software Foundation	python.org
matplotlib version 3.3.0	Hunter 2007	matplotlib.org
numpy version 1.19.1	Oliphant 2006	numpy.org
pandas version 1.1.0	McKinney, et al. 2010	pandas.pydata.org
SCANPY version 1.5.1	Wolf, et al. 2018	pypi.org/project/scanpy/
scikit-learn version 0.23.1	Pedregosa, et al. 2011	scikit-learn.org
scipy version 1.5.2	Oliphant 2007	scipy.org
seaborn version 0.10.1	Waskom, et al. 2014	seaborn.pydata.org
umap-learn version 0.4.3	McInnes and Healy 2018	github.com/lmcinnes/umap
R version 3.6.3	The R Foundation	r-project.org
Seurat version 3.0.0	Butler, et al. 2018	satijalab.org/seurat
DropletUtils version 3.11	Lun, et al. 2019	10.18129/B9.bioc.DropletUtils
CellBender version 0.2.0	Fleming, Marioni, and Babadi 2019	github.com/broadinstitute/CellBender
UpSetR version 1.4.0	Lex, et al. 2014	doi:10.1109/TVCG.2014.2346248
sc-UniFrac version 0.9.6	Liu, et al. 2018	github.com/liuqivandy/scUnifrac
dropkick version 1.2.3	this manuscript	pypi.org/project/dropkick/