# Differential dropout analysis captures biological variation in single-cell RNA sequencing data

Gerard A. Bouland[1,2] (g.a.bouland@tudelft.nl)

Ahmed Mahfouz[1,2,3,*] (a.mahfouz@lumc.nl)

Marcel J.T. Reinders[1,2,3,*] (m.j.t.reinders@tudelft.nl)

[1] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[2] Department of Human Genetics, Leiden University Medical Center, Leiden 2333ZC, The Netherlands

[3.] Leiden Computational Biology Center, Leiden University Medical Center, Leiden 2333ZC, The Netherlands

* Corresponding authors: Ahmed Mahfouz and Marcel J.T. Reinders

**Abstract**

Single-cell RNA sequencing data is characterized by a large number of zero counts, yet there is growing evidence that these zeros reflect biological variation rather than technical artifacts. We propose differential dropout analysis (DDA) to identify the effects of biological variation in single-cell RNA sequencing data. Using 16 publicly available and simulated datasets, we show that DDA accurately detects biological variation and can assess the relative abundance of transcripts more robustly than methods relying on counts. DDA is available at https://github.com/gbouland/DDA.

22 **Main**

23 Single cell RNA sequencing (scRNAseq) data is highly sparse, and the common belief is that the zero
24 values are primarily caused by technical artifacts (i.e. dropouts). Although more zeros are observed in
25 scRNAseq data than expected, these can largely be explained by biological rather than technical factors[1].
26 Also, the amount of zeros in scRNAseq is in line with distributional models of molecule sampling counts[2].
27 Methods that utilize dropout patterns for feature selection[3,4] and cell type clustering[5] have recently
28 been developed and perform better or comparable with methods relying on highly variable genes. For
29 instance, Qiu[5] binarized scRNAseq count data, where dropouts remain zero and every non-zero value
30 was assigned a one. With this binary representation, cell type clusters were identified based on co-
31 occurrence of transcripts. Yet, it is not clear whether dropouts also reflect differences across distinct
32 biological cell populations. Therefore, we investigated whether biological differences across cell
33 population can be identified using differential dropout analysis (DDA), rather than the commonly used
34 differential expression analysis (DEA). Instead of relying on changes in the expression value of genes
35 across cell populations, which can be sparse and are subject to pre-processing steps, we analyzed the
36 binary dropout patterns across biological distinct cell populations, i.e. are there more (or less) dropouts
37 for a gene in condition *A* compared to condition *B*.

38 As proof of concept, we performed DDA with a simple logistic regression on binarized expression
39 profiles from 16 scRNAseq datasets (662,825 cells in total, Supplementary Table 1). We compared the
40 results of DDA with those of differently expressed genes (DEGs) detected using the commonly used
41 Wilcoxon Rank Sum test, which is also top ranked for single cell analyses[6,7]. We tested each gene using
42 both DDA and DEA for differences between conditions (6 datasets), cell-types (6 datasets) and normal-
43 versus cancerous tissues (4 datasets, Fig. 1a). Across all datasets, a total of 96,275 significant genes ($P_{FDR}$
44 $\leq 0.05$) were identified with either DDA (92,381 genes) or DEA (91,521 genes). Of these, 87,627 were
45 identified by both tests, resulting in a Jaccard index of 0.91. This high degree of agreement is also

46 reflected in each individual dataset (median = 0.92, minimum = 0.76, and maximum = 0.99). We did not
47 use a log fold-change (logFC) or log odds-ratio (logOR) threshold, as for each dataset and comparison
48 different thresholds are appropriate. In all datasets, the logFC and logOR were significantly (spearman)
49 correlated (median(ρ) = 0.90, minimum(ρ) = 0.49, and maximum(ρ) = 0.98, $P \leq 5 \times 10^{-100}$). The three
50 datasets with the lowest correlation coefficient between logOR and logFC ($\rho \leq 0.62$) were datasets
51 generated using the Smart-seq protocol (Supplementary Table 1). Across the datasets, we observed an
52 average increase of 1.80 in logOR (median = 1.70, $Q_1 = 1.59$, $Q_3 = 2.10$), for every increase in logFC (see
53 Fig. 1b for the *cancer atlas (2)* dataset[8]). The high degree of agreement of detected genes shows that
54 DDA performs on par with the Wilcoxon Rank Sum test and the strong correlation of the logFC and
55 logOR across all datasets shows that the results can be interpreted in a similar way.

56 To compare the performance of DDA and DEA in a controlled manner, we simulated scRNAseq data with
57 muscat[9] using the provided dataset as reference[10]. We generated scRNAseq data with varying number of
58 cells and 25% of differentially expressed genes. With these simulated data, DEA detected consistently
59 more genes, both the true positive rate (TPR) and false positive rate (FPR) were higher than the TPR and
60 FPR of DDA (Supplementary Fig. 1a, Supplementary Fig. 1b). DDA, however, had a higher precision (PPV,
61 Supplementary Fig. 1c) and was more accurate (F1 score, Fig. 1c).

62 Despite the observed association between mean expression and dropout rate, which has been
63 previously described[2], and similar performance of the two tests, there were 4,754 and 3,894 genes
64 uniquely identified using DDA and DEA, respectively, across all datasets. To better understand these
65 differences, we highlighted two differential dropout genes (DDGs). In the *cortex dataset*[11], *MTUS2* had
66 significantly less dropouts in excitatory neurons (logOR = 1.30, $P_{FDR|DDA} = 5.27 \times 10^{-115}$, Fig. 1d) compared
67 to inhibitory neurons, while the median expression levels were not significantly different (logFC = $-1.70 \times$
68 $10^{-3}$, $P_{FDR|DEA} = 5.70 \times 10^{-2}$), implying additional high ranked expressions for every additional zero. In the
69 *aging mouse atlas droplet* dataset[12], *Gnb2l1* had significantly less dropouts in the 3-month-old mice

70 (logOR = -0.67, $P_{FDR|DDA} = 1.81 \times 10^{-122}$, Fig. 1e) compared to the 24-month-old mice, while again the
71 median expression levels were not significantly different (logFC = $3.15 \times 10^{-3}$, $P_{FDR|DEA} = 6.97 \times 10^{-1}$).
72 These examples show that differences in variance between contrasting cell populations can interfere
73 with the association between dropout rate and mean expression, resulting in disparities between DDA
74 and DEA.

75 To exclude that the differentially behaving genes between DDA and DEA associate with a specific
76 technological or biological process, we investigated whether there were genes repeatedly detected by a
77 one of the two methods. In most cases, genes that were identified as DDG-only (or DEG-only) were
78 found within a single dataset (Supplementary Fig. 2a, Supplementary Fig. 2b), suggesting the absence of
79 a driving process for them.

80 To validate the DDA results, we compared the results of the *Alzheimer's Disease* (AD) dataset[13]
81 (entorhinal cortex) with DEA analysis performed on a bulk RNAseq AD dataset[14], comprised of samples
82 from the fusiform gyrus. For genes measured in both, the scRNAseq dataset and the bulk dataset (N =
83 2,228), the majority of DDG-only (60.7%) were also differentially expressed in bulk (Fig. 2a). The logOR
84 of the single cell analysis was also significantly correlated with the logFC of the bulk analysis ($\rho = 0.39$, P
85 $= 9.70 \times 10^{-79}$, Fig. 2b). Similarly, in a second dataset[15], 65.5% of the DDG-only genes were differentially
86 expressed in bulk samples from the frontal cortex, temporal cortex and hippocampal formation
87 (Supplementary Fig 3). The high correlation between the logOR and the logFC of a bulk dataset, despite
88 being from different brain region, and that the majority of DDG-only were still detected in bulk, further
89 confirms that DDA can be used to detect differentially abundant genes.

90 To test the binarization scheme, we performed a DDA on the AD dataset for binary profiles generated
91 with different thresholds for binarization (thresholds ranging from one to ten counts). Naturally, for
92 every increase in the threshold, the number of genes with zero measurements across all cells increased,

93 resulting in a decreasing number of tested and significant genes (Supplementary Fig 4a, Supplementary
94 Fig 4b). With higher thresholds, we found a decrease in correlation of the logORs from DDA with the
95 logFC from DEA (Supplementary Fig 4c). These results show that the default binarization scheme where
96 dropouts remain zero and every non-zero value is assigned a one, is indeed appropriate.

97 Finally, a runtime benchmark comparing DDA with all DE methods available in Seurat revealed that DDA
98 is among the most time efficient methods for detecting differentially abundant genes (Supplementary
99 Fig 5).

100 Altogether, our results show that dropout patterns across cell populations represent biological variation
101 and can be used as measure of relative abundance of transcripts. Across 16 datasets and a variety of
102 contrasting cell populations (disease vs healthy, cell-types, and cancer status), DDA detected biologically
103 relevant genes that were missed by DEA. While the performance of DDA and DEA on real data is largely
104 comparable, with a known ground truth, DDA performed better than DEA on simulated data.
105 Additionally, DDA benefits reproducibility and is more robust than DEA, since the only pre-processing
106 step required for DDA is the binarization of counts. In contrast, DEA requires normalization and
107 transformation of counts, where an analyst can choose from an excess of equally valid methods[16].
108 Performing DDA on datasets generated using the Smart-seq protocol should be approached with more
109 caution: although the agreement of detected genes between DDA and DEA was high, we observed the
110 lowest correlation between the logOR and logFC for these datasets.

111 DDA is a normalization-free, time efficient and accurate alternative to DEA for which we see three
112 potential use cases. First, DDA could be performed in isolation as a fast and accurate alternative to DEA.
113 Besides logistic regression, alternative tests such as the Fisher's exact test or the chi-squared test could
114 also be explored. However, logistic regression is more versatile as it allows to adjust for covariates.
115 Second, DDA could be performed in addition to DEA to identify more genes. Finally, DDA could be used

116 to validate pre-processing, normalization and DEA as a big discrepancy between the DDGs and DEGs
117 could indicate an aberration in the DEA results.

118 **Methods**

119 **Single cell RNA datasets**

120 In total, 16 scRNAseq dataset (14 human and 2 mouse) were used to investigate the utility and biological
121 relevance of binarized expression profiles of genes (Supplementary Table 1). All datasets had pre-
122 annotated cell types and conditions. From the corresponding references, un-normalized count matrices
123 were acquired, and only annotated cells were kept for further analysis. For each dataset, we extracted
124 the annotated cell-type, patient ID, and to which of contrasting cell population the cell belonged from
125 the included meta data. This was slightly different for the aging mouse atlases[12] and cancer atlas[8]. For
126 the aging mouse atlases, instead of annotated cell-types we retrieved the tissue names. For the cancer
127 atlas, the contrasting cell populations were defined by cell-type, so we retrieved the tissue and the
128 cancer-type for each cell. Each dataset was separately pre-processed. For the differential dropout
129 analysis (DDA), the count matrices were transformed to a binary representation, where dropouts remain
130 zero and every non-zero value was assigned a one. For the differential expression analysis (DEA), each
131 count matrix was log-normalized using Seurat 3.2.2[6], such that: $y_{ij} = \log\left(\frac{x_{ij}}{\sum_j x_{ij}} \times 10^4\right)$, where $x_{ij}$ and
132 $y_{ij}$ are the raw and normalized values for every gene $i$ in every cell $j$, respectively. This normalizes the
133 feature expression measurements for each cell by the total expression, multiplies this by a scale factor
134 (10,000 by default), and log-transforms the result. The cancer atlas was already normalized, as it was a
135 merger of multiple datasets.

136 **Statistical analysis**

137 Association p-values were corrected for multiple tests with the Benjamini-Hochberg procedure and
138 significance was assumed at an adjusted P-value of $P_{FDR} \leq 0.05$. Spearman's rank correlation coefficient
139 and the associated p-values were calculated using the `cor.test` function in R v4.0.2.

140 **Differential expression analysis**

141 DEA was performed using the Wilcoxon Rank Sum test using the FindMarkers function in Seurat 3.2.2[6].
142 Note that the Wilcoxon Rank Sum test from Seurat takes into account zero measurements, and handles
143 them as ties between contrasting cell populations. Genes coding for ribosomal proteins were excluded
144 and we only tested genes that were expressed in at least 10% of the cells in either of the respective
145 groups of interest. This is the default option in the FindMarkers function and speeds up testing by
146 ignoring infrequently expressed genes. Genes coding for ribosomal proteins were excluded. Association
147 p-values were corrected for multiple tests.

148 **Differential dropout analysis**

149 DDA was performed using logistic regression with the `glm()` function in R v4.0.2. All genes that were
150 tested with DEA were also tested with DDA. The resulting association p-values were corrected for
151 multiple tests.

152 **DDA – DEA comparison**

153 For the comparison between DDA and DEA, we investigated agreement and disagreement between
154 detected genes and the linear association between the logOR and logFC. Agreement was calculated by
155 the Jaccard index, i.e. number of genes that both tests commonly detected, divided by the total number
156 of genes that were detected. Agreement was calculated on the combination of all datasets and for each
157 individual dataset. The disagreement was investigated by means of inspecting characteristics of DDGs-
158 only and DEGs-only. DDGs-only were defined as genes that were detected ($P_{FDR} \leq 0.05$) by DDA and were

159 not detected ($P_{FDR} > 0.05$) by DEA. Conversely, DEGs-only were defined as genes that were detected ($P_{FDR}$
160 $\leq 0.05$) by DEA and were not detected ($P_{FDR} > 0.05$) by DDA. The Spearman's rank correlation coefficients
161 between the logOR and logFC were calculated with the estimates of all tested genes of the respective
162 datasets. The scale differences for every dataset, between logOR and logFC, were calculated with a
163 linear model on the estimates of all tested genes of the respective datasets, using the `lm` function in R
164 v4.0.2. The logOR was specified as outcome variable and the logFC as predictor variable. The resulting
165 slopes were interpreted as scale differences between the logOR and logFC

166 **Simulation**

167 Data was simulated with muscat 1.2.1[9]. The provided dataset[10] was used as reference. 100 simulated
168 datasets were generated with varying sample sizes (500 cells, 1,000 cells, 1,500 cells and 2,000 cells), 25
169 simulations per sample size. For each simulation 1,000 genes were generated of which 25% were
170 differently expressed between two groups of equal size. For both tests we calculated the True Positive
171 Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PVV) and accuracy (F1-score) per
172 simulation.

173 **Running times**

174 The runtime benchmark was performed with simulated data from muscat[9]. Runtimes of DDA and all DE
175 methods (wilcox, bimod, roc, t, negbinom, poisson, LR, MAST and DESeq2) included in Seurat were
176 measured based on 500 genes and 1,000 cells up to 10,000 cells with increments of 1,000. Runtime was
177 measured with the `system.time` function in R.

178 **Validation with existing bulk RNA-seq data**

179 The AD bulk RNA-seq datasets were acquired from Gemma[17]. The first dataset from Friedman, et al[14]
180 consisted of 33 controls (CT) and 84 samples from individuals diagnosed with Alzheimer's Disease (AD)

181 collected from the fusiform gyrus. This dataset was reprocessed by Gemma and no batch effects were
182 present. For the differential expression analysis in bulk we used the Wilcoxon Rank Sum test from limma
183 v3.44.3[18]. In total, 2,228 genes were tested for differential expression, as these genes were also included
184 in the scRNAseq AD dataset[13] analysis. The second dataset from Hokama, et al[15] consisted of 47 controls
185 and 32 AD samples. The samples originated from the frontal cortex ($N_{CT}$ = 18, $N_{AD}$ = 15), temporal cortex
186 ($N_{CT}$ = 19, $N_{AD}$ = 10) and hippocampal formation ($N_{CT}$ = 10, $N_{AD}$ = 7). The data was reprocessed and batch
187 corrected by Gemma. For the differential expression analysis no distinction was made between brain
188 regions. In total, 2,001 genes were tested for differential expression. All resulting association p-values
189 were corrected for multiple tests. For validation, the significant DDGs and DEGs from the scRNAseq AD
190 dataset analyses were compared with the significantly differentially expressed genes from the bulk
191 analyses. Venn diagrams were plotted with ggVennDiagram v0.3[19]. Correlations were calculated
192 between the logOR and logFC of the single-cell analysis with the bulk logFC.

193 **Availability of data and materials**

194 The datasets used and prepared for this study can be downloaded from Zenodo
195 (http://doi.org/10.5281/zenodo.4487320). The results are also made available in an interactive shiny
196 dashboard: http://insyprojects.ewi.tudelft.nl:5000/DifferentialDropoutAnalysis/. The scripts, functions
197 and source data for the figures are available at the Github repository:
198 https://github.com/gbouland/differential-dropout-analysis, including two vignettes describing a DDA
199 starting from a Seurat object and a raw count matrix. DDA is also implemented in an R-package and is
200 available at: https://github.com/gbouland/DDA.

201 **Acknowledgements**

202 This research was supported by an NWO Gravitation project: BRAINSCAPES: A Roadmap from
203 Neurogenetics to Neurobiology (NWO: 024.004.012)

204 **Author contributions**

205 GAB, AM, and MJTR conceived the study and designed the experiments. GAB performed all experiments
206 and drafted the manuscript. GAB, AM, and MJTR reviewed and approved the manuscript.

207 **Competing interests**

208 The authors declare no competing interests.

209 **References**


210 1. Choi, K., Chen, Y., Skelly, D. A. & Churchill, G. A. Bayesian model selection reveals biological
211 origins of zero inflation in single-cell transcriptomics. *Genome Biol.* **21**, 183 (2020).

212 2. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* vol. 38 147–150
213 (2020).

214 3. Andrews, T. S., Hemberg, M. & Birol, I. M3Drop: Dropout-based feature selection for scRNASeq.
215 *Bioinformatics* **35**, 2865–2867 (2019).

216 4. Li, R. & Quon, G. ScBFA: Modeling detection patterns to mitigate technical noise in large-scale
217 single-cell genomics data. *Genome Biol.* **20**, 193 (2019).

218 5. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1–9 (2020).

219 6. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

220 7. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential
221 expression analysis. *Nat. Methods* **15**, 255–261 (2018).

222 8. Nieto, P. *et al.* A Single-Cell Tumor Immune Atlas for Precision Oncology. *bioRxiv*
223 2020.10.26.354829 (2020) doi:10.1101/2020.10.26.354829.

224 9. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample
225 multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 1–12 (2020).

226 10. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation.
227 *Nat. Biotechnol.* **36**, 89–94 (2018).

228 11. Agarwal, D. *et al.* A single-cell atlas of the human substantia nigra reveals cell-specific pathways
229 associated with neurological disorders. *Nat. Commun.* **11**, 1–11 (2020).

230 12. Almanzar, N. *et al.* A single-cell transcriptomic atlas characterizes ageing tissues in the mouse.
231 *Nature* **583**, 590–595 (2020).

232 13. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's
233 disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097
234 (2019).

235 14. Friedman, B. A. *et al.* Diverse Brain Myeloid Expression Profiles Reveal Distinct Microglial
236 Activation States and Aspects of Alzheimer's Disease Not Evident in Mouse Models. *Cell Rep.* **22**,
237 832–847 (2018).

238 15. Hokama, M. *et al.* Altered expression of diabetes-related genes in Alzheimer's disease brains: The
239 Hisayama study. *Cereb. Cortex* **24**, 2476–2488 (2014).

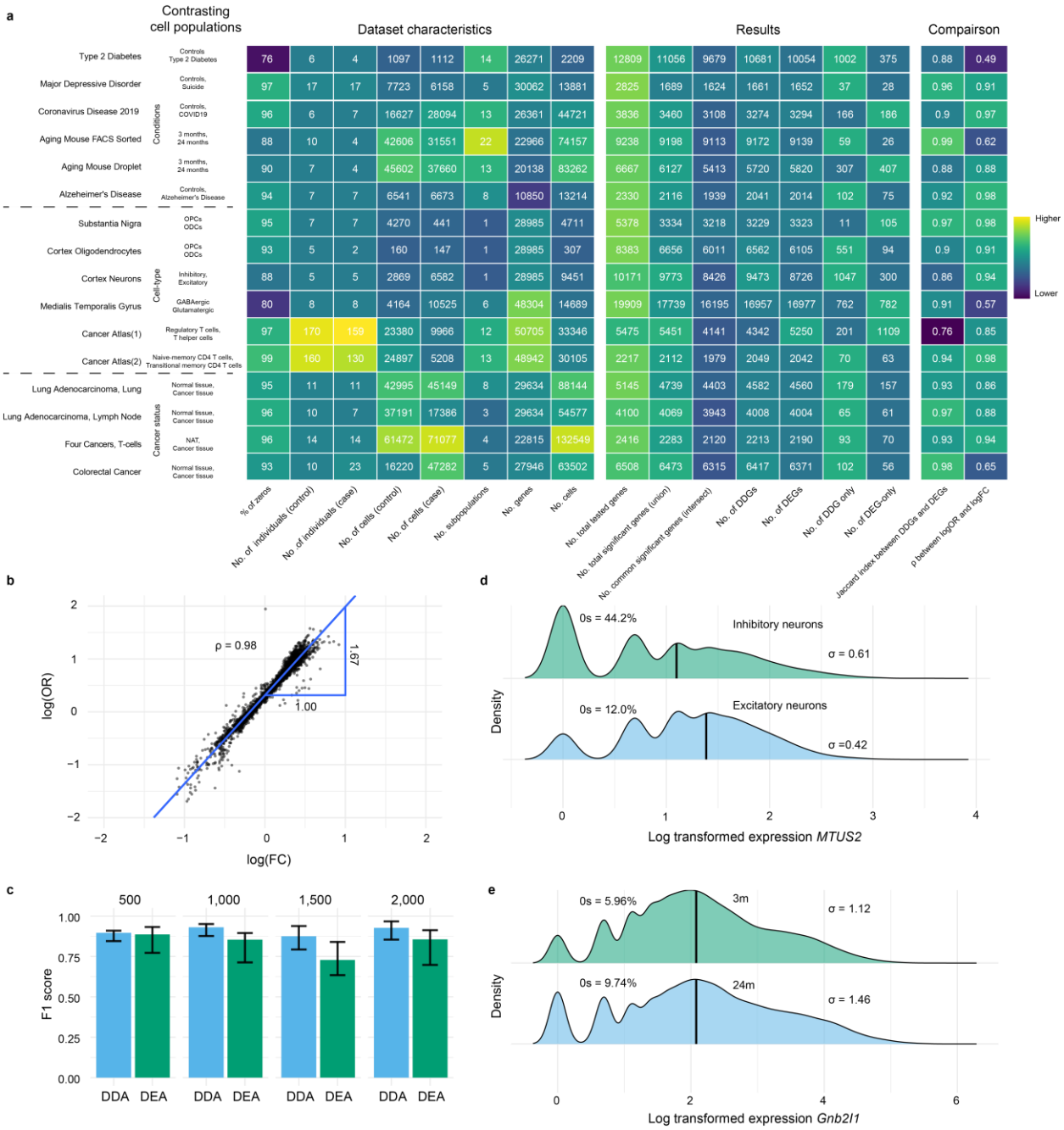
240 16. Lytal, N., Ran, D. & An, L. Normalization Methods on Single-Cell RNA-seq Data: An Empirical
241 Survey. *Front. Genet.* **11**, 41 (2020).

242 17. Zoubarev, A. *et al.* Gemma: a resource for the reuse, sharing and meta-analysis of expression
243 profiling data. *Bioinformatics* **28**, 2272–2273 (2012).

244 18. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and

245 microarray studies. *Nucleic Acids Res.* **43**, (2015).

246 19. Gao, C.-H. ggVennDiagram: A 'ggplot2' Implement of Venn Diagram. (2019).

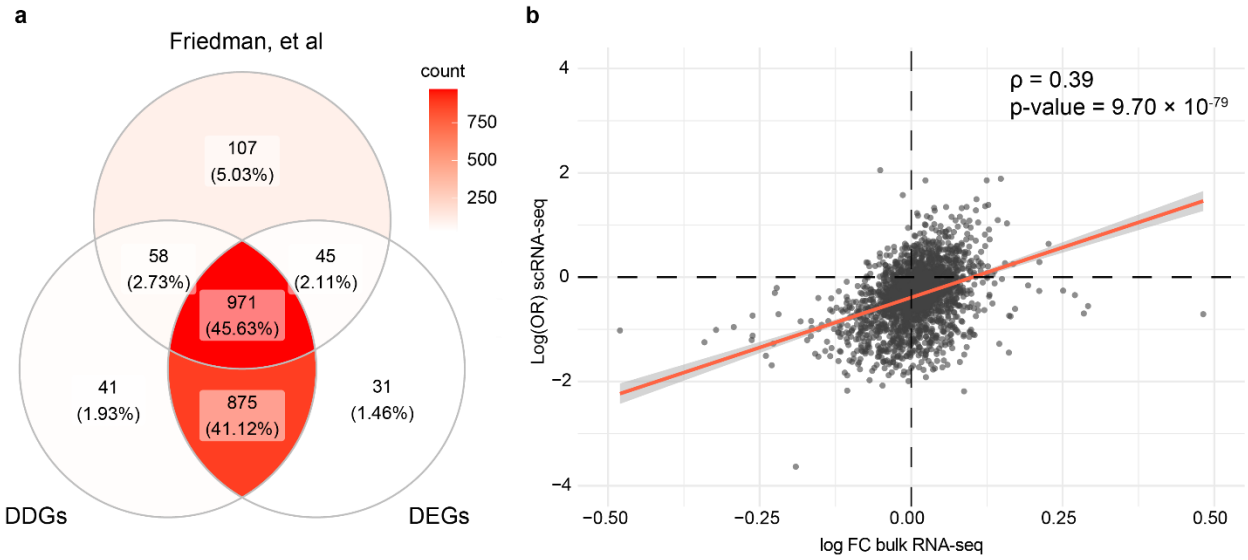



247

248 *Figure 1* ***a,*** Heatmap of the dataset characteristics, general overview of the results of both DDA and DEA

249 per dataset, and a comparison of the results. The rows represent the datasets, the first column shows

250 the cell populations that were used as contrast for testing. **b,** Plot of the logOR and logFC of the Cancer
251 Atlas (2) dataset. The x-axis represent the logFCs of each tested gene, and the y-axis represent the
252 logORs for the same genes. The blue lines shows the linear association between the logFC and logOR.
253 The Spearman's rank correlation coefficient ($\rho$) is also shown in the plot. **c** Barplots of the F-score of DDA
254 based on simulated data. Numbers above barplots shows the number of cells that were generated for
255 the repsective simulations simulation. Height of bar is defined the median value from 25 simulations,
256 error bars are defined by the first and thrid quartile. **d** Two density plots of *MTUS2* from cortex neuron
257 dataset. The top plot shows the density of *MTUS2* in inhibitory neurons and the bottom plot shows the
258 density of *MTUS2* in excitatory neurons. **e** Two density plots of *Gnb2I1* from the aging mouse atlas
259 droplet dataset. The top plot shows the density of *Gnb2I1* in 3-month-old mice and the bottom plot
260 shows the density of *Gnb2I1 in* 24-month-old mice. Both **d** and **e** are supported with fraction of zeros
261 and variance of each cell population.



262

263 Figure 2 **a,** Venn diagram of genes detected ($P_{FDR} \leq 0.05$) in a bulk AD dataset (Friedman, et al), in the
264 single cell AD dataset with DDA (DDGs) and with DEA (DDGs). Each section shows the number and
265 percentage of genes belonging to that section. **b** Plot of the logFC from the AD bulk dataset (x-axis) and

266 the logOR from the single cell AD dataset (y-axis). The red line represents the linear association between

267 the bulk logFC and logOR**.** The Spearman's rank correlation coefficient (ρ) and corresponding association

268 p-value are also shown. Outlier genes (n = 9, **b**) were removed from the plots.