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

Gerard A. Bouland[1,2], Ahmed Mahfouz[1,2,3,*] & Marcel J.T. Reinders[1,2,3,*]

[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 author: Ahmed Mahfouz (a.mahfouz@lumc.nl) and Marcel J.T. Reinders (m.j.t.reinders@tudelft.nl)

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 rather than technical artifacts. We propose differential dropout analysis (DDA), as an alternative to differential expression analysis (DEA), to identify the effects of biological variation in single-cell RNA sequencing data. Using 16 publicly available datasets, we show that dropout patterns are biological in nature and can assess the relative abundance of transcripts more robustly than counts.

Main

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

As proof of concept, we performed DDA with a simple logistic regression on binarized expression profiles from 16 scRNAseq datasets (662,825 cells in total, Supplementary Table 1). We compared the results of DDA with those of differently expressed genes (DEGs) detected using the commonly used Wilcoxon Rank Sum test[6,7]. We tested each gene using both DDA and DEA for differences between conditions (6 datasets), cell-types (6 datasets) and normal- versus cancerous tissues (4 datasets, Fig. 1a). Across all datasets, a total of 98,013 significant genes ($P_{FDR} \leq 0.05$) were identified with either DDA (93,856 genes) or DEA (92,975 genes). Of these, 88,818 were identified by both tests, resulting in a Jaccard index of 0.91. This high degree of agreement is also reflected in each individual dataset (median

= 0.92, minimum = 0.76, and maximum = 0.99). We did not use a log fold-change (logFC) or log odds-ratio (logOR) threshold, as for each dataset and comparison different thresholds are appropriate. In all datasets, the logFC and logOR were significantly correlated (median(R) = 0.78, minimum(R) = 0.17, and maximum(R) = 0.98, $P \leq 7.75 \times 10^{-84}$). The three datasets with the lowest correlation coefficient between logOR and logFC ($R \leq 0.53$) were datasets generated using the Smart-seq protocol (Supplementary Table 1). Across the datasets, we observed an average increase of 1.79 in logOR (median = 1.70, Q_1 = 1.56, Q_3 = 2.10), for every increase in logFC. As an example, the correlation between the logFC and logOR in the *cancer atlas (2)* dataset[8] was r=0.98 and for every increase in logFC, the logOR increased 1.67 (Fig. 1b). Given that observed power and p-values are directly associated[9] and DDA detected more genes than DEA at identical p-value thresholds and population sizes, we argue that DDA has more statistical power than DEA to detect differentially abundant genes.

Despite the observed association between mean expression and dropout rate, which has been previously described[2], there were 5,038 and 4,157 genes uniquely identified using DDA and DEA, respectively, across all datasets. To better understand these differences, we highlighted two differential dropout genes (DDGs) from the *aging mouse atlas droplet* dataset[10] where discrepancies between the results of DDA and DEA were the largest. *Rpl5* had significantly less dropouts in the 3-month-old mice (logOR = -0.81, $P_{FDR|DDA} = 3.76 \times 10^{-158}$, Fig. 1c) compared to the 24-month-old mice, while the expression levels were not significantly different (logFC = -5.53×10^{-3}, $P_{FDR|DEA} = 5.01 \times 10^{-1}$). Also, *Rps15* had significantly less dropouts in the 3-month-old mice (logOR = -1.64, $P_{FDR|DDA} = 1.62 \times 10^{-123}$, Fig. 1d) compared to the 24-month-old mice, while the expression levels were not significantly different (logFC = -7.16×10^{-4}, $P_{FDR|DEA} = 2.44 \times 10^{-1}$). It should be noted that the Wilcoxon Rank Sum test from Seurat takes into account zero measurements, and handles them as ties between contrasting cell populations. However, in both cases, for every additional zero measurement in the 24-month-old mice compared to the 3-month-old mice, a high rank measurement is present in the 24-month-old mice, which is reflected

in the higher variability of *Rpl5* and *Rps15* in the 24-month-old mice. These results show that differences in variance between contrasting cell populations can interfere with the association between dropout rate and mean expression, resulting in disparities between DDA and DEA.

To exclude that the differentially behaving genes between DDA and DEA associate with a fixed biological process, we investigated whether there were recurring genes across datasets for either of the two methods. In most cases, genes that were identified as DDG-only or DEG-only were found within a single dataset (Supplementary Fig. 1a, Supplementary Fig. 1b). Next, we took the union of all DDG-only genes across all datasets and performed a GO term enrichment analysis. We found 16 GO terms to be enriched ($P_{FDR} \leq 0.05$, Supplementary Table 2), including mRNA catabolic process ($N_{genes} = 131$, $P_{FDR} = 8.34 \times 10^{-18}$) and RNA splicing ($N_{genes} = 149$, $P_{FDR} = 1.70 \times 10^{-18}$). For the union of DEG-only genes, 37 GO terms were enriched (Supplementary Table 3), including regulation of mRNA metabolic process ($N_{genes} = 97$, $P_{FDR} = 1.81 \times 10^{-12}$) and RNA splicing ($N_{genes} = 119$, $P_{FDR} = 6.06 \times 10^{-10}$). The observation that most DDG-only and DEG-only were found with a single dataset and that few generic GO terms were enriched, suggests that the behaviour of these genes is not associated with a specific biological process.

To validate the DDA and DEA results, we compared the results for the *Alzheimer's Disease* (AD) dataset[11] (entorhinal cortex) with DEA analysis performed on a bulk RNAseq AD dataset[12], comprised of samples from the fusiform gyrus. For genes measured in both the scRNAseq dataset and the bulk dataset (N = 2,228), the majority of DDG-only (60.7%) and DEG-only (57.9%) were also differentially expressed in bulk (Fig. 2a). Both, the logOR and logFC of the single cell analysis were significantly correlated with the logFC of the bulk analysis ($r \geq 0.22$, $P \leq 1.27 \times 10^{-25}$, Fig. 2b, Fig. 2c). Similarly, in a second dataset[13], 65.5% of the DDG-only and 65.2% of the DEG-only genes were differentially expressed in bulk samples from the frontal cortex, temporal cortex and hippocampal formation (Supplementary Fig 2). Given that the logOR and logFC are correlated with the logFC of a bulk dataset from a different brain region, and that the

majority of DDG-only and DEG-only were also detected in bulk, further suggests that these genes are associated with AD.

To test the binarization scheme, we performed a DDA on the AD dataset for binary profiles generated with different thresholds for binarization (thresholds ranging from one to ten counts). Naturally, for every increase in the threshold, the number of genes with zero measurements across all cells increased, resulting in a decreasing number of tested and significant genes (Supplementary Fig 3a, Supplementary Fig 3b). With higher thresholds, we found a decrease in correlation of the logORs from DDA with the logFC from DEA (Supplementary Fig 3c). These results show that the default binarization scheme where dropouts remain zero and every non-zero value is assigned a one, is indeed appropriate.

Altogether, our results show that dropout patterns across cell populations are biological in nature and can be used as measure of relative abundance of transcripts. Across 16 datasets and a variety of contrasting cell populations (disease vs healthy, cell-types, and cancer status), DDA detected biologically relevant genes that were missed by DEA. Moreover, with identical number of tested genes, p-value thresholds and cell population sizes, DDA detected more genes than DEA, suggesting that DDA is statistically more powerful than DEA. While the performance of DDA and DEA is largely comparable, DDA benefits reproducibility and is more robust than DEA, since the only pre-processing step required for DDA is the binarization of counts. In contrast, DEA requires normalization and transformation of counts, where an analyst can choose from an excess of equally valid methods[14]. Performing DDA on datasets generated using the Smart-seq protocol should be approached with caution, as for these datasets we identified large differences between the DDA and DEA.

We do see three potential use cases for differential dropout analysis. First, DDA could be performed in isolation as a fast and accurate alternative to DEA. Besides logistic regression, alternative tests such as the Fisher's exact test or the chi-squared test can also be used. However, logistic regression is more

versatile as it allows to adjust for covariates. Second, DDA could be performed in addition to DEA to identify more genes. When only DDA is considered, it is advisable to calculate fold-changes in addition to odd-ratios. Although they are highly correlated, odd-ratios cannot be interpreted as fold-changes since there is no one-to-one correspondence. Finally, DDA could be used to validate pre-processing, normalization and DEA as a big discrepancy between the DDGs and DEGs could indicate an aberration in the DEA results.

Methods

Single cell RNA datasets

In total, 16 scRNAseq dataset (14 human and 2 mouse) were used to investigate the utility and biological relevance of binarized expression profiles of genes (Supplementary Table 1). All datasets had pre-annotated cell types and conditions. From the corresponding references, un-normalized count matrices were acquired, and only annotated cells were kept for further analysis. For each dataset, we extracted the annotated cell-type, patient ID, and to which of contrasting cell population the cell belonged from the included meta data. This was slightly different for the aging mouse atlases[10] and cancer atlas[8]. For the aging mouse atlases, instead of annotated cell-types we retrieved the tissue names. For the cancer atlas, the contrasting cell populations were defined by cell-type, so we retrieved the tissue and the cancer-type for each cell. Each dataset was separately pre-processed. For the differential dropout analysis (DDA), the count matrices were transformed to a binary representation, where dropouts remain zero and every non-zero value was assigned a one. For the differential expression analysis (DEA), each 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 y_{ij} are the raw and normalized values for every gene i in every cell j, respectively. This normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor

(10,000 by default), and log-transforms the result. The cancer atlas was already normalized, as it was a merger of multiple datasets.

Statistical analysis

Association p-values were corrected for multiple tests with the Benjamini-Hochberg procedure and significance was assumed at a False Discovery Rate (FDR) of $P_{FDR} \leq 0.05$. Pearson's correlation coefficients and the associated p-values were calculated using the `cor.test` function in R v4.0.2.

Differential expression analysis

DEA was performed using the Wilcoxon Rank Sum test using the FindMarkers function in Seurat 3.2.2[6]. We only tested genes that were expressed in at least 10% of the cells in either of the respective groups of interest. This is the default option in the FindMarkers function and speeds up testing by ignoring infrequently expressed genes. We utilized no threshold for the log-fold change, as for each dataset and comparison different thresholds are appropriate. DEA was performed per dataset between the specified contrasting cell populations (See Supplementary Table 1). Association p-values were corrected for multiple tests.

Differential dropout analysis

DDA was performed using logistic regression with the `glm()` function in R v4.0.2. All genes that were tested with DEA were also tested with DDA. In concordance with the DEA, we use no regression coefficient threshold. The resulting association p-values were corrected for multiple tests. Two vignettes have been made and are deposited at https://github.com/gbouland/differential-dropout-analysis describing a DDA starting from a Seurat object and a raw count matrix.

DDA – DEA comparison

For the comparison between DDA and DEA, we investigated agreement and disagreement between detected genes and the linear association between the logOR and logFC. Agreement was calculated by the number of genes that both tests commonly detected, divided by the total number of genes that were detected. In other words, the intersect divided by the union, otherwise known as the Jaccard index. Agreement was calculated on the combination of all datasets and for each individual dataset. The disagreement was investigated by means of inspecting characteristics of DDGs-only and DEGs-only. DDGs-only were defined as genes that were detected ($P_{FDR} \leq 0.05$) by DDA and were not detected ($P_{FDR} > 0.05$) by DEA. Conversely, DEGs-only were defined as genes that were detected ($P_{FDR} \leq 0.05$) by DEA and were not detected ($P_{FDR} > 0.05$) by DDA. The Pearson's correlation coefficient between the logOR and logFC were calculated with the estimates of all tested genes of the respective datasets. The scale differences for every dataset, between logOR and logFC, were calculated with a linear model on the estimates of all tested genes of the respective datasets, using the `lm` function in R v4.0.2. The logOR was specified as outcome variable and the logFC as predictor variable. The resulting slopes were interpreted as scale differences between the logOR and logFC

Recurring genes and GO term enrichment

Recurring genes were defined as genes that were repeatedly detected as DDG- or DEG-only with the included datasets. For this we excluded the aging mouse atlases, as genes detected with these datasets, naturally will not be detected in the human datasets. Also, GO term enrichments were performed on the union of all DDG-only and DEG-only from all datasets, excluding the aging mouse atlases. GO term enrichments were performed with clusterProfiler v3.16.1[15]. The gene symbols were converted to entrez IDs based on org.Hs.eg.db v3.11.4[16]. The function `enrichGO` was used with biological process (BP) as selected sub-ontology. The association p-values were corrected for multiple tests. Processes that did not

contain a gene from all datasets were removed, as such, the remaining terms represented shared enriched process that each dataset at least contributed one gene to.

Validation with existing bulk RNA-seq data

The AD bulk RNA-seq datasets were acquired from Gemma[17]. The first dataset from Friedman, et al[12] consisted of 33 controls (CT) and 84 samples from individuals diagnosed with Alzheimer's Disease (AD) collected from the fusiform gyrus. This dataset was reprocessed by Gemma and no batch effects were present. For the differential expression analysis in bulk we used the Wilcoxon Rank Sum test from limma v3.44.3[18]. In total, 2,228 genes were tested for differential expression, as these genes were also included in the scRNAseq AD dataset[11] analysis. The second dataset from Hokama, et al[13] consisted of 47 controls and 32 AD samples. The samples originated from the frontal cortex (N_{CT} = 18, N_{AD} = 15), temporal cortex (N_{CT} = 19, N_{AD} = 10) and hippocampal formation (N_{CT} = 10, N_{AD} = 7). The data was reprocessed and batch corrected by Gemma. For the differential expression analysis no distinction was made between brain regions. In total, 2,001 genes were tested for differential expression. All resulting association p-values were corrected for multiple tests. For validation, the significant DDGs and DEGs from the scRNAseq AD dataset analyses were compared with the significantly differentially expressed genes from the bulk analyses. Venn diagrams were plotted with ggVennDiagram v0.3[19]. Correlations were calculated between the logOR and logFC of the single-cell analysis with the bulk logFC.

Changing binary threshold

To investigate the effect of different thresholds for the binarization scheme on the results, we performed DDA on ten variants of the scRNAseq AD dataset. These ten variants were generated by increasing the threshold for assigning a zero or a one, with increments of one. The first variant of the dataset, which was also the default used for this study, was generated with a threshold of one, where zero remains zero and every non-zero value was assigned a one. The last variant was generated with a

threshold of ten, where every measurement in which the count was below ten was assigned a zero and every measurement with counts equal or above ten was assigned a one. To evaluate the results, we investigated the number of tested and significant genes, per dataset variant. Additionally, the logORs from each dataset variant was tested on correlation with the logFC from the DEA.

Availability of data and materials

The datasets used and prepared for this study can be downloaded from Zenodo (http://doi.org/10.5281/zenodo.4487320). The results are also made available in an interactive shiny dashboard: http://insyprojects.ewi.tudelft.nl:5000/DifferentialDropoutAnalysis/. The scripts, functions and source data for the figures are available at the Github repository: https://github.com/gbouland/differential-dropout-analysis.

Acknowledgements

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

Author contributions

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

Competing interests

The authors declare no competing interests.

References

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

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

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

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

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

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

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

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

9. Hoenig, J. M. & Heisey, D. M. The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am. Stat.* **55**, 19–24 (2001).

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

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

12. Friedman, B. A. *et al.* Diverse Brain Myeloid Expression Profiles Reveal Distinct Microglial Activation States and Aspects of Alzheimer's Disease Not Evident in Mouse Models. *Cell Rep.* **22**,

832–847 (2018).

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

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

15. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

16. Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. (2020).

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

18. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, (2015).

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

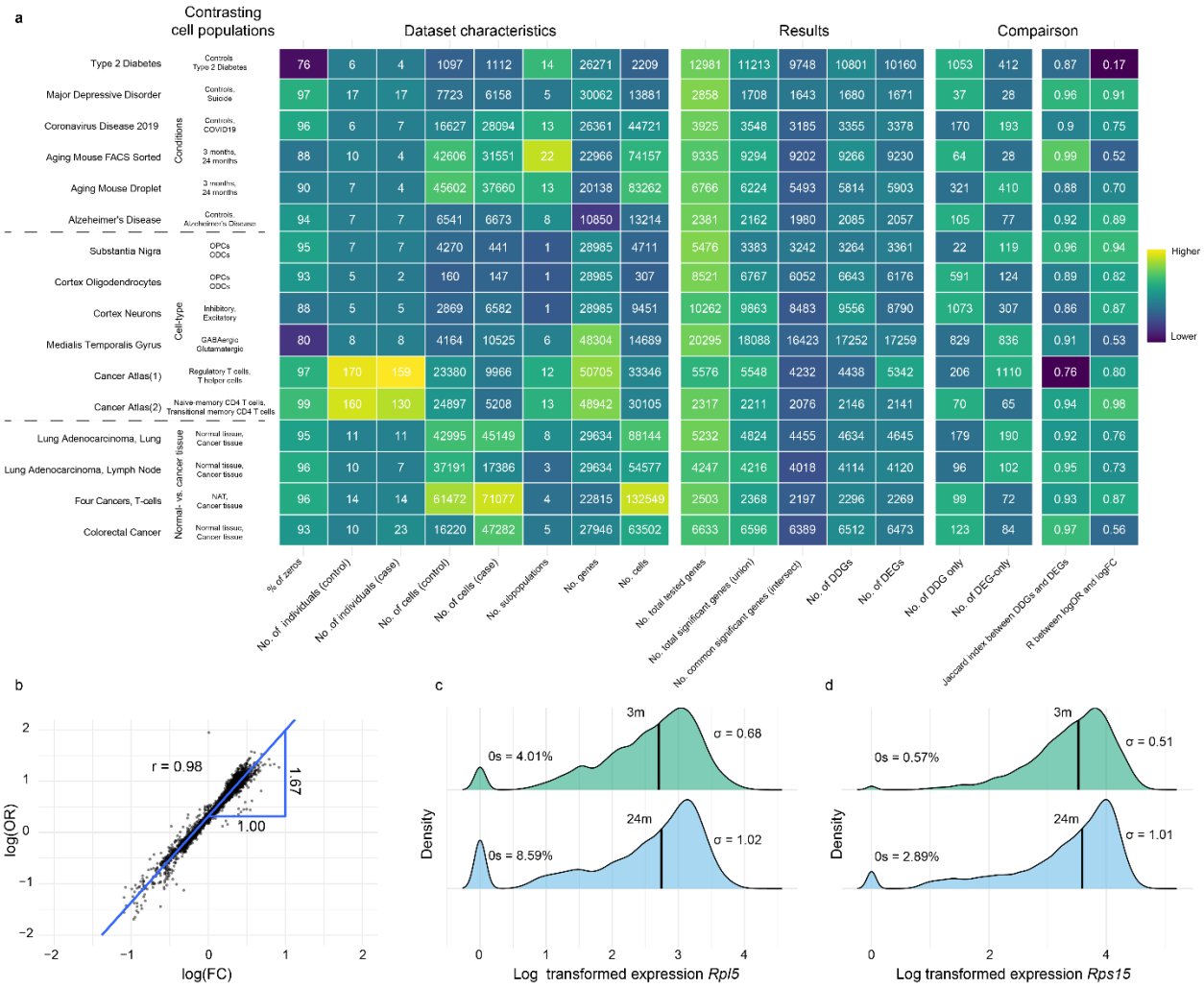

Figure 1 ***a,*** Heatmap of the dataset characteristics, general overview of the results of both DDA and DEA per dataset, and a comparison of the results. The rows represent the datasets, the first column shows the cell populations that were used as contrast for testing. **b,** Plot of the logOR and logFC of the Cancer Atlas (2) dataset. The x-axis represent the logFCs of each tested gene, and the y-axis represent the logORs for the same genes. The blue lines shows the linear association between the logFC and logOR. The pearson's correlation coefficient (r) is also shown in the plot. **c,d** Two density plots of *Rpl5* (**c**) and *Rps15* (**d**) from the aging mouse atlas droplet dataset. The top plot shows the density of *Rpl5/Rps15* in 3-month-old mice and the bottom plot shows the density of *Rpl5/Rps15* in the 24-month-old mice. The x-axis represents the log transformed expression of *Rpl5/Rps15* in the respective cell populations. The plot is supported with fraction of zeros and variance of each cell population.

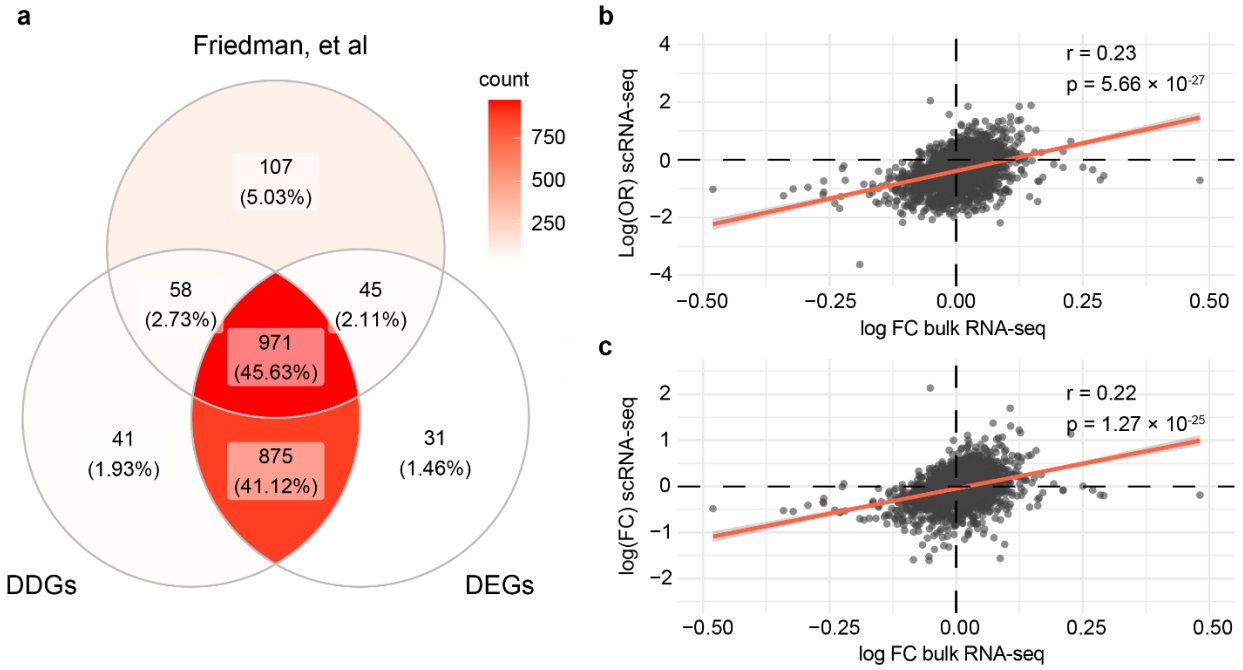

Figure 2 **a,** Venn diagram of genes detected ($P_{FDR} \leq 0.05$) in a bulk AD dataset (Friedman, et al), in the single cell AD dataset with DDA (DDGs) and with DEA (DDGs). Each section shows the number and percentage of genes belonging to that section. **b,c** Plot of the logFC from the AD bulk dataset (x-axis) and the logOR (**b**) and logFC (**c**) from the single cell AD dataset (y-axis). The red line represents the linear association between the bulk logFC and single cell logOR**.** The Pearson's correlation coefficient (r) and corresponding association p-value are also shown. Outlier genes (n = 9, **b**) and (n = 9, **c**) were removed from the plots.