Supplementary material

`distinct`: a novel approach to differential distribution analyses

Simone Tiberi[1*], Helena L Crowell[1], Lukas M Weber[2], Pantelis Samartsidis[3] and Mark D Robinson[1]

[1]*Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland.*

[2]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.*

[3]*MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK.*

[*] e-mail: Simone.Tiberi@uzh.ch

1 Supplementary Tables

	DE	DP	DM	DB	DV	null	average
scDD-perm.vstresiduals	2094.7	2008.8	2019.7	1603.2	1951.9	2142.1	1970.1
scDD-perm.log2-cpm	454.2	439.9	460.0	490.1	397.3	443.5	447.5
MM-nbinom	318.0	314.7	322.0	298.2	207.2	323.7	297.3
MM-vstresiduals	85.5	90.8	87.8	95.0	66.8	81.5	84.6
MM-dream2	30.3	30.7	28.6	31.4	25.2	30.4	29.4
distinct.vstresiduals	6.7	6.0	5.9	3.7	2.7	2.0	4.5
distinct.log2-cpm	5.3	4.4	4.5	3.2	2.2	1.5	3.5
distinct.cpm	5.2	4.5	4.5	3.0	1.9	1.5	3.4
scDD-KS.vstresiduals	0.5	0.4	0.4	0.4	0.5	0.4	0.5
scDD-KS.log2-cpm	0.4	0.4	0.3	0.3	0.4	0.4	0.4
edgeR.cpm	0.2	0.2	0.2	0.2	0.2	0.2	0.2
edgeR.counts	0.2	0.2	0.2	0.1	0.2	0.2	0.2
limma-voom.counts	0.1	0.1	0.1	0.0	0.1	0.1	0.1
limma-trend.vstresiduals	0.1	0.1	0.1	0.0	0.0	0.0	0.1
limma-trend.log2-cpm	0.1	0.1	0.1	0.0	0.0	0.0	0.1

Supplementary Table 1: Computing time, expressed in minutes, in *muscat* main simulations. For each column, times are averaged across the five replicate simulations; in each replicate simulation 3,600 cells are available (200, on average, per cluster-sample combination). 'MM' refers to mixed models, *scDD-KS* indicates *scDD* based on the Kolmogorov-Smirnov test, while *scDD-perm* denotes *scDD* based on the permutation test (100 permutations). *distinct*, MM models and *scDD* were run on 3 cores, while pseudo-bulk methods based on *edgeR* and *limma* used a single core since they do not allow for parellel computing. Note that computing times for *distinct* are significantly larger in non-null datasets (which contain, on average, 10% of differential genes in each cluster), because *distinct* increases the number of permutations used (and hence the computational cost) for significant results. Note that *scDD-perm* required much longer on vstresiduals than on log2-cpms because *scDD* performs differential testing on non-zero values: vstresiduals, unlike log2-cpms, are not zero-inflated and, therefore, many more cells have to be used for differential testing.

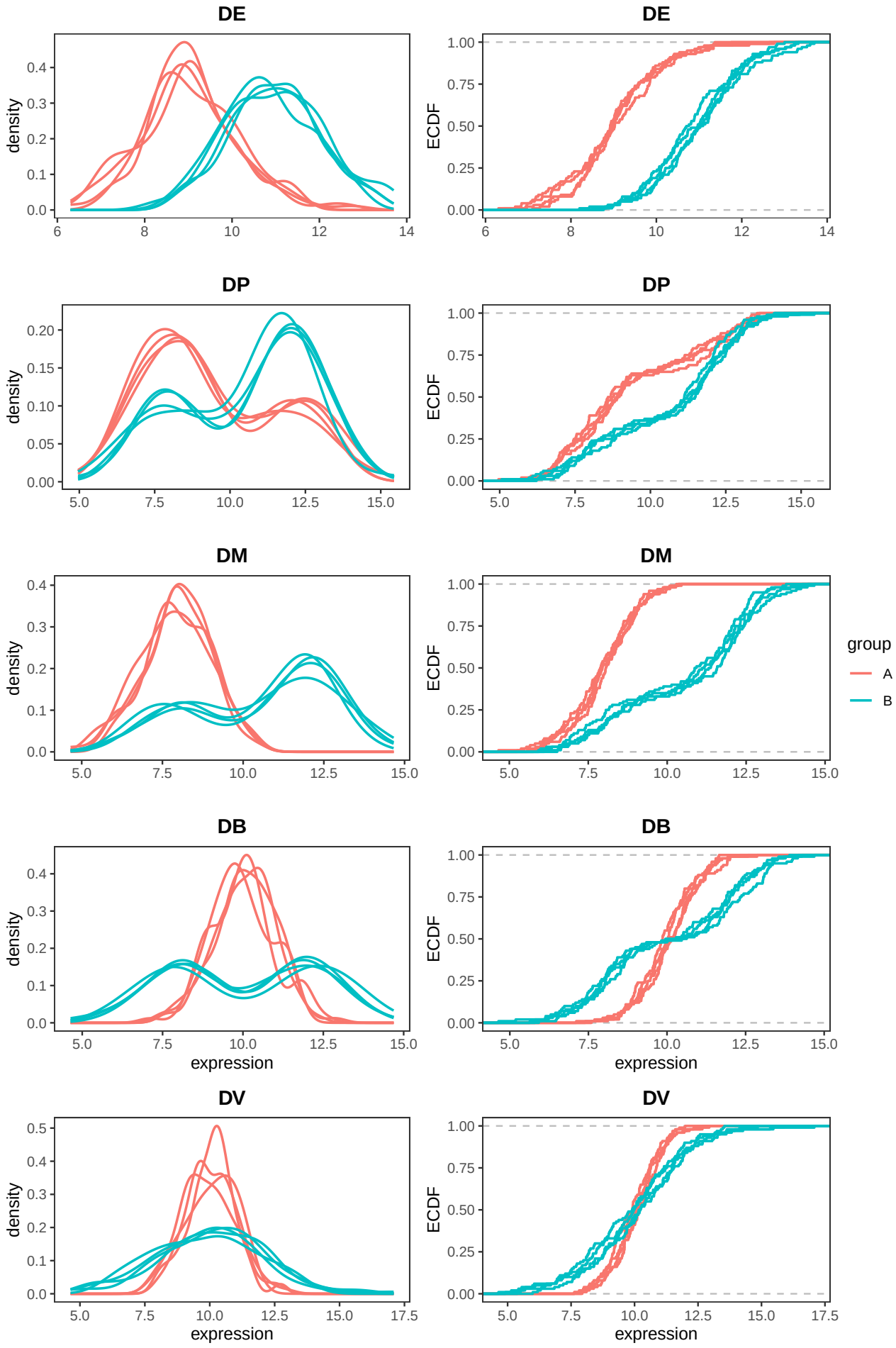
	raw p-value			globally adjusted p-value			locally adjusted p-value		
threshold	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
distinct.cpm	0.12	0.07	0.02	0.00	0.00	0.00	0.00	0.00	0.00
distinct.log2-cpm	0.12	0.08	0.02	0.00	0.00	0.00	0.01	0.00	0.00
distinct.vstresiduals	0.12	0.08	0.02	0.00	0.00	0.00	0.00	0.00	0.00
edgeR.counts	0.07	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00
edgeR.cpm	0.07	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00
limma-trend.log2-cpm	0.28	0.20	0.10	0.02	0.02	0.02	0.11	0.08	0.06
limma-trend.vstresiduals	0.25	0.18	0.08	0.02	0.02	0.02	0.24	0.18	0.08
limma-voom.counts	0.12	0.07	0.02	0.00	0.00	0.00	0.02	0.01	0.01

Supplementary Table 2: Fraction of false positive tests returned from each method, at the 0.01, 0.05 and 0.1 significance thresholds, in the null *T-cells* experimental data. Values are averages across the three replicates.

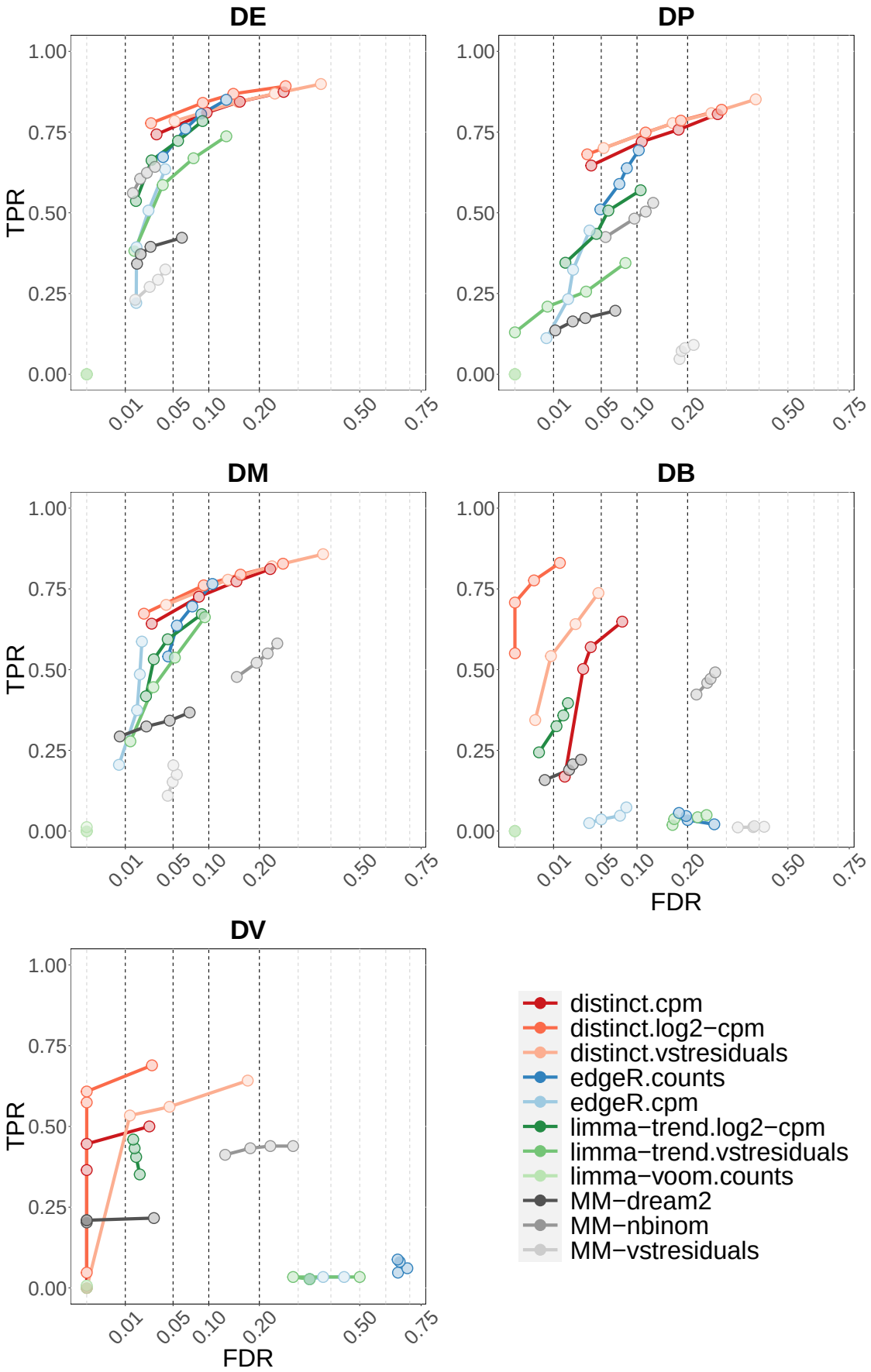
	raw p-value			globally adjusted p-value			locally adjusted p-value		
threshold	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
distinct.cpm	0.13	0.08	0.02	0.01	0.00	0.00	0.01	0.01	0.00
distinct.log2-cpm	0.13	0.08	0.03	0.01	0.01	0.00	0.01	0.01	0.00
distinct.vstresiduals	0.14	0.09	0.03	0.01	0.00	0.00	0.01	0.01	0.00
edgeR.counts	0.08	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00
edgeR.cpm	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
limma-trend.log2-cpm	0.19	0.10	0.02	0.00	0.00	0.00	0.03	0.02	0.00
limma-trend.vstresiduals	0.23	0.13	0.03	0.00	0.00	0.00	0.02	0.02	0.00
limma-voom.counts	0.08	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Supplementary Table 3: Fraction of false positive tests returned from each method, at the 0.01, 0.05 and 0.1 significance thresholds, in the null *Kang* experimental data. Values are averages across the three replicates.

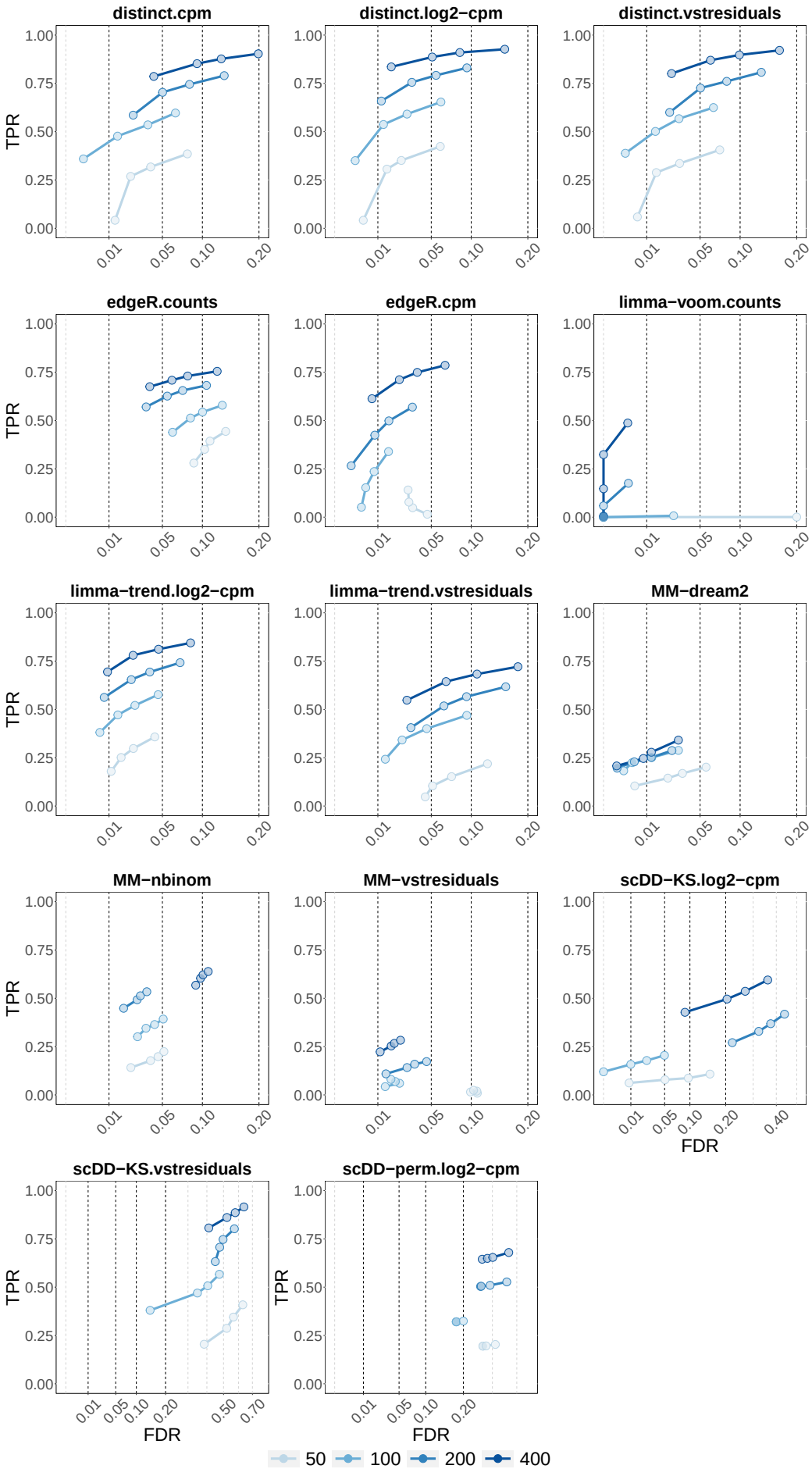
2 Supplementary Figures



Supplementary Figure 1: Density and corresponding empirical cumulative distribution function (ECDF) for differential patterns DE, DP, DM, DB and DV, as illustrated in Figure 1. Two groups of four samples each are compared; 100 cells were simulated from each sample.

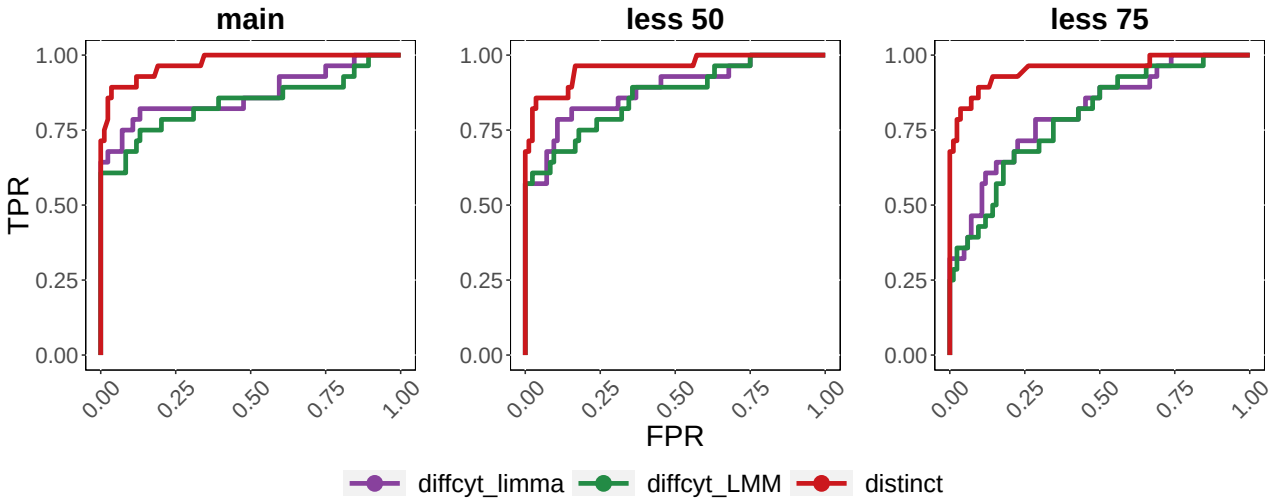


Supplementary Figure 2: TPR vs. FDR in *muscat* simulated data with batch effects; DE, DP, DM, DB and DV refer to the differential profiles illustrated in Figure 1. Each simulation compares two groups of 3 samples each; on average, 200 cells are available for every sample in each of three clusters. Results are averages across the five simulation replicates. Circles indicate observed FDR for 0.01, 0.05, 0.1 and 0.2 significance thresholds. Note that *scDD* methods were excluded from this analysis because *scDD* does not directly handle covariates.

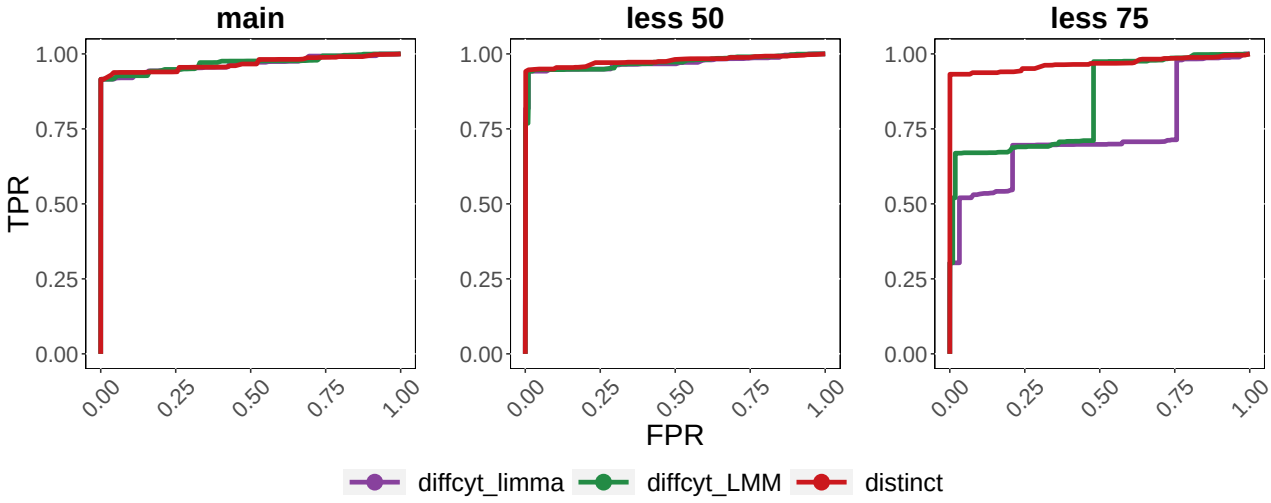


Supplementary Figure 3: True positive rate (TPR) vs. false discovery rate (FDR) in *muscat* simulated data, with 50, 100, 200 and 400 cells per cluster-sample combination (the simulation with 200 cells the main *muscat* simulation, in Figure 2b), corresponding to a total of 900, 1,800, 3,600 and 7,200 cells, respectively. Each plot represents results are aggregated over the five replicate simulations of the four differential types, DE, DP, DM, DB and DV, contributing in equal fraction. Each individual simulation replicate consists of 4,000 genes, 3 cell clusters and two groups of 3 samples each. Circles indicate observed FDR for 0.01, 0.05, 0.1 and 0.2 significance thresholds. Note that *scDD-perm.vstresiduals* was excluded from this analysis due to its computational cost.

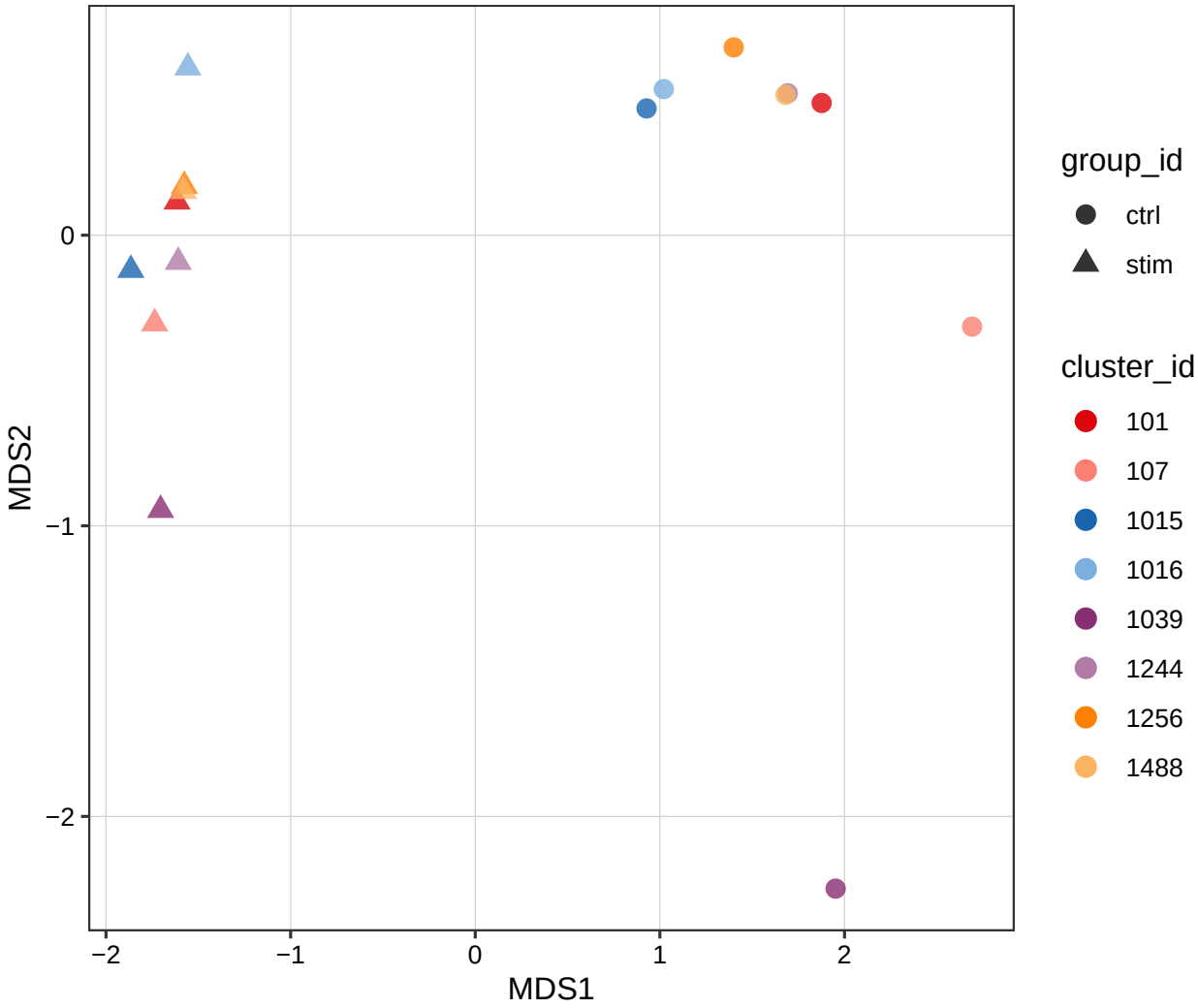
a



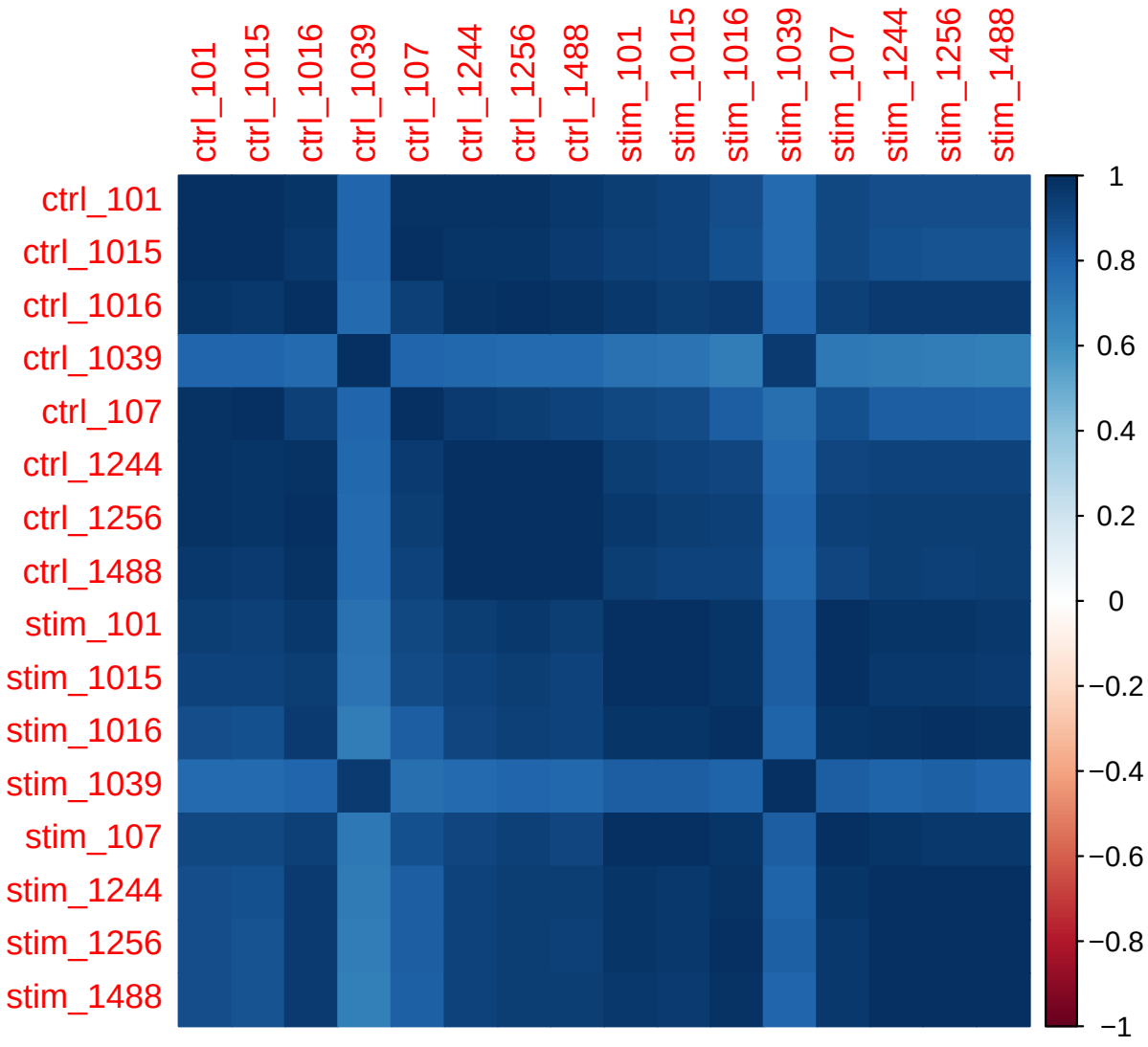
b



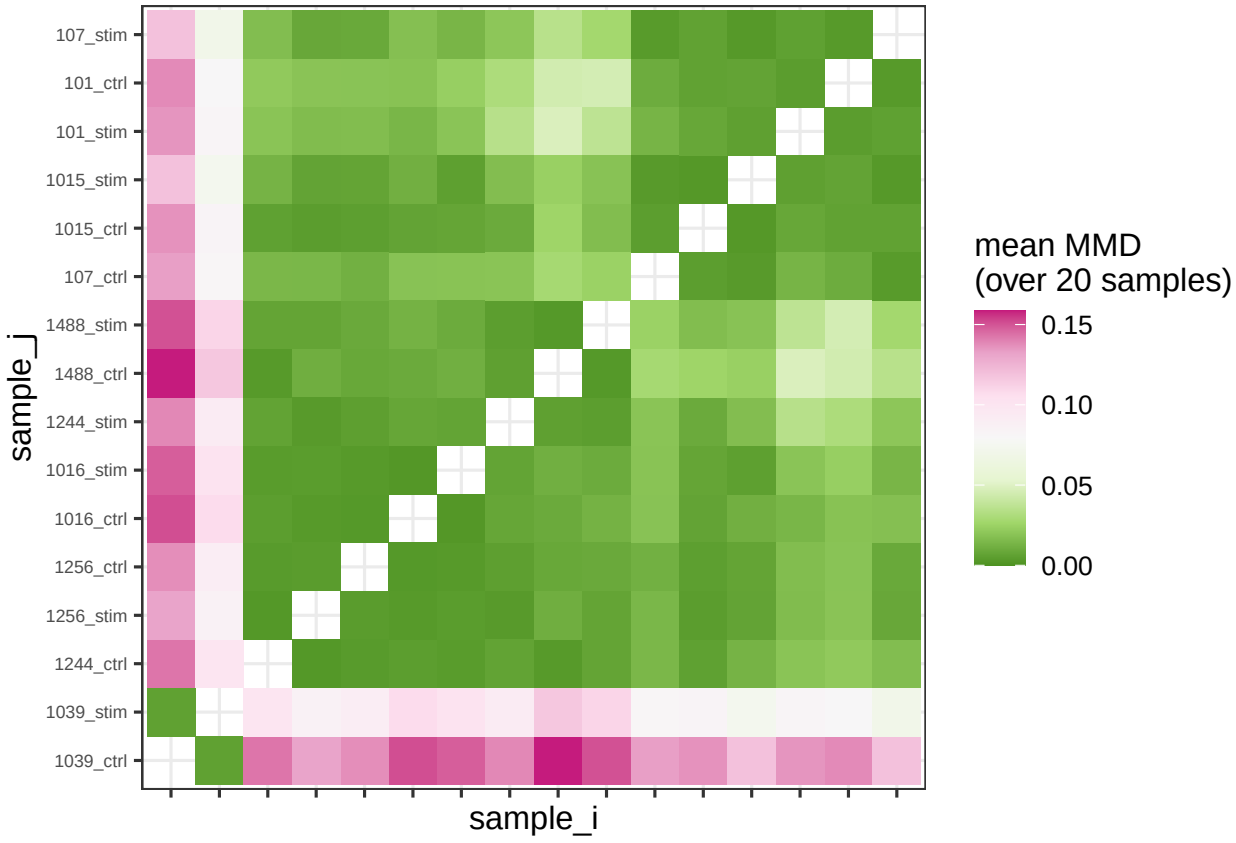
Supplementary Figure 4: Receiver operating characteristic (ROC) curve in *diffcyt* non-null semi-simulated data. Each simulation consists of 88,435 cells and two groups of 8 samples each. 'main', 'less 50' and 'less 75' indicate the main simulation, and those where differential effects are diluted by 50 and 75%, respectively. We evaluated methods' performance in terms of detecting DS for phosphorylated ribosomal protein S6 (pS6) in B cells, which is the strongest differential signal across the cell types in this dataset [5,6]. **(a)** As in the *muscat* simulation study, cells were clustered based on manually annotated cell types [6]. **(b)** As in Weber *et al.* [6], cells were clustered in an unsupervised manner.



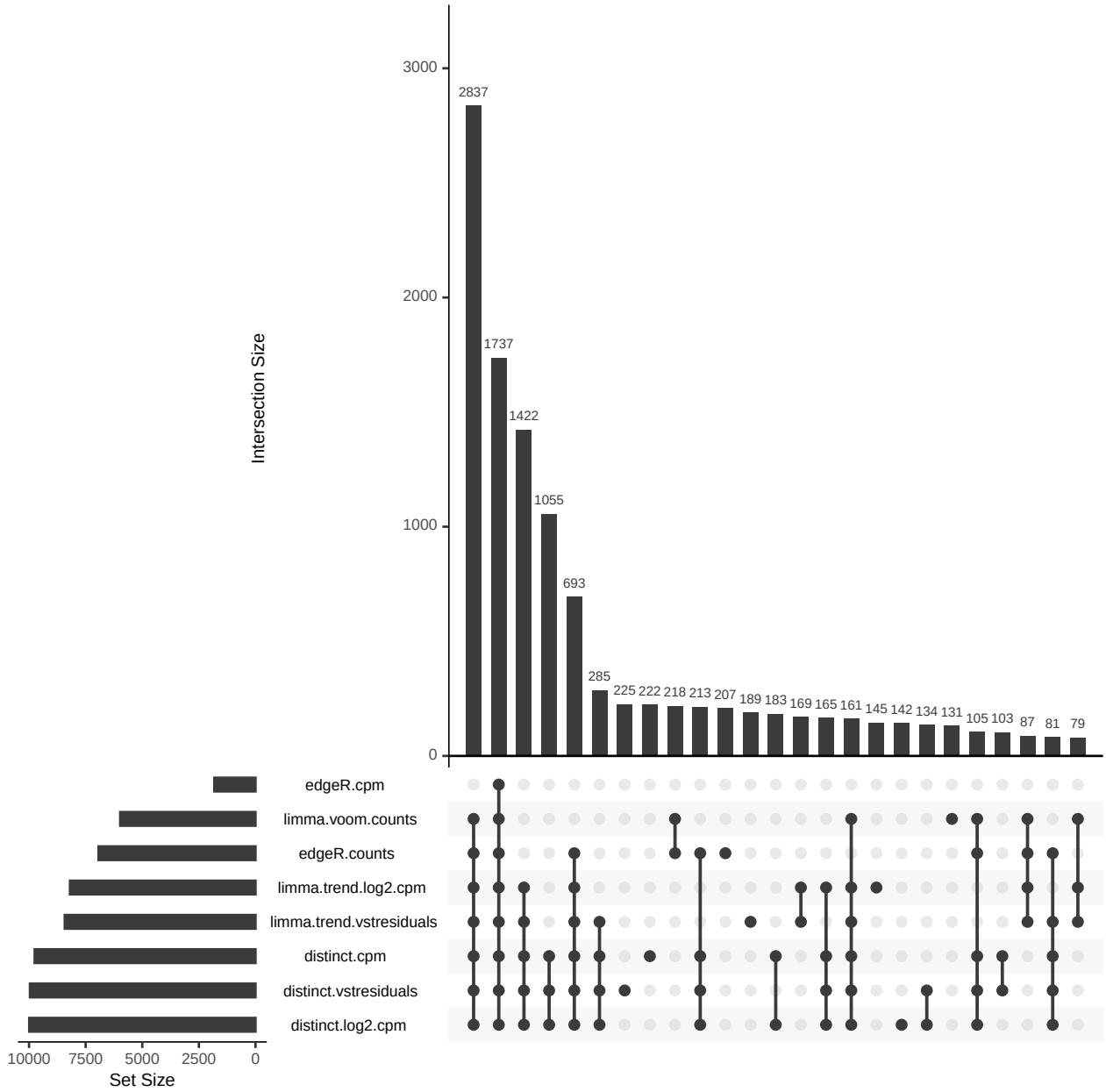
Supplementary Figure 5: Multidimensional scaling (MDS) plot, performed via *muscat*'s *pbMDS* function [1], of the pseudo-bulk counts per million (CPM) from the *Kang* dataset [3]. Legend label *cluster_id* refers to the id of samples, where sample 1039 emerges as a potential outlier in both control and stimulated conditions.



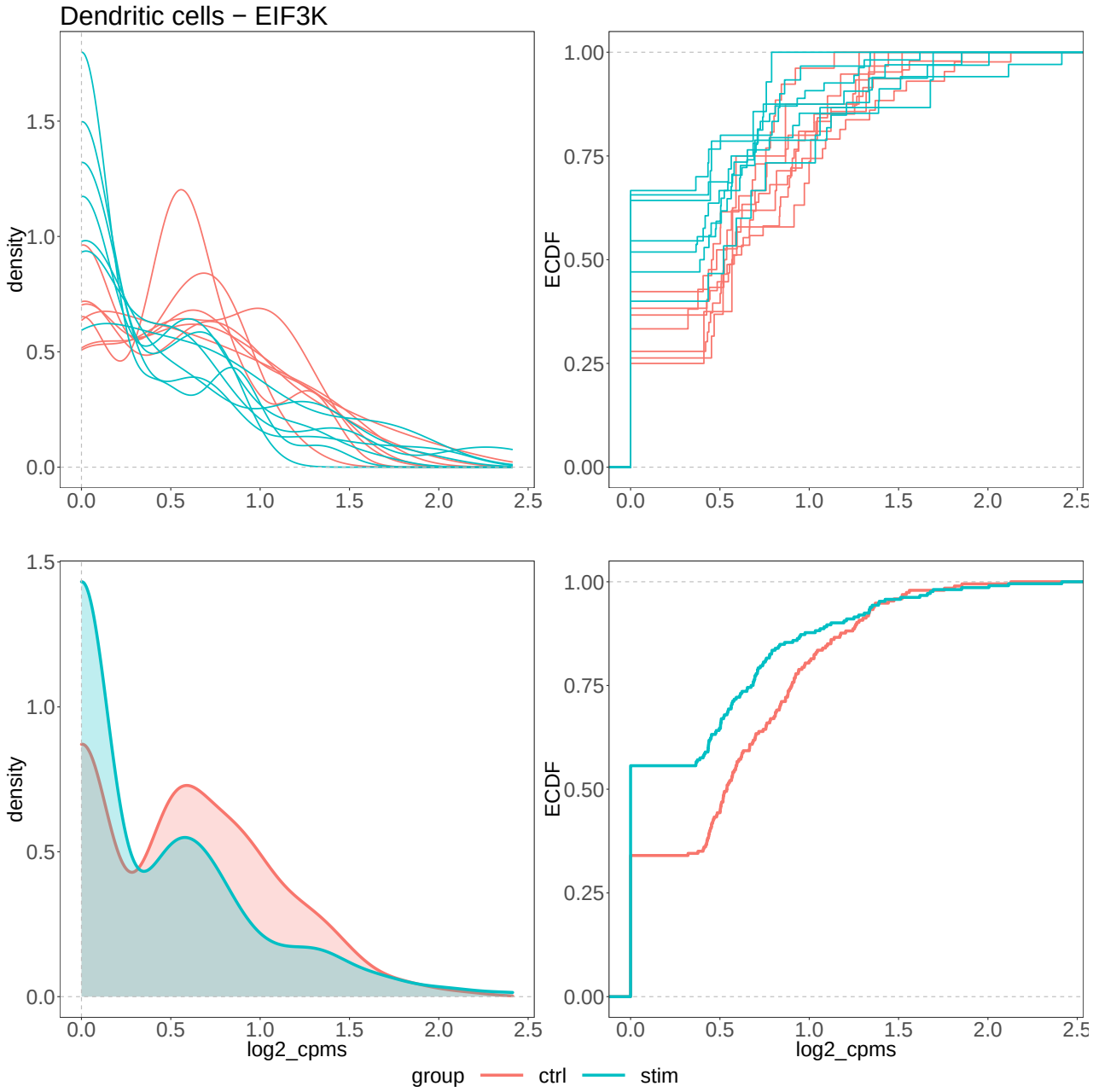
Supplementary Figure 6: Visual representation of the correlation, between samples, of the pseudo-bulk CPMs from the *Kang* dataset [3]. Again, sample 1039 behaves as a potential outlier in both control (ctrl_1039) and stimulated (stim_1039) conditions.



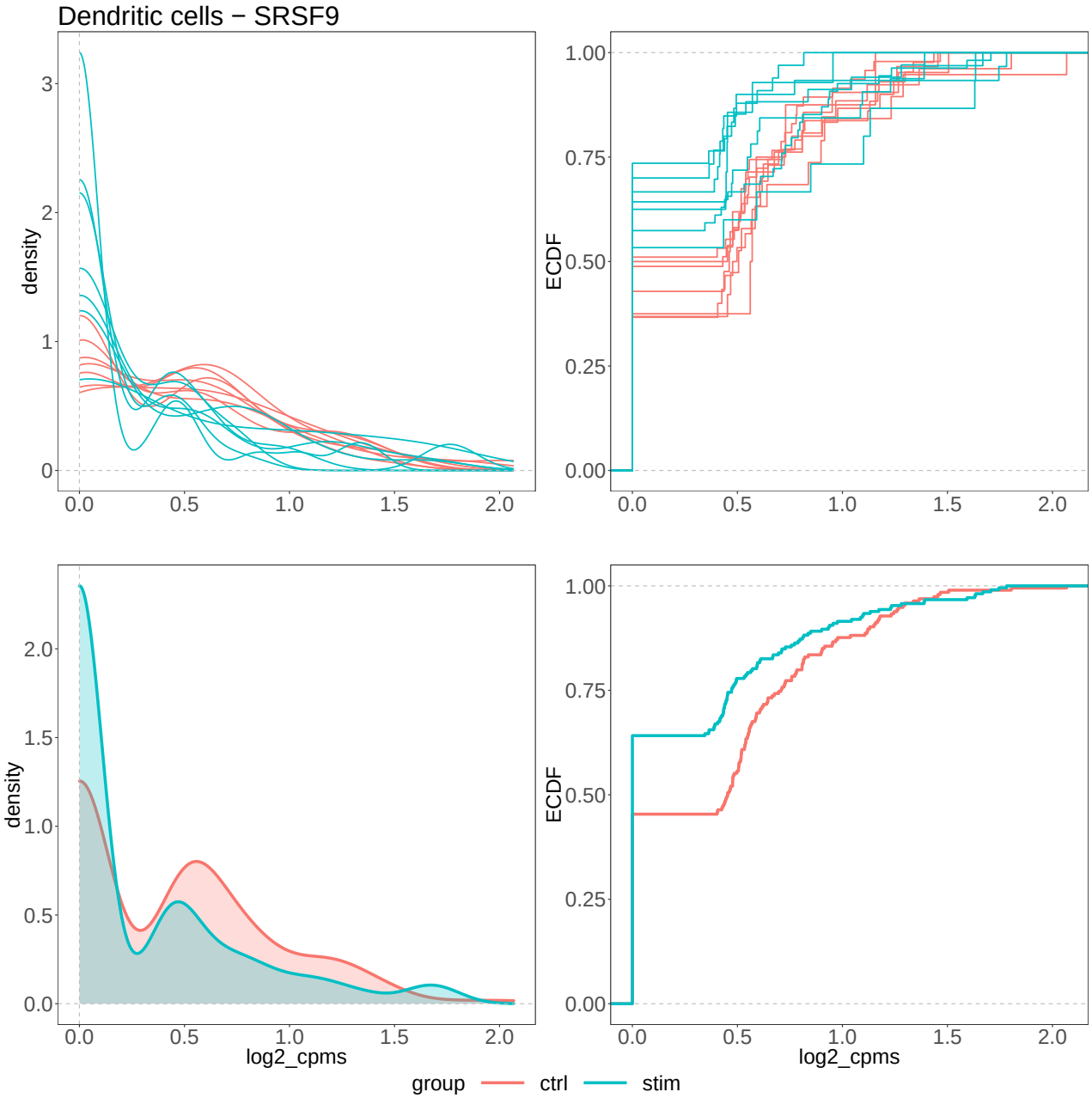
Supplementary Figure 7: Maximum mean discrepancy (MMD) of quality controls of samples, performed via *SampleQC*'s *plot_mmd_heatmap* function [4], from the *Kang* dataset [3]. Again, sample 1039 appears to be a potential outlier in both control (ctrl_1039) and stimulated (stim_1039) conditions.



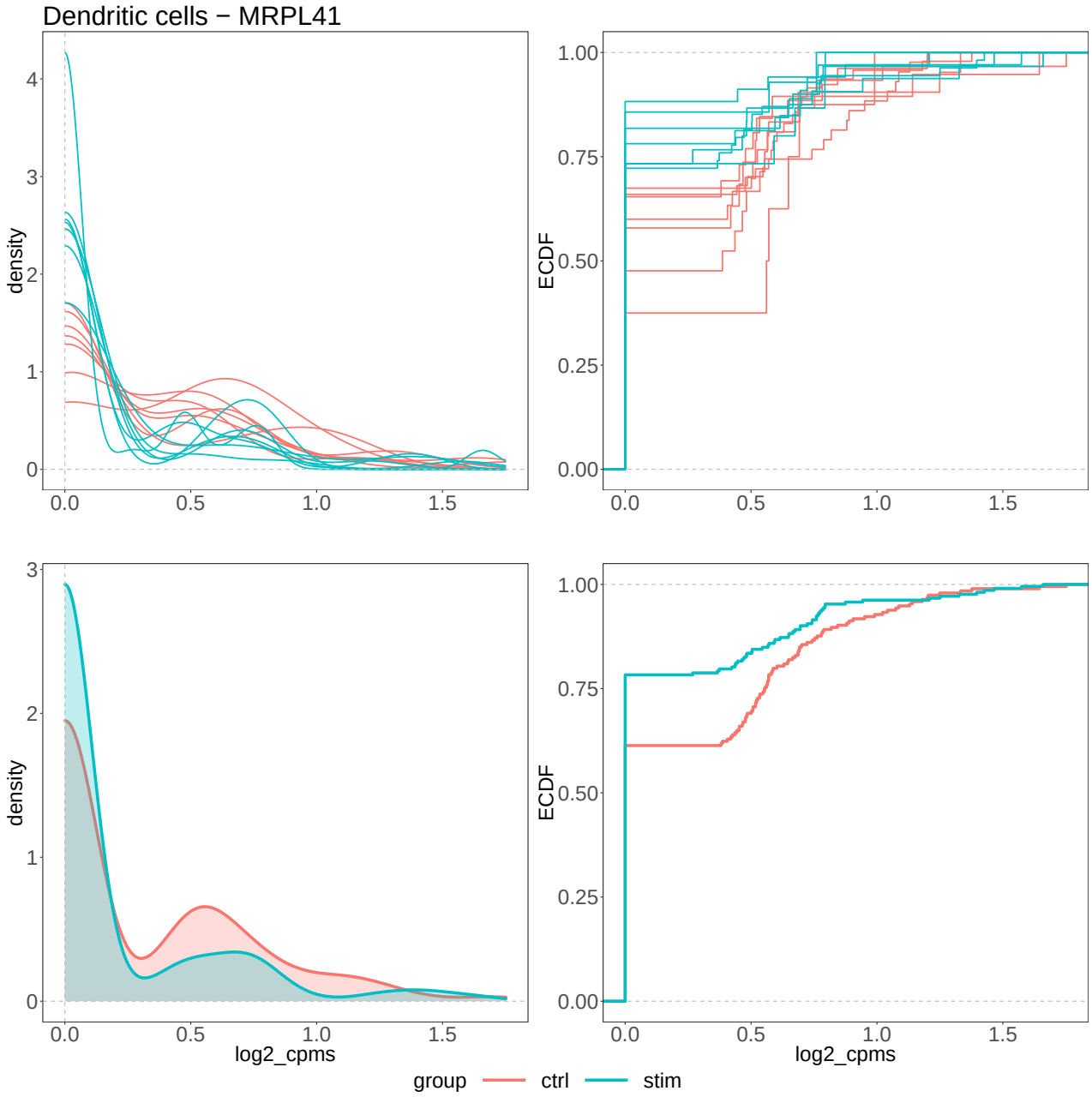
Supplementary Figure 8: Visualization (obtained via *UpSetR* R package [2]) of set intersections between differential patterns identified by each method on the *Kang* dataset, when comparing controls and stimulated samples (adjusted p-value < 0.05). *distinct* identifies more patterns than PB methods, with *edgeR* based on CPMs being the most conservative approach. Note that *distinct* results are highly coherent across different input data.



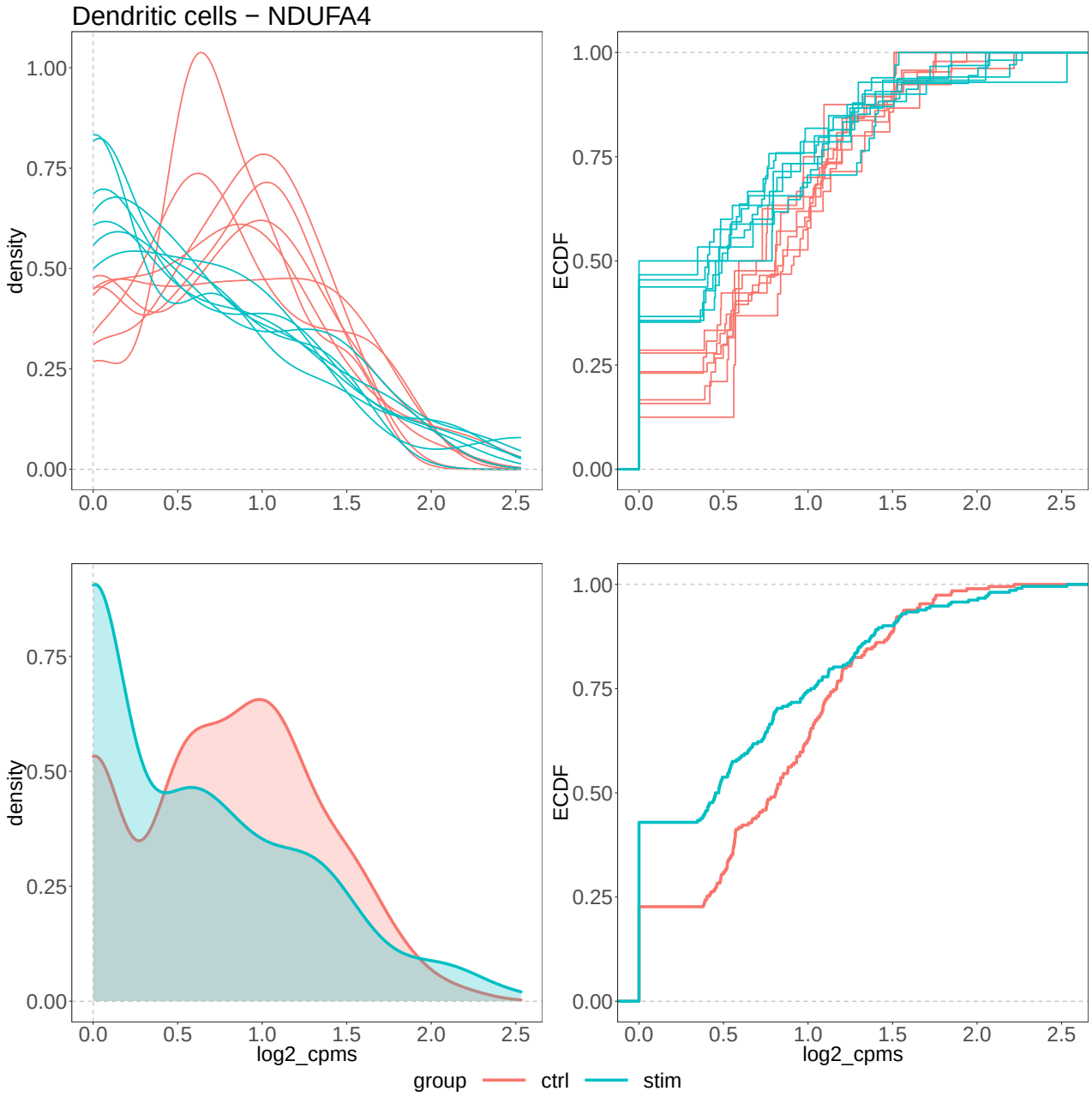
Supplementary Figure 9: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene EIF3K in Dendritic cells, resembling a DP differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



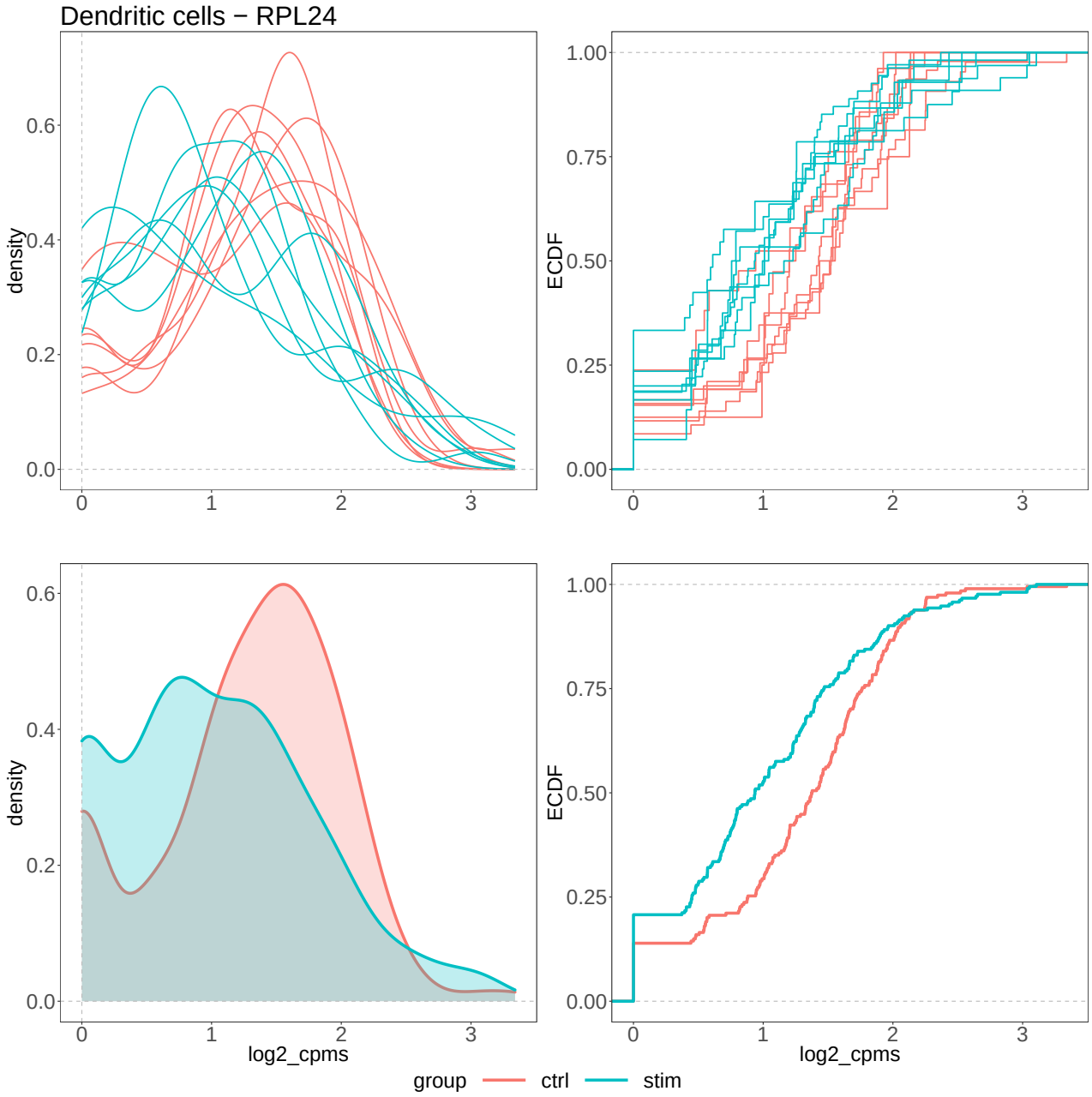
Supplementary Figure 10: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene SRSF9 in Dendritic cells, resembling a DP differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



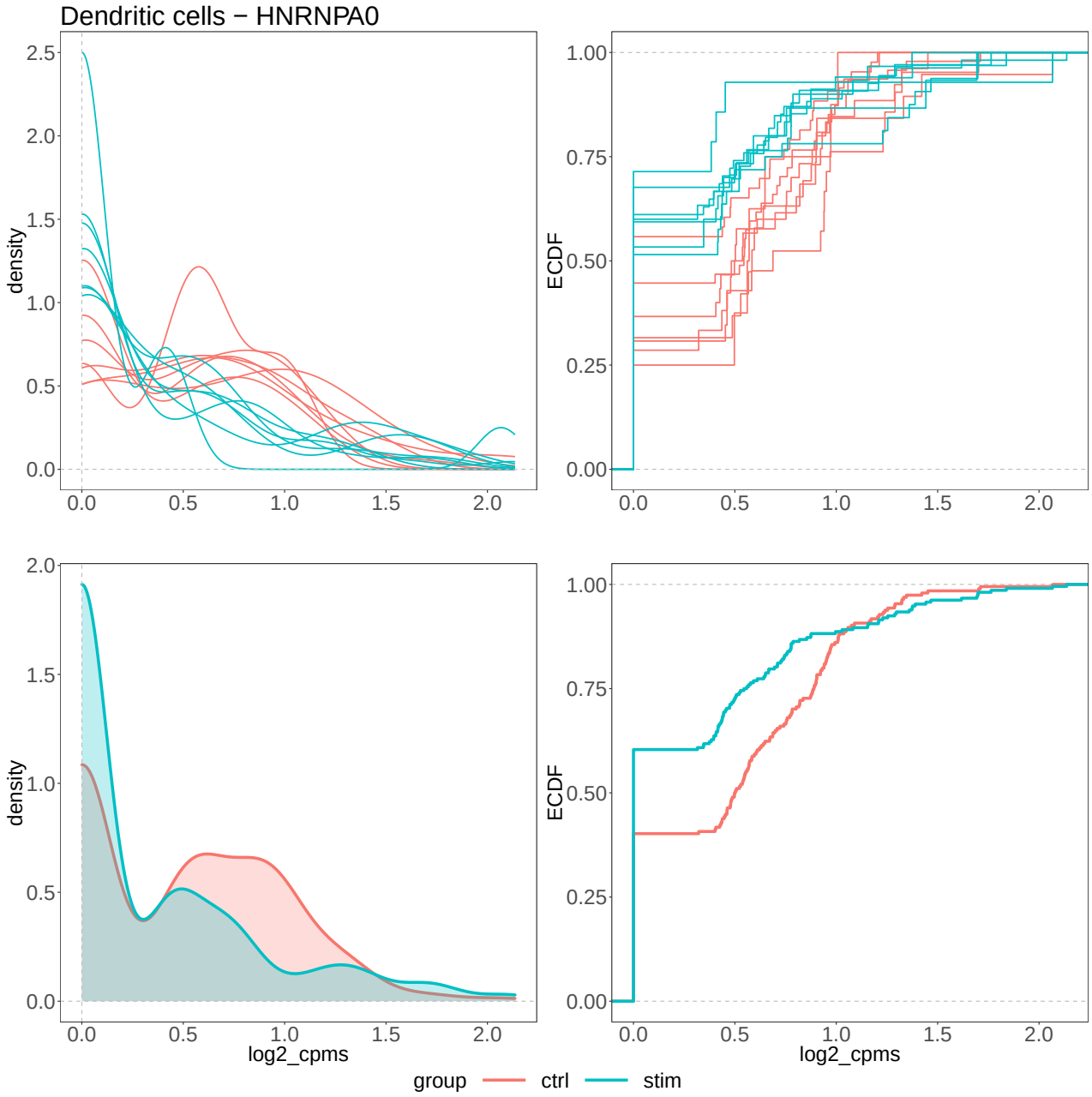
Supplementary Figure 11: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene MRPL41 in Dendritic cells, resembling a DP differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



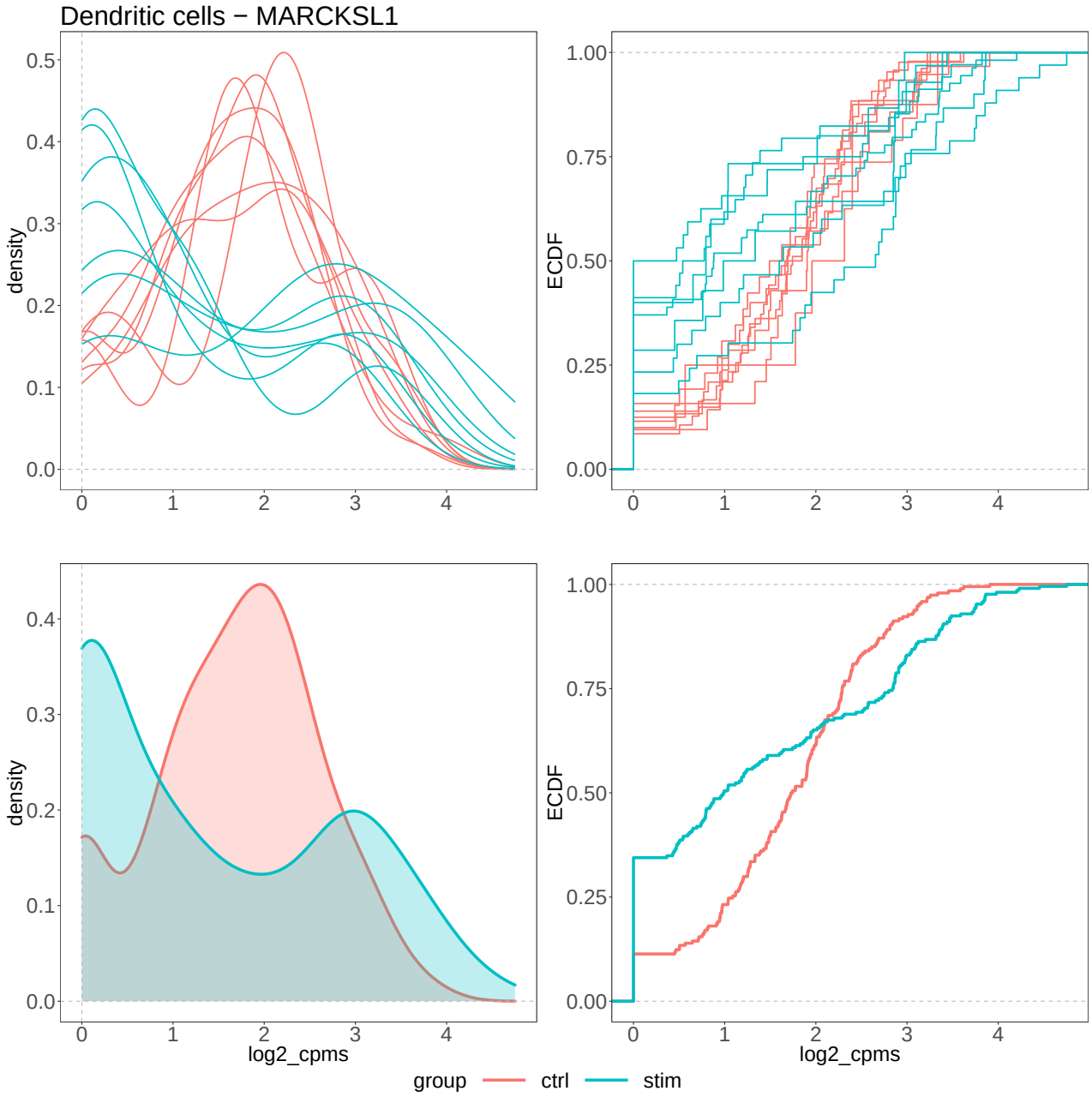
Supplementary Figure 12: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene NDUFA4 in Dendritic cells, resembling a DP differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



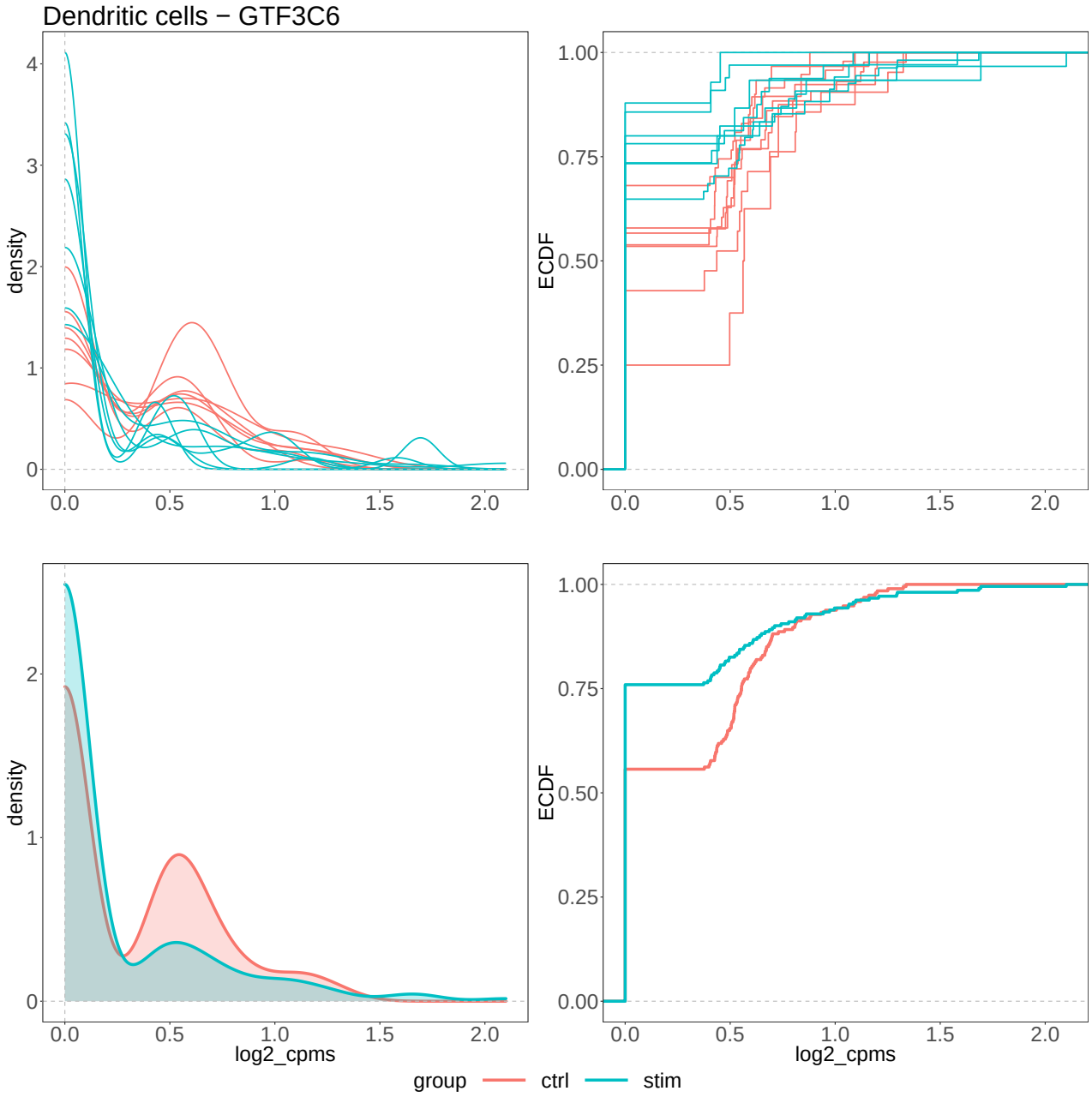
Supplementary Figure 13: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene RPL24 in Dendritic cells, resembling a DP differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



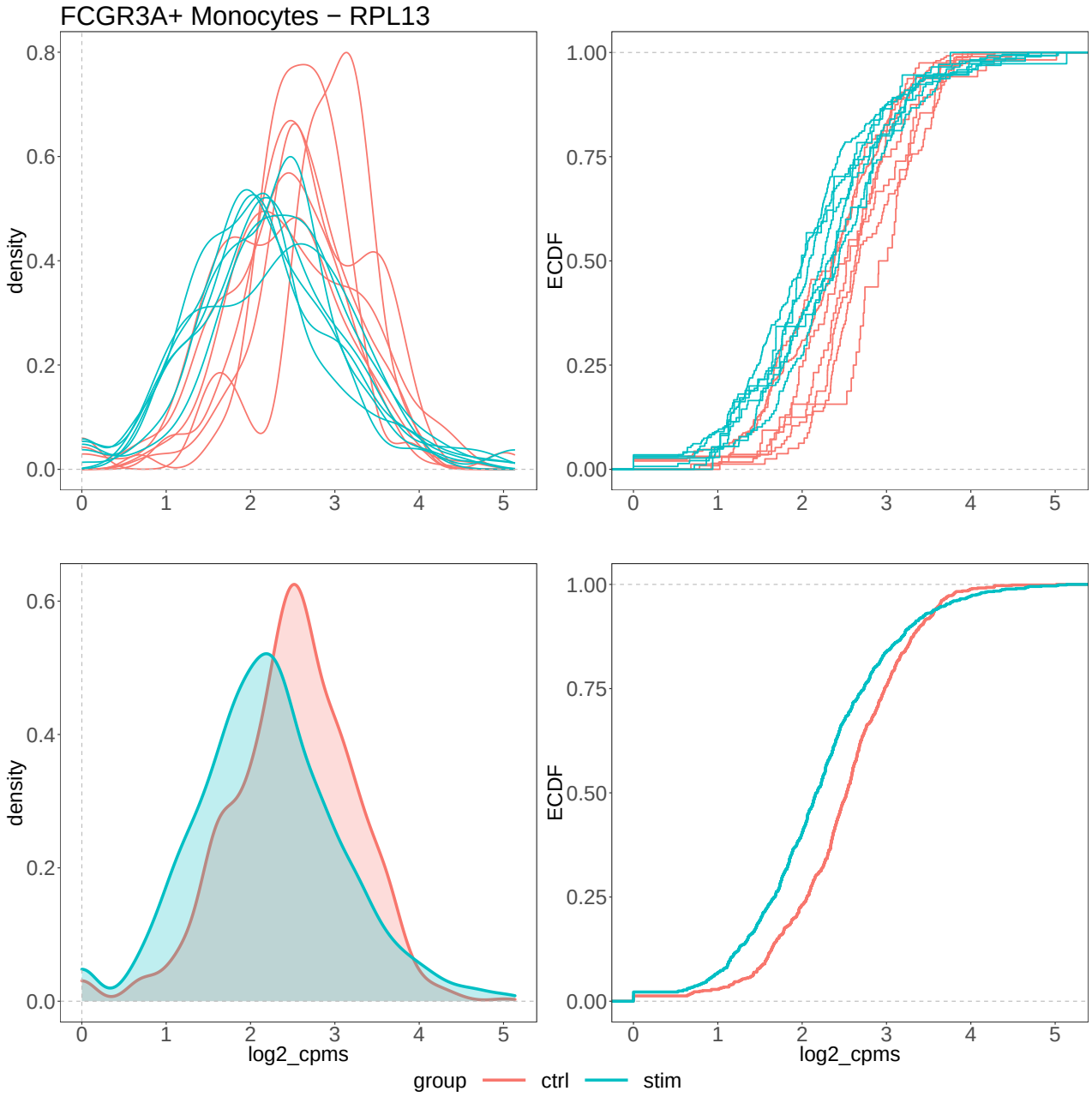
Supplementary Figure 14: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene HNRNPA0 in Dendritic cells, resembling a DP differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



Supplementary Figure 15: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene MARCKSL1 in Dendritic cells, resembling a DB differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



Supplementary Figure 16: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene GTF3C6 in Dendritic cells, resembling a DP differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.



Supplementary Figure 17: Density (left panels) and ECDF (right panels), for the log2-CPMs of gene RPL13 in FCGR3A+ Monocytes cells, resembling a DE differential pattern. Top panels display one line per sample, while bottom panels show results aggregated at the group level. This result is identified by *distinct* on all input data (adjusted p-value < 0.05), and not by any PB tool (adjusted p-value > 0.05), on the *Kang* dataset, when comparing controls and stimulated samples.

References

[1] H. L. Crowell, C. Soneson, P.-L. Germain, D. Calini, L. Collin, C. Raposo, D. Malhotra, and M. D. Robinson. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1):1–12, 2020.

[2] N. Gehlenborg. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*, 2019. R package version 1.4.0.

[3] H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89, 2018.

[4] W. Macnair. *SampleQC: Robust, multivariate, multi-sample, multi-celltype QC for single cell RNA-seq data*, 2020. R package version 0.1.0.

[5] M. Nowicka, C. Krieg, L. M. Weber, F. J. Hartmann, S. Guglietta, B. Becher, M. P. Levesque, and M. D. Robinson. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6, 2017.

[6] L. M. Weber, M. Nowicka, C. Soneson, and M. D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications biology*, 2(1):1–11, 2019.