Analysis of CITE sequencing data of human brain cells

Anastasiia Popova

Outline

Outline

Introduction

motivation

Understanding which specific proteins are responsible for the development of autoimmune diseases can aid in early diagnosis and treatment.

One of the most progressive research methods for this today is **single-cell RNA sequencing** (scRNA-seq), which provides a detailed and accurate picture of gene expression.

goal

In this project, I will analyze single-cell sequencing data from the human brain organoid. *Since the cells were also bound with antibodies from patients with autoimmune diseases, I hope to discover a correlation between gene expression and binding.*

method

I will use the **augmented transcriptome mapping** of RNA sequencing data, a fast and highly resolved method for analyzing gene expression [1].

Because RNA-seq data analysis mainly relies on packages for R language, here I also aim to entirely **rely on Python**. The main reason for this is the belief that big data processing can be more effective using Python. It is also a more flexible and fastly developing tool, using which also might be crucial for cutting-edge data analysis.

[1] Workflow is mainly based on recommendations in https://www.sc-best-practices.org/ book.

Introduction

CITE-seq protocol

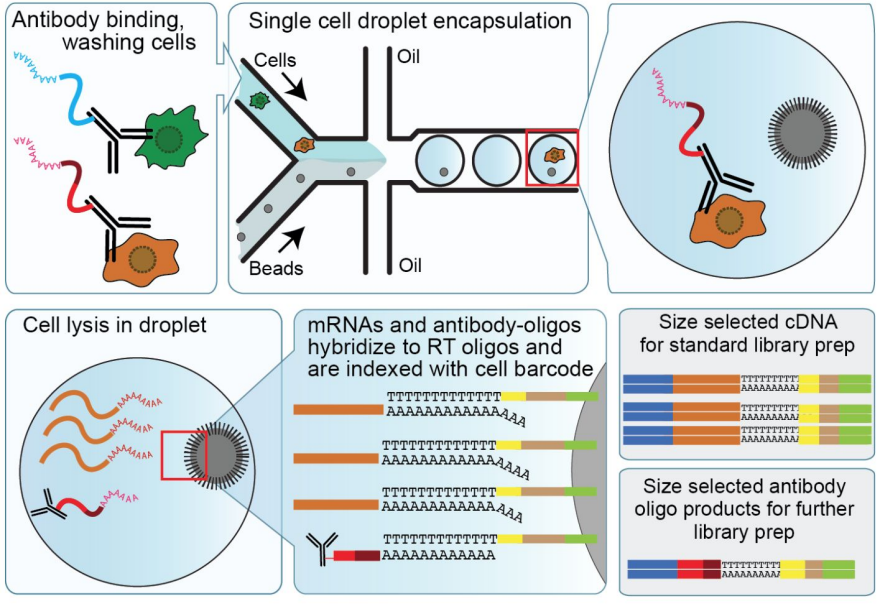
The CITE-seq protocol integrates cellular protein and single-cell RNA (scRNA) measurements into a single assay using oligonucleotide-labeled antibodies.

design of experiment

Cerebral (brain) organoid cells from four donors were bound to four types of antibodies.

data processing approach

In general, scRNA and antibody/cell-hashtag data are processed separately. Here they will only be merged at the dimensionality reduction stage.



picture is from https://cite-seq.com/

Outline

Raw Data Processing

steps of raw scRNA data processing

1. Generation so-called the **splici index** for the reference genome (a.k.a. augmented transcriptome: build spliced transcripts + introns using pyroe [1] and index it using salmon [2])

2. Generation a **permit list** for cell barcode correction 10x Chromium v2 chemistry

3. **Mapping** of the sequencing reads against an index of the reference (alevin-fry [3])

4. **UMI resolution**: allocation a molecular count to each gene in each cell. UMIs with equal matches to multiple genes are treated as a group (a.k.a. *equivalence class*), using an expectation maximization algorithm (alevin-fry [3]).

[1]https://pyroe.readthedocs.io/en/latest/

[2]https://combine-lab.github.io/salmon/

[3]https://alevin-fry.readthedocs.io/en/latest/

Raw Data Processing

steps of raw antibodies data processing

1. Generation of the index the feature barcodes according the barcode sequences of the antibody-derived tags (ADT) and the hash antibody tag oligos (HTO)[2]. **Further here is referred as "antibodies data"**

2. Generation a **permit list** for cell barcode correction 10x Chromium v2 chemistry

3. **Mapping** of the sequencing reads against the index [3]

4. **UMI resolution**. UMIs with equal matches to multiple genes are treated as an equivalence class (alevin-fry [3]).

[1]https://pyroe.readthedocs.io/en/latest/

[2]https://combine-lab.github.io/salmon/

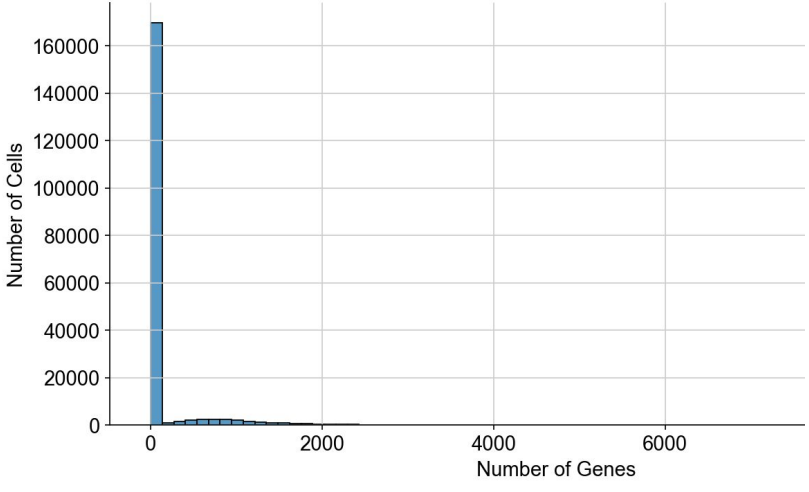
[3]https://alevin-fry.readthedocs.io/en/latest/

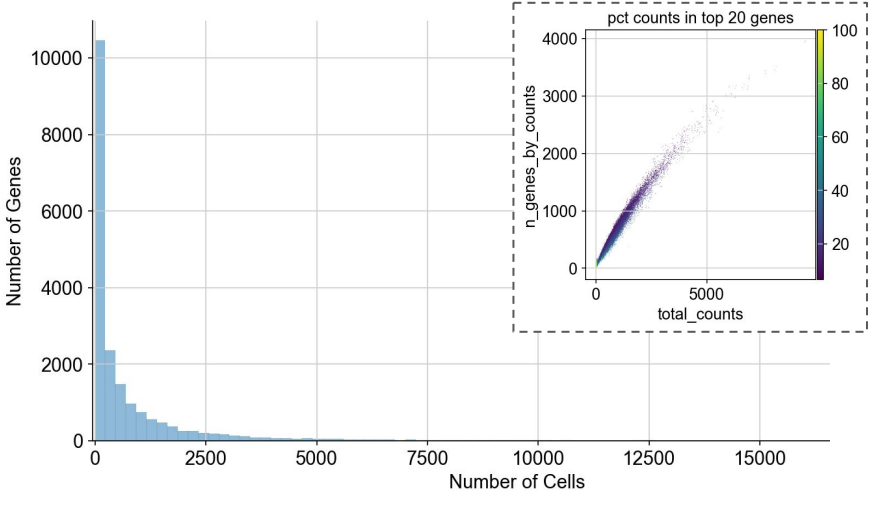
Outline

Exploratory RNA Data Analysis

After raw data processing, the **scRNA dataset** consists of **190,954 cells and 62,700 genes**.

To work only with protein-coding genes, preselection of genes was done here. The list of genes is from **Ensembl BioMart database**. Number of genes on this stage is **20,054**.



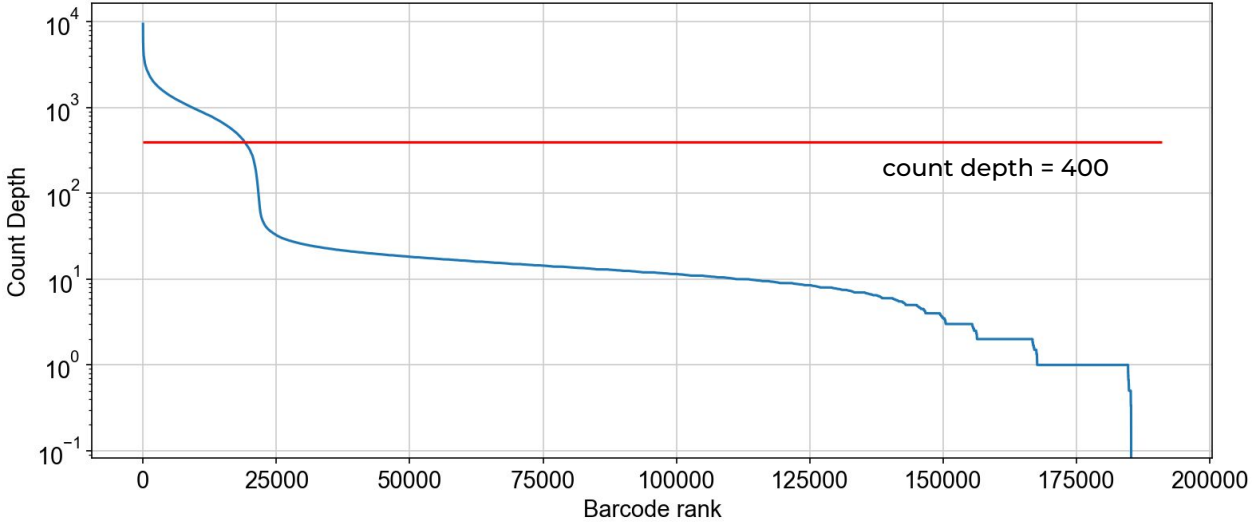


Here and further scanpy and anndata Python libraries were used.

Outline

Ambient RNA Correction

Droplet-based scRNA-seq protocol assumes that UMI enables to identify of the number of molecules for each gene and each cell. But cell-free mRNA molecules can be present in the dilution (people refer to this as a "soup"), which are also sequenced.



Number of cells after ambient RNA correction: **19,026**

We also will filter out cells for which less than **1000** genes were found.

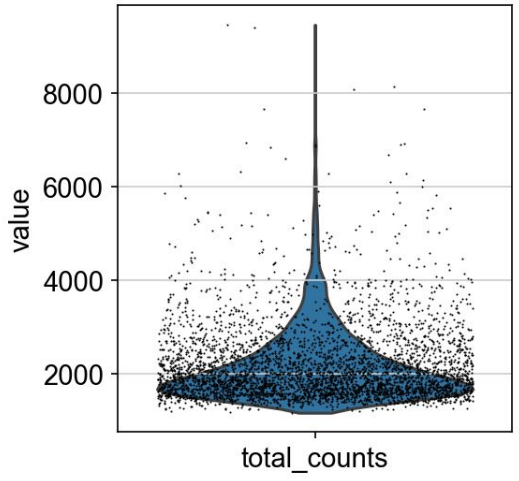
Number of cells after filtration of low-quality cells: **4,326**

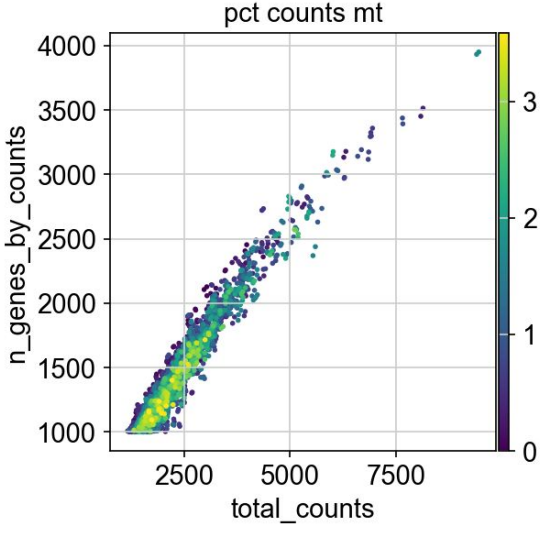
Damaged Cells Filtration

To filter out cells with broken membranes, which usually have a high amount of mitochondrial counts, few detected genes and a low count depth (the number of counts per barcode). To ensure accurate representation of informative mRNA transcripts, it is necessary to exclude cells with high ribosomal RNA reads from further analysis.

A quality control threshold: outliers are data points that differ by **5 median absolute deviations**. Also, cells with a percentage of mitochondrial or ribosomal counts exceeding **8** % are filtered out.

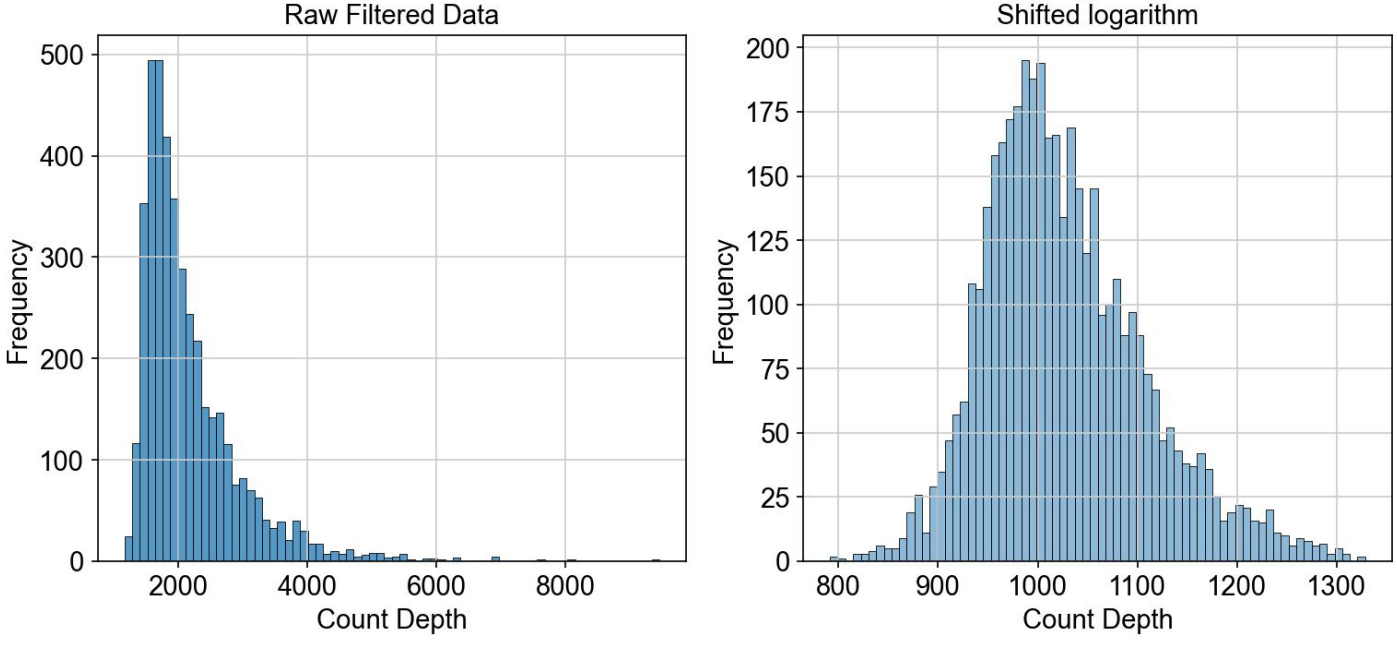
- Number of mitochondrial genes: 13;
- Number of ribosomal genes: 102
- Number of cells after filtering of low quality cells: **4,205**





Normalization of RNA Data

Shifted logarithm normalization aims to make the variances across the dataset more similar. It used for subsequent dimensionality reduction and identification of differentially expressed genes.



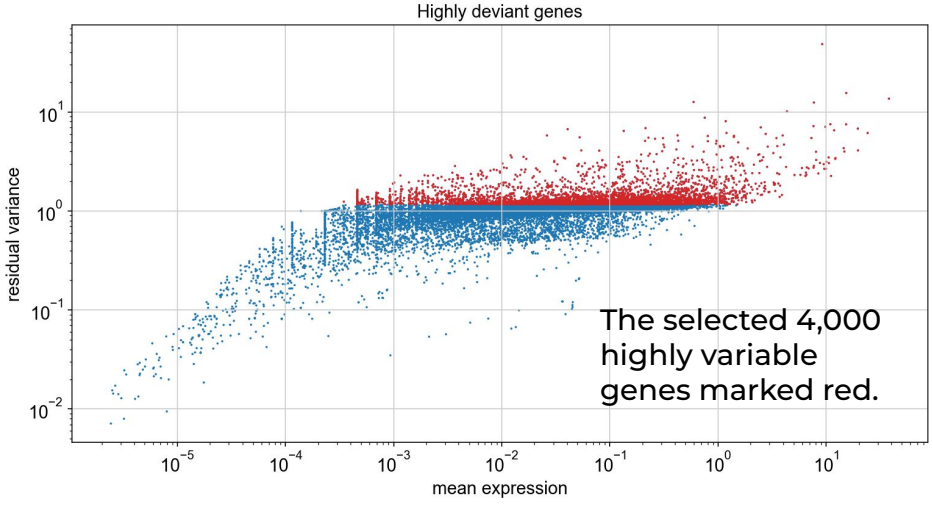
Gene Selection

Some number of genes might be detected in a few cells (the count matrix is sparse), but usually, it is interesting to consider genes with high spread of expression values relative to the mean.

Analytic Pearson residuals help detect how much each gene deviates from the constant-expression model [1].

```
Top-20 highly variable genes
----------------------------------------

gene_name
H2BC4        0.0
ERBB4        1.0
CCSER1       2.0
H2AC13       3.0
COPG2        4.0
TJP1         5.0
NKAIN2       6.0
CNTNAP2      7.0
DLGAP1       8.0
MT-CO1       9.0
NXPH1       10.0
BLCAP       11.0
NEGR1       12.0
VIM         13.0
DGKI        14.0
SH3TC1      15.0
LSAMP       16.0
KCND2       17.0
IL6ST       18.0
TRABD2A     19.0
```



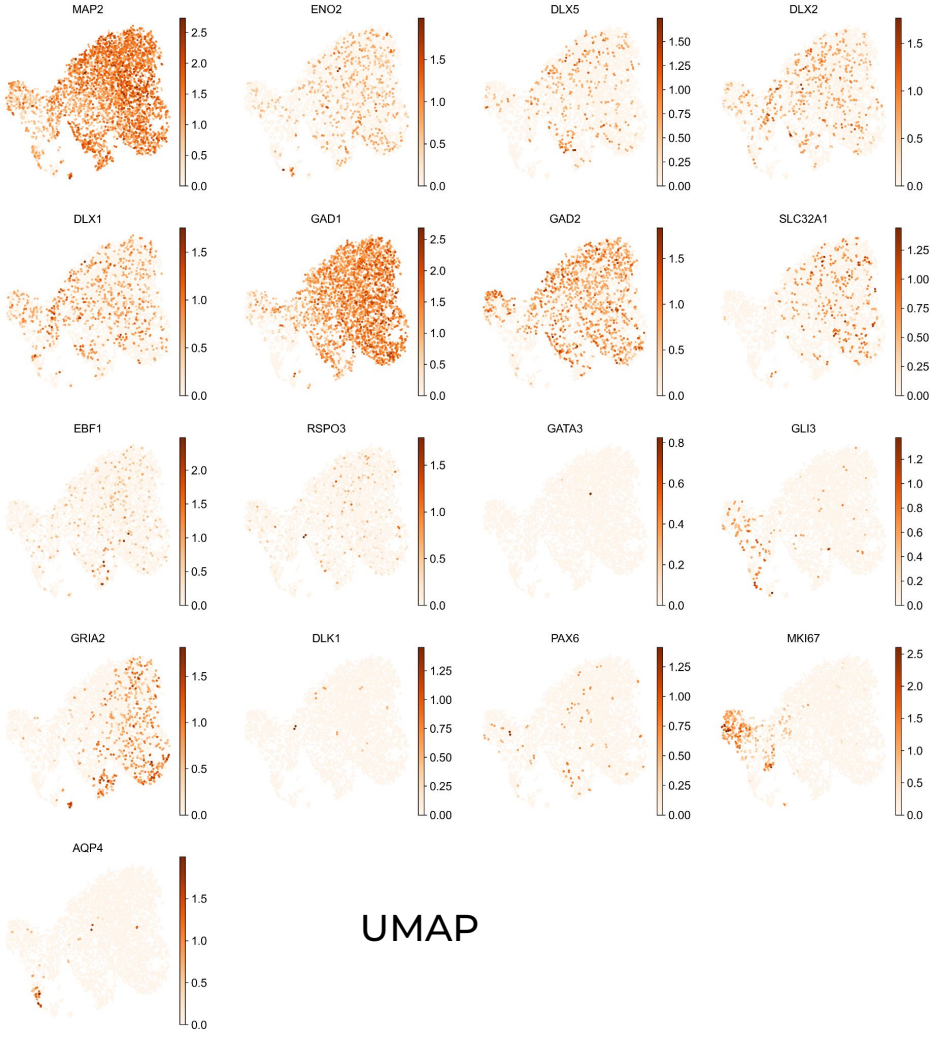
[1] For more information and comparisons to other gene selection methods, refer to Lause et al. (2021).

Brain Organoid Genes

Here, we want to check how many genes associated with brain organoids included in [1] are observed in our sample.

The number of highly variable marker brain organoid genes in our data is 17 (from a list of 37 genes).

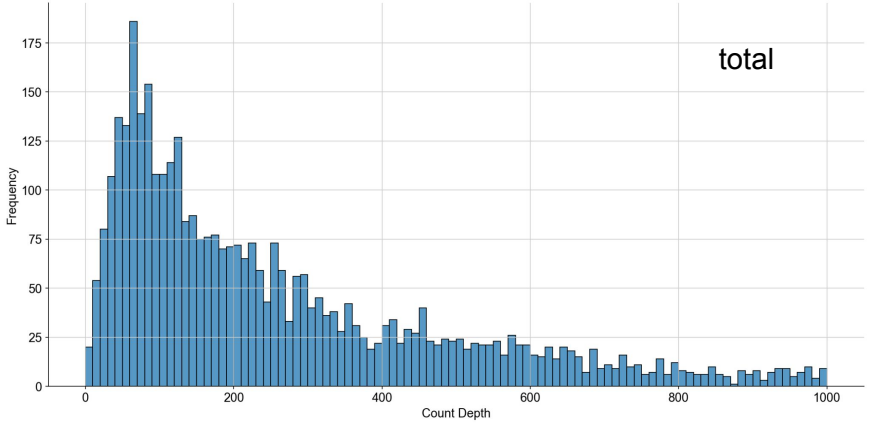
Their expression is shown on the right UMAP figure.



[1] Kanton, Sabina, et al. "Organoid single-cell genomic atlas uncovers human-specific features of brain development." Nature 574.7778 (2019): 418-422.

Outline

Exploratory Antibodies Data Analysis



- Figures are for an intersection between RNA and antibodies data

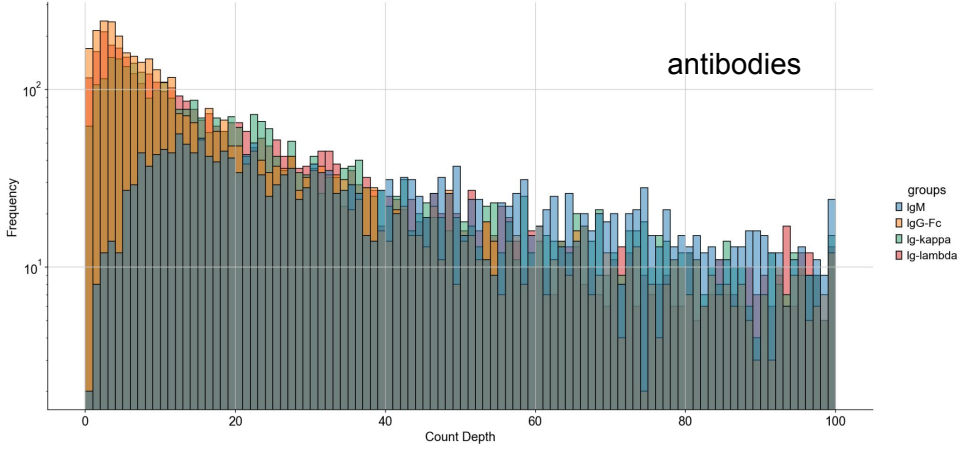
Total number of cells: **229,621**

Total Counts for Donor-1: 534,118

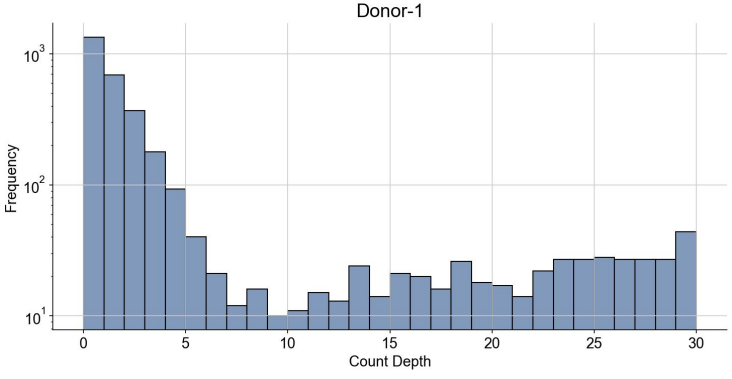
Total Counts for Donor-2: 673,080

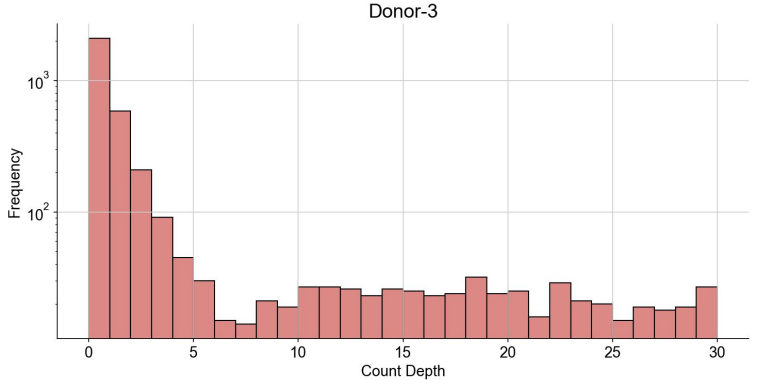
Total Counts for Donor-3: 273,355

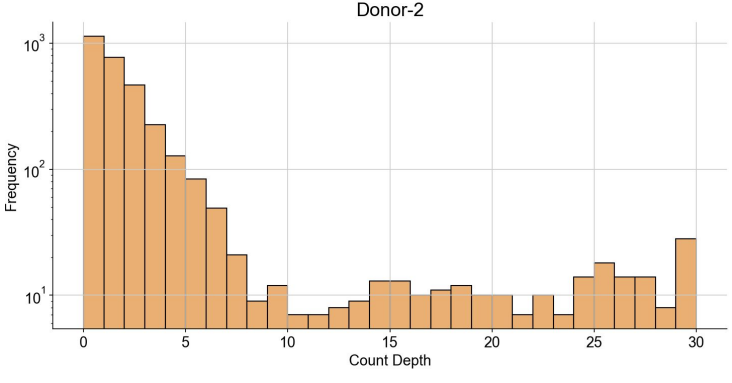
Total Counts for Donor-4: 257,469

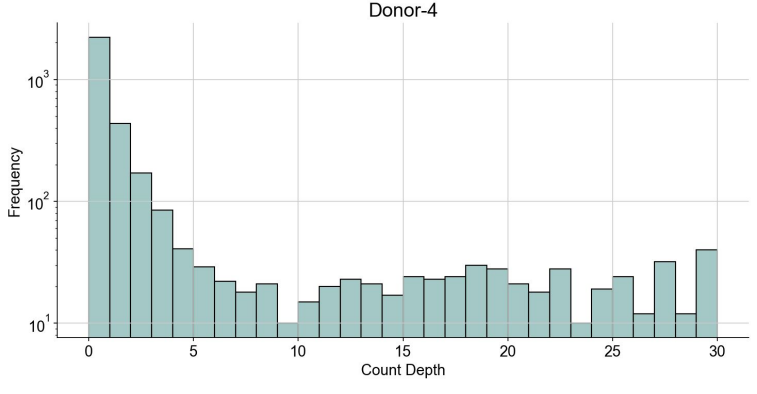


Exploratory Antibodies Data Analysis



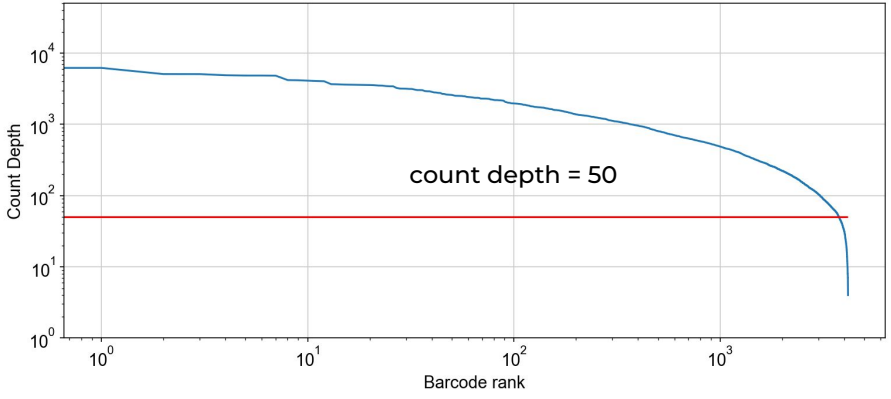


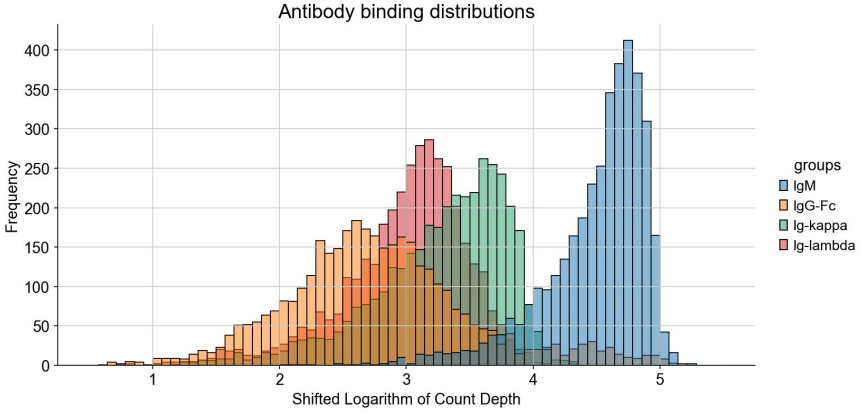




Outline

Filtering of Antibodies Data





1. Filter out cells with total count depth for antibody binding less than 50
2. If a count for a donor < threshold, then put this count to 0.

Threshold for Donor-1 is 9
Threshold for Donor-2 is 8
Threshold for Donor-3 is 7
Threshold for Donor-4 is 9

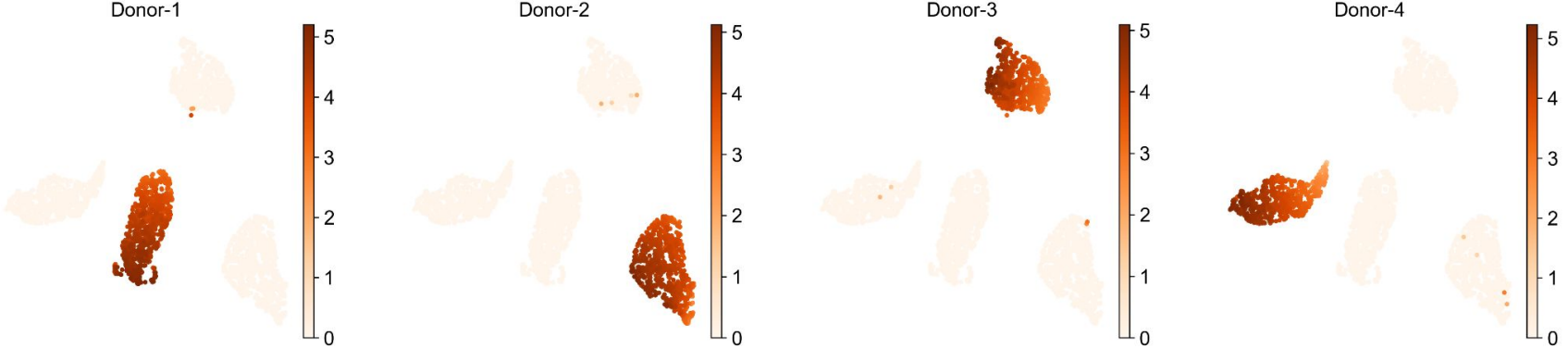
Number of cells after filtering of low quality counts: **3,761**

Heterogeneous Doublet Detection

When several cells from different donors are in the same droplet
(a.k.a. doublet), sequencing results may be misrepresented and spoil the analysis.

If we have counts more than zero for any 2 donors, it must be a doublet and we filter it out.

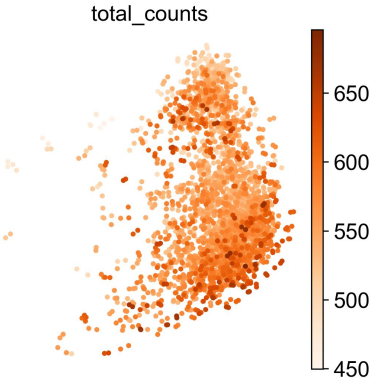
Number of doublets: 926

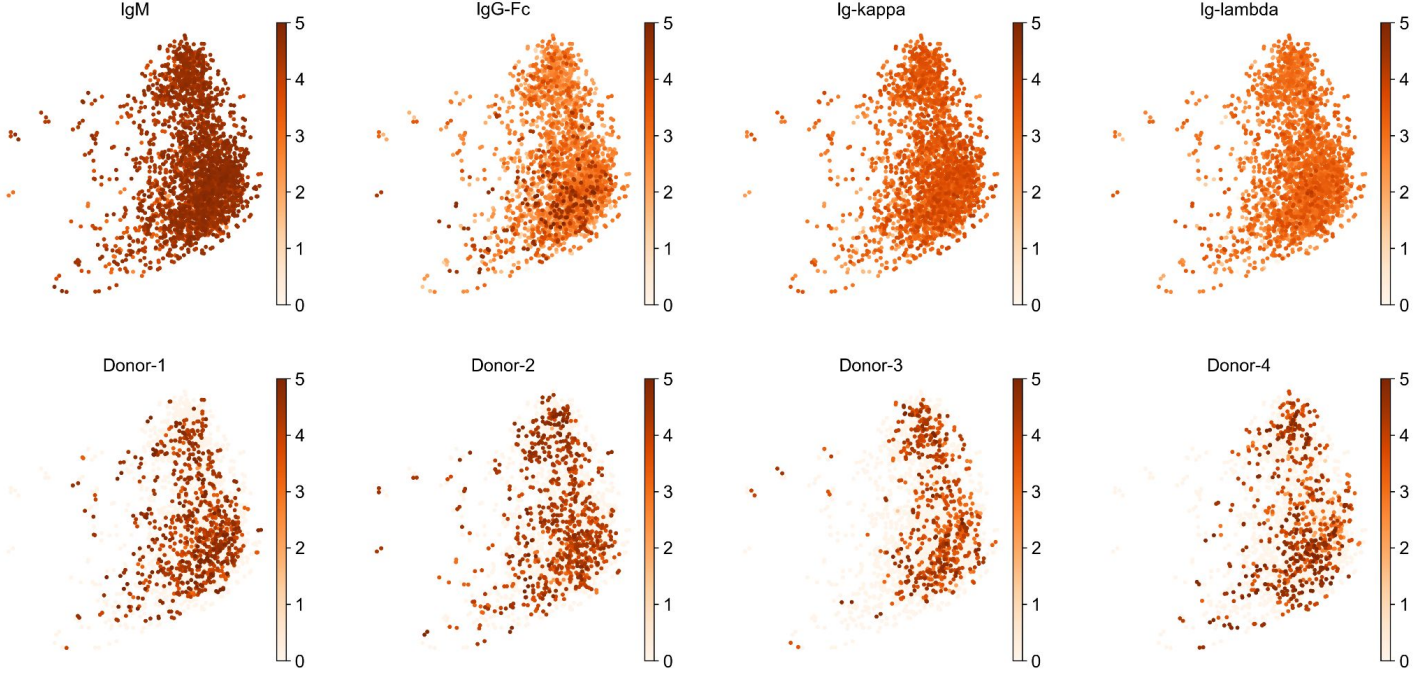


Outline

Dimensionality Reduction

PCA

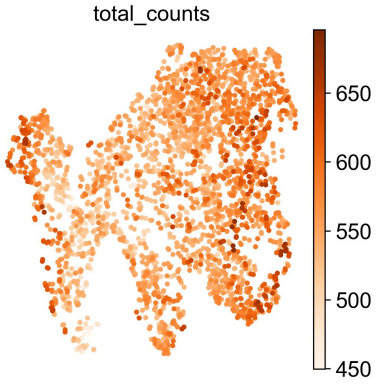


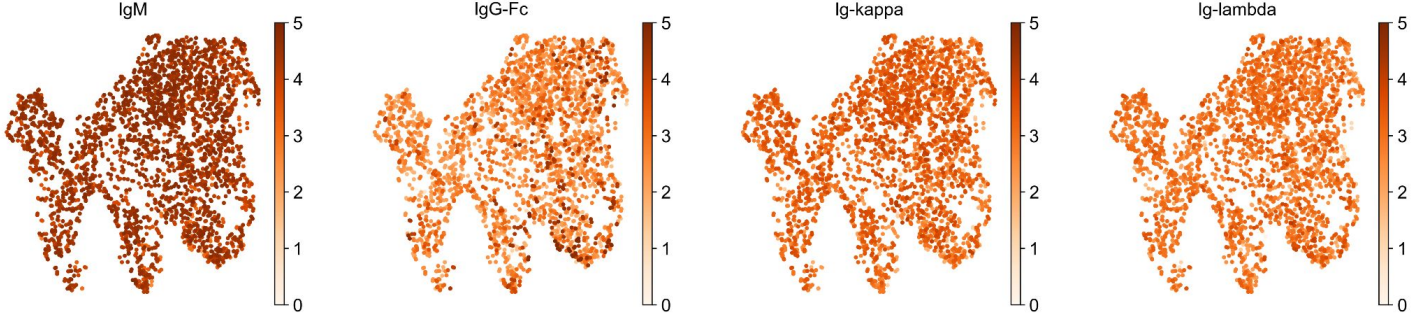


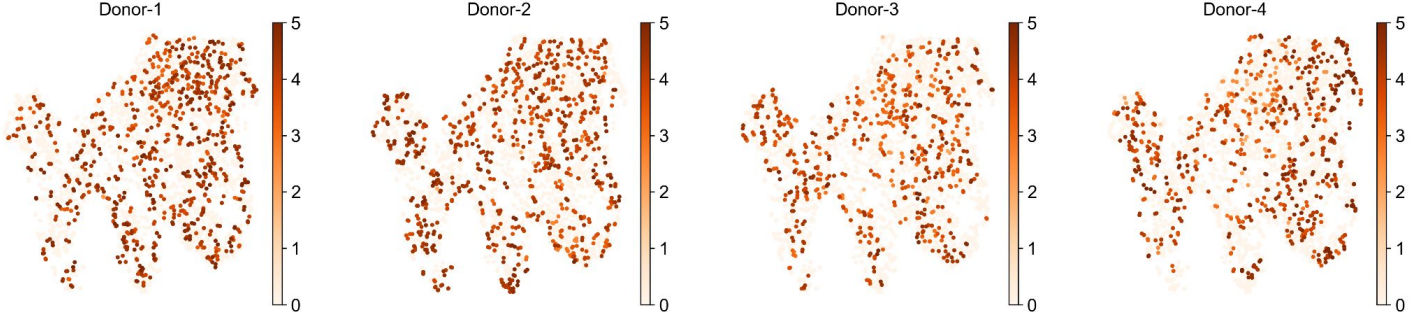
- Normalized counts

Dimensionality Reduction

UMAP



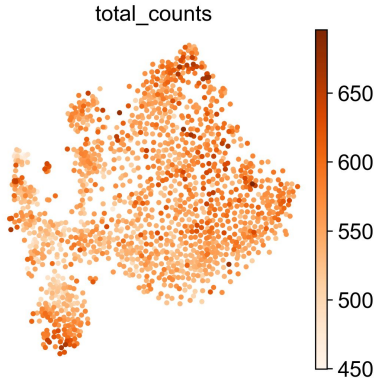


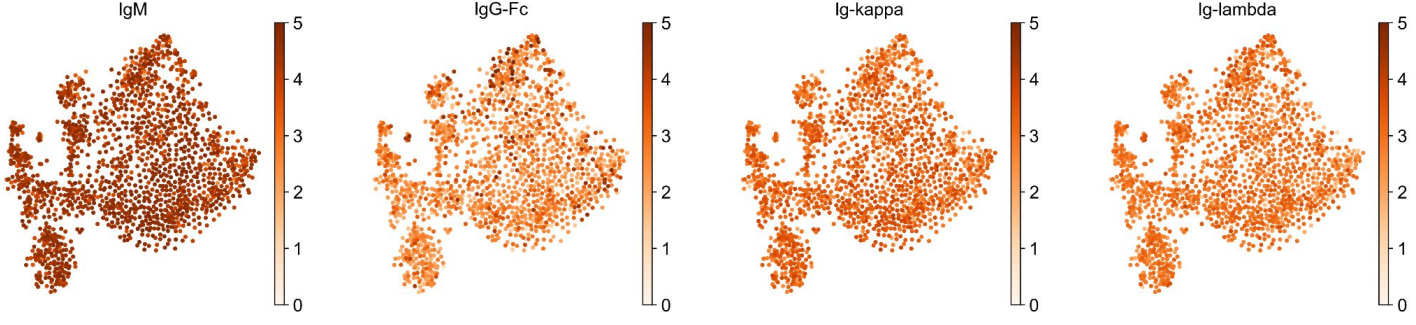


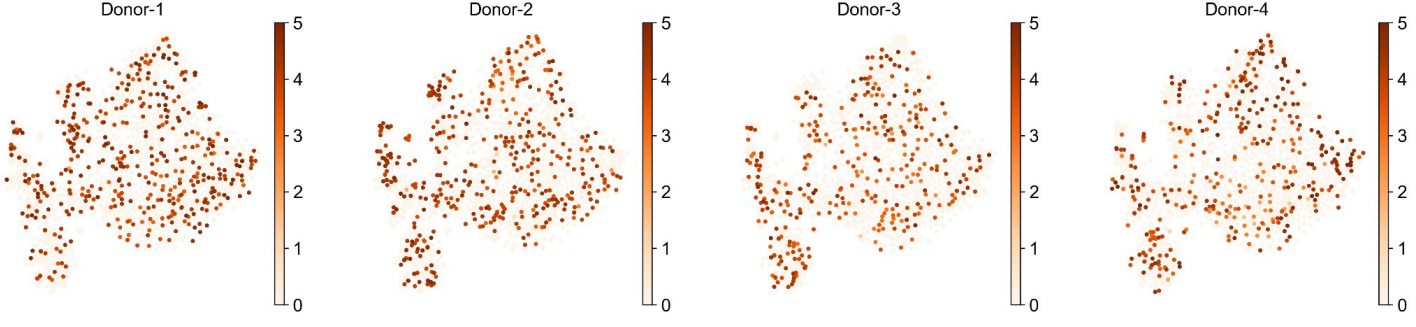
- Normalized counts

Dimensionality Reduction

t-SNE



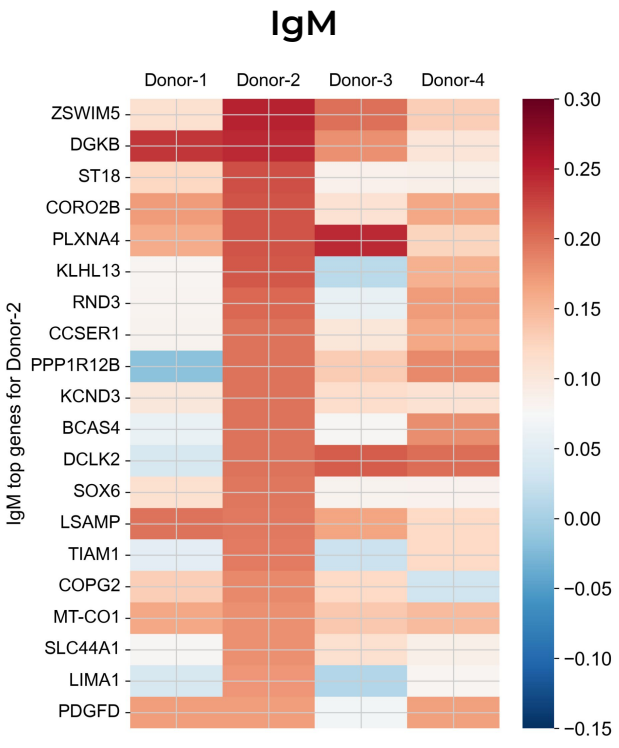


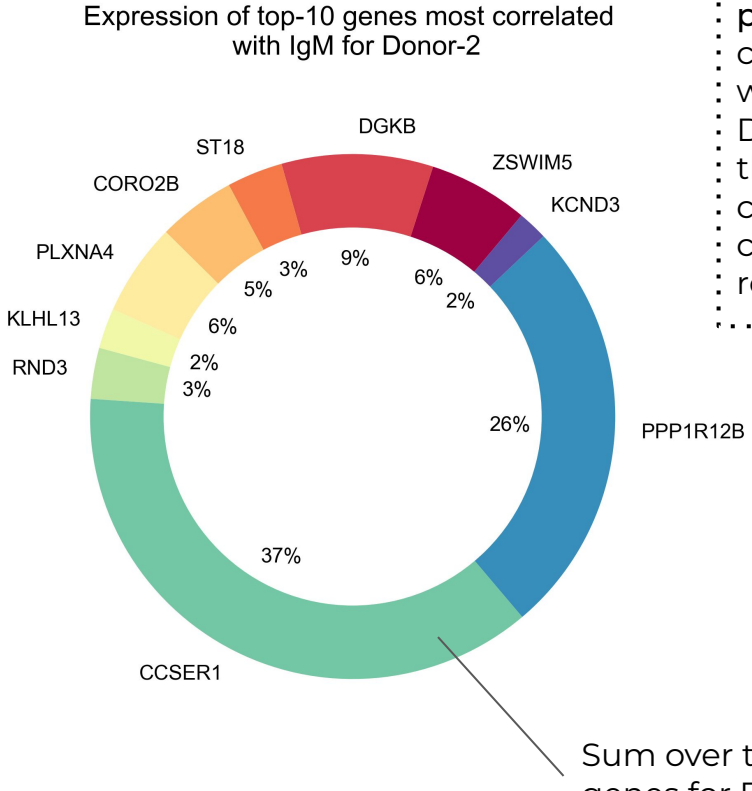


- Normalized counts

Outline

Spearman's Rank Correlation

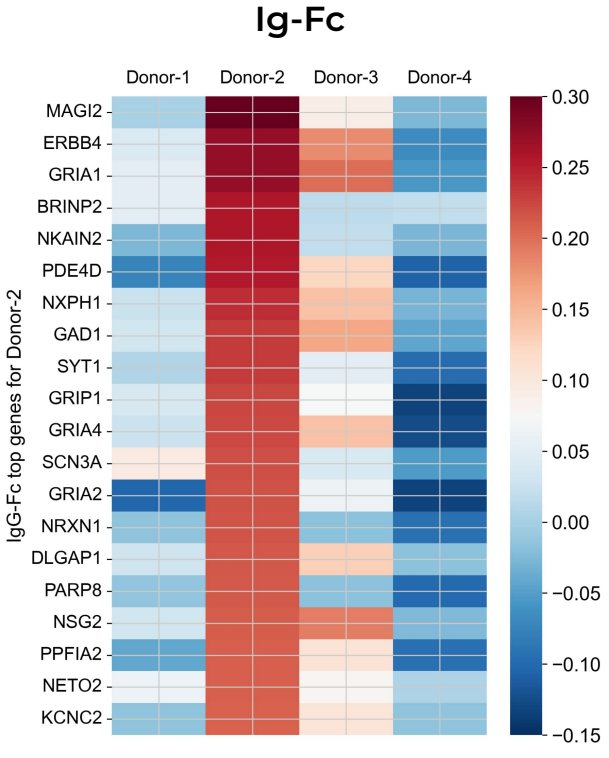


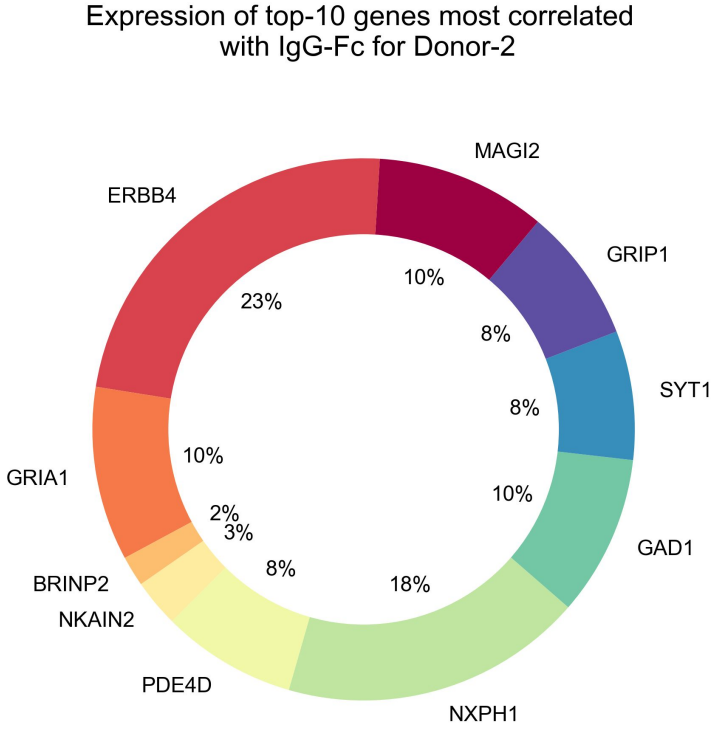


The top 20 **positively** correlated genes were selected for Donor-2, and for this list correlation coefficients were computed for the rest of the donors.

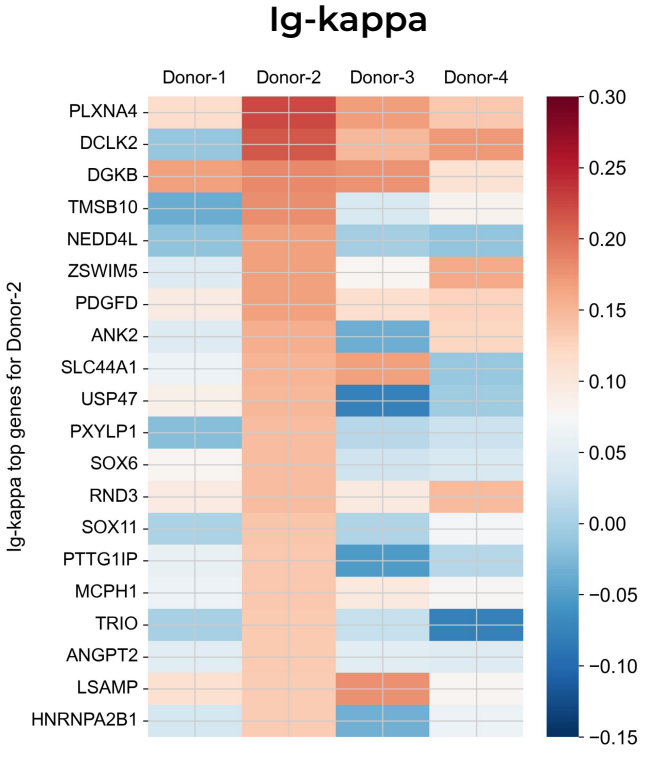
Sum over the top correlated genes for Donor-2

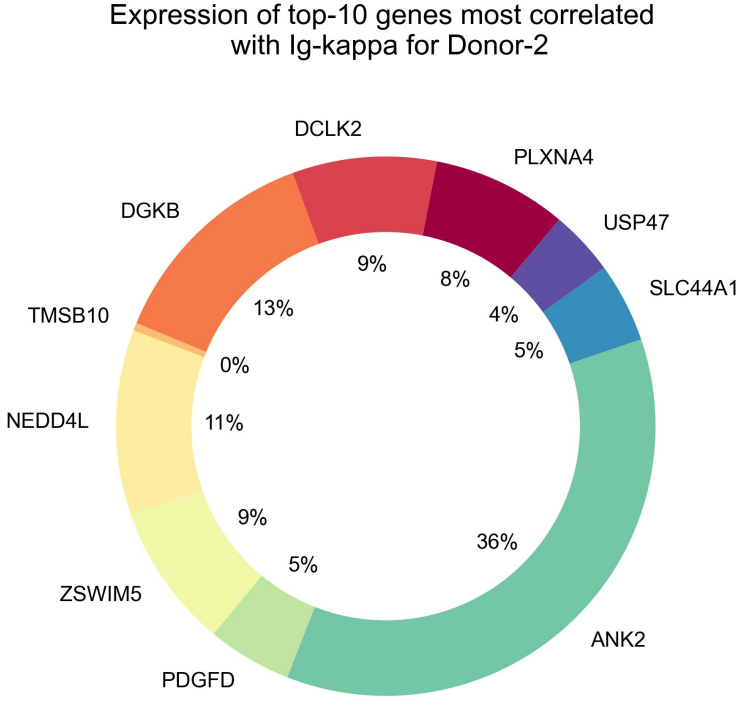
Spearman’s Rank Correlation



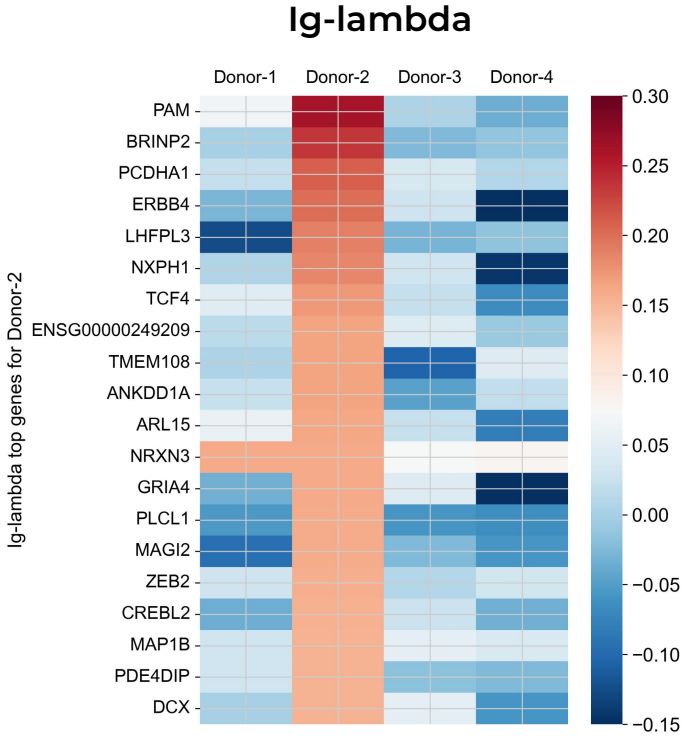


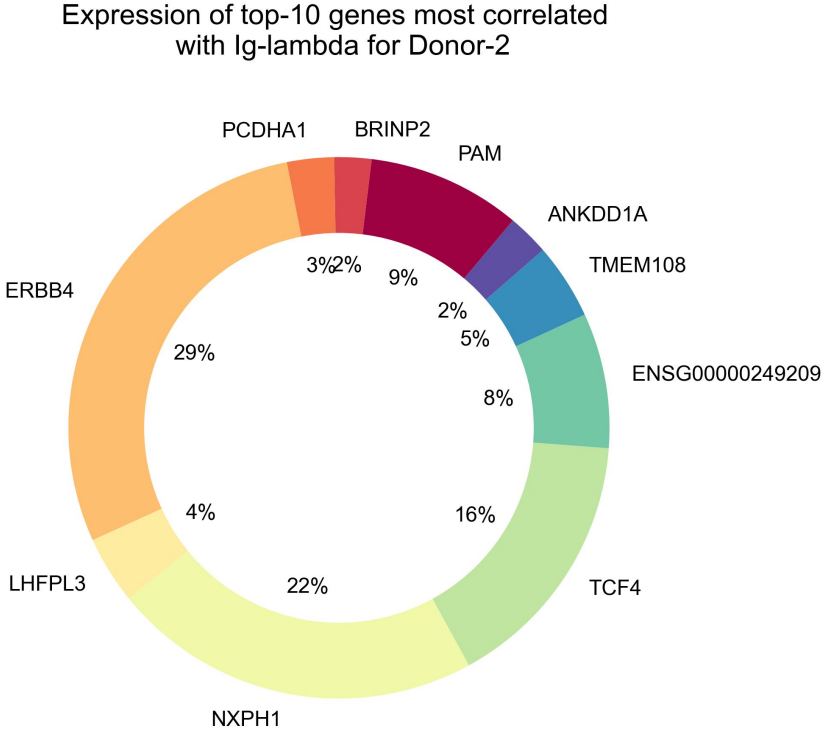
Spearman’s Rank Correlation



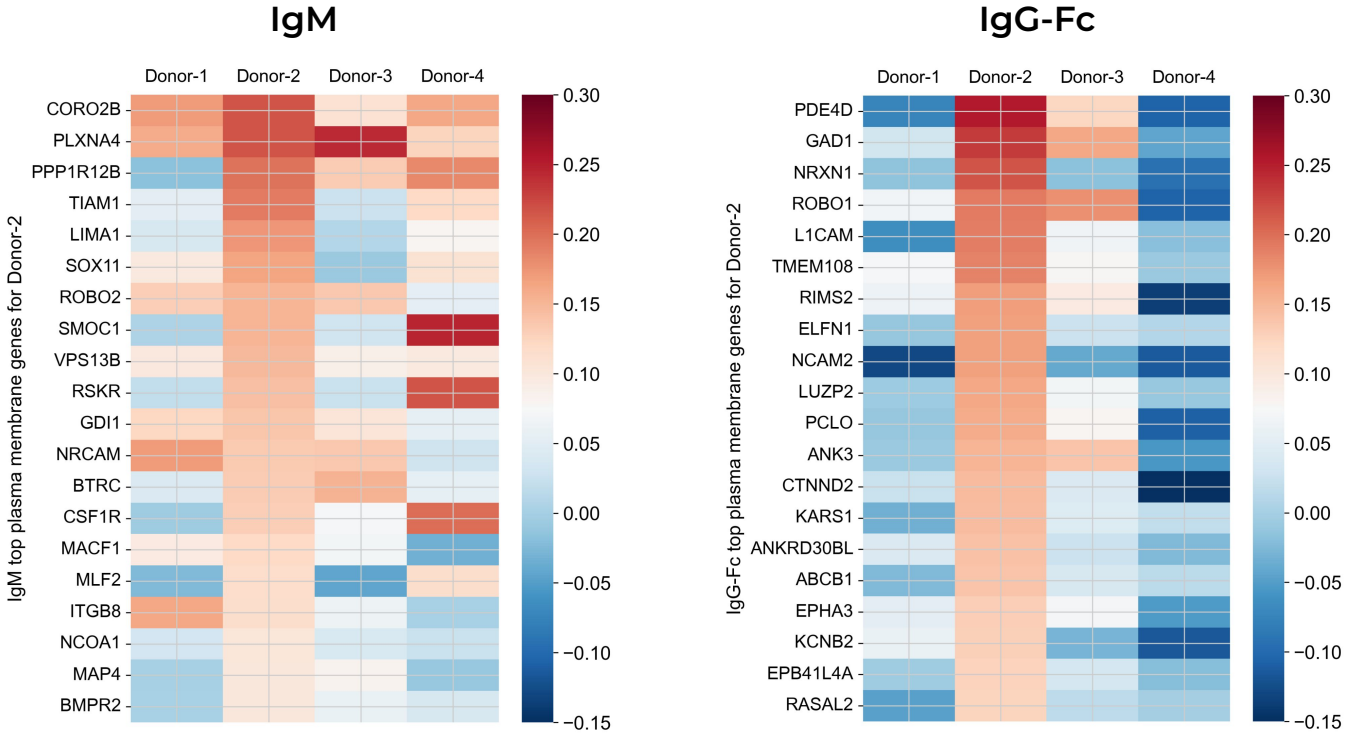


Spearman’s Rank Correlation



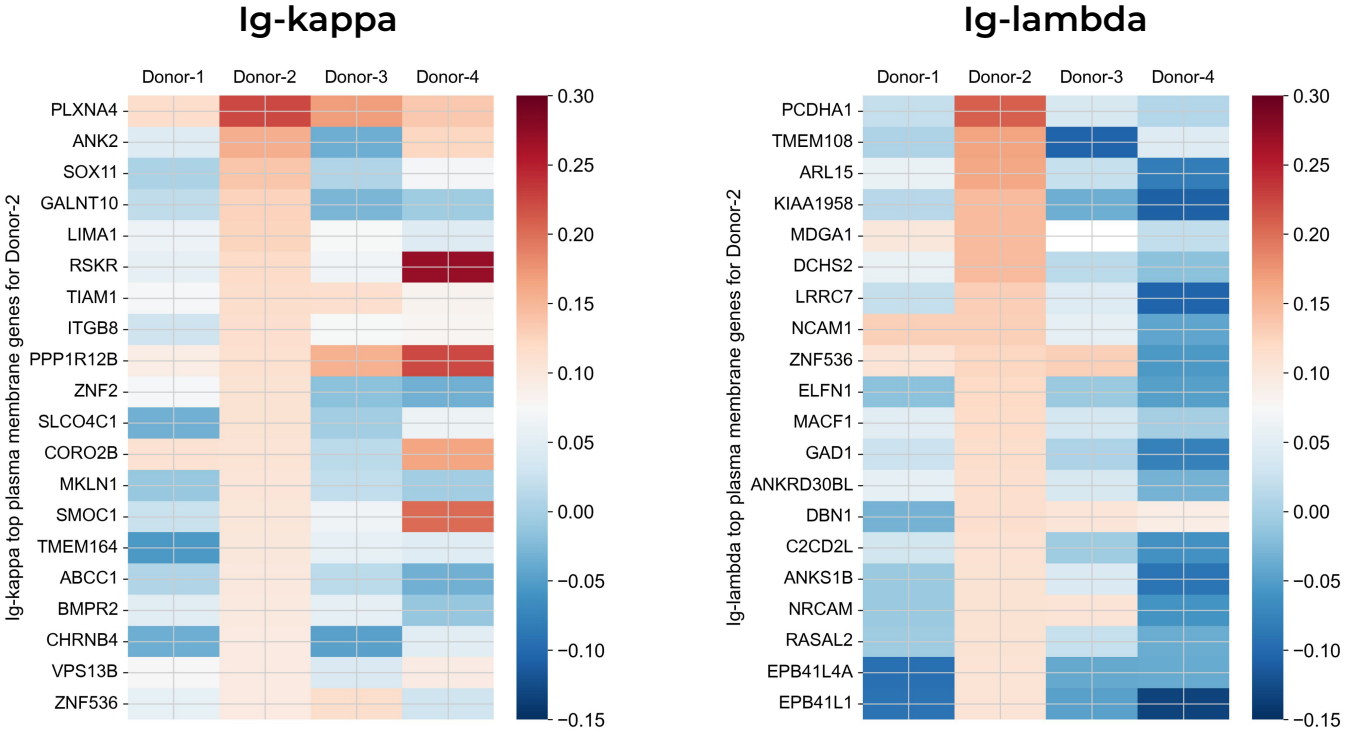


Plasma Membrane Gene Selection



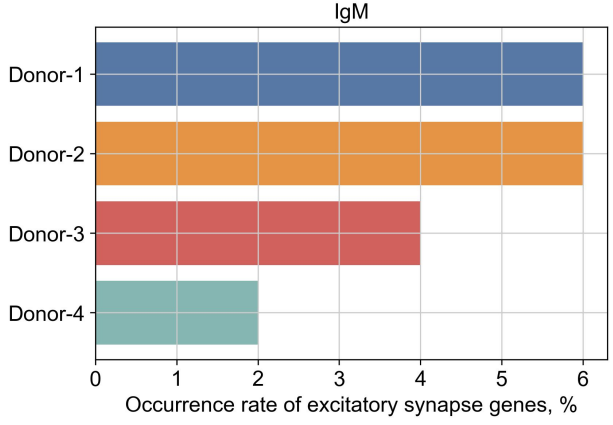
The top 20 positively correlated genes were selected for Donor-2, and for this list correlation coefficients were computed for the rest of the donors.

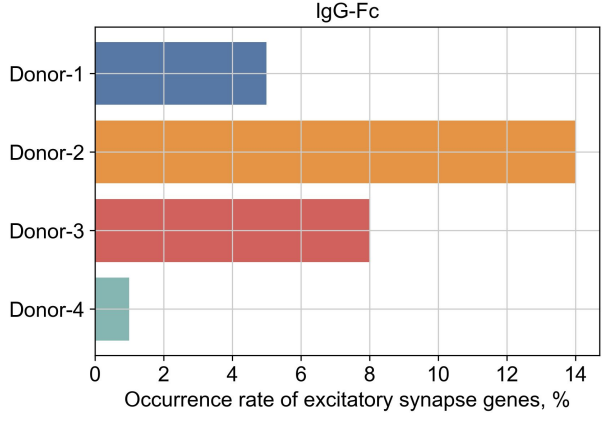
Plasma Membrane Gene Selection

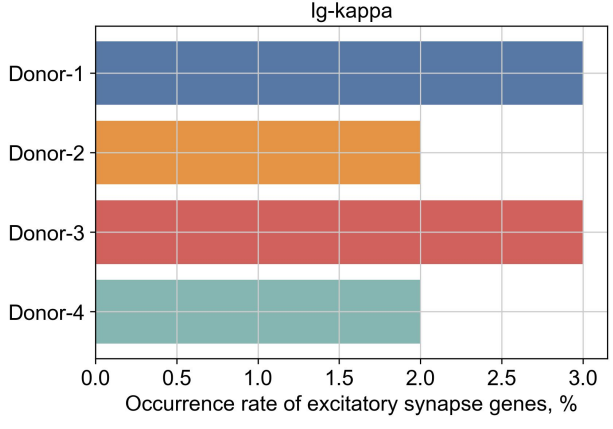


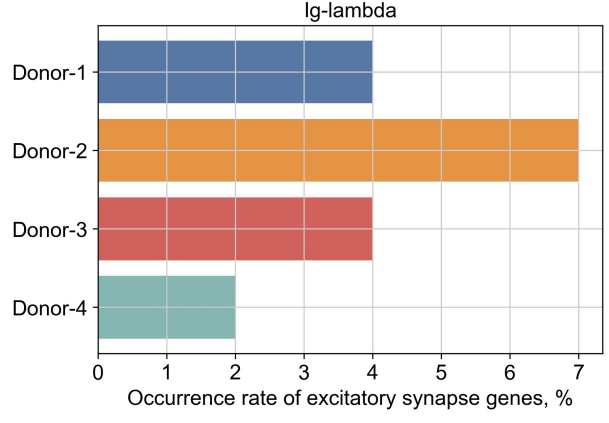
The top 20 positively correlated genes were selected for Donor-2, and for this list correlation coefficients were computed for the rest of the donors.

Antibody Binding and Gene Expression









⁎ From the top-100 correlated genes

Conclusions

In conclusion, this project represents a step forward in understanding autoimmune diseases through the integration of single-cell RNA sequencing and antibody binding data. By analyzing human brain organoid cells bound with antibodies from autoimmune disease patients, the study aims to uncover correlations between gene expression and binding.

Even though the absolute values of the correlation coefficients for genes are not very high, the list of genes is reproducible. Moreover, there is still room for improvement in the results: implementing stricter quality control measures and annotating clusters to examine local correlations between antibody binding and gene expression.

The choice to rely solely on Python for data analysis reflects the potential for more effective big data processing and the flexibility it offers for cutting-edge research. Using the CITE-seq protocol, promises to provide valuable insights into the proteins responsible for autoimmune diseases.