Analysis of CITE-seq data from a brain organoid

Anastasiia Popova

30.01.2024

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 today is the **CITE-seq protocol**, which integrates cellular protein and single-cell RNA (scRNA) measurements into a single assay using oligonucleotide-labeled antibodies.

design of experiment

We have 4 samples each of it consists of cerebral (brain) organoid cells and blood serum belonging to one of 4 of donors.

method

Chromium Next GEM Single Cell 5' Reagent Kits v2 (Dual Index) and data preprocessing using Cell Ranger.

goal

We are searching for a gene (or genes) responsible for producing a target protein bound by the antibodies in the blood serum from a donor with a disease.

Our hypothesis is that there should be a significant correlation between the counts of antibodies binding and the gene reads encoding this protein.

Antibody Data

We already have 8,913 cells uniquely assigned to each donor from Cell Ranger.

Quality Control

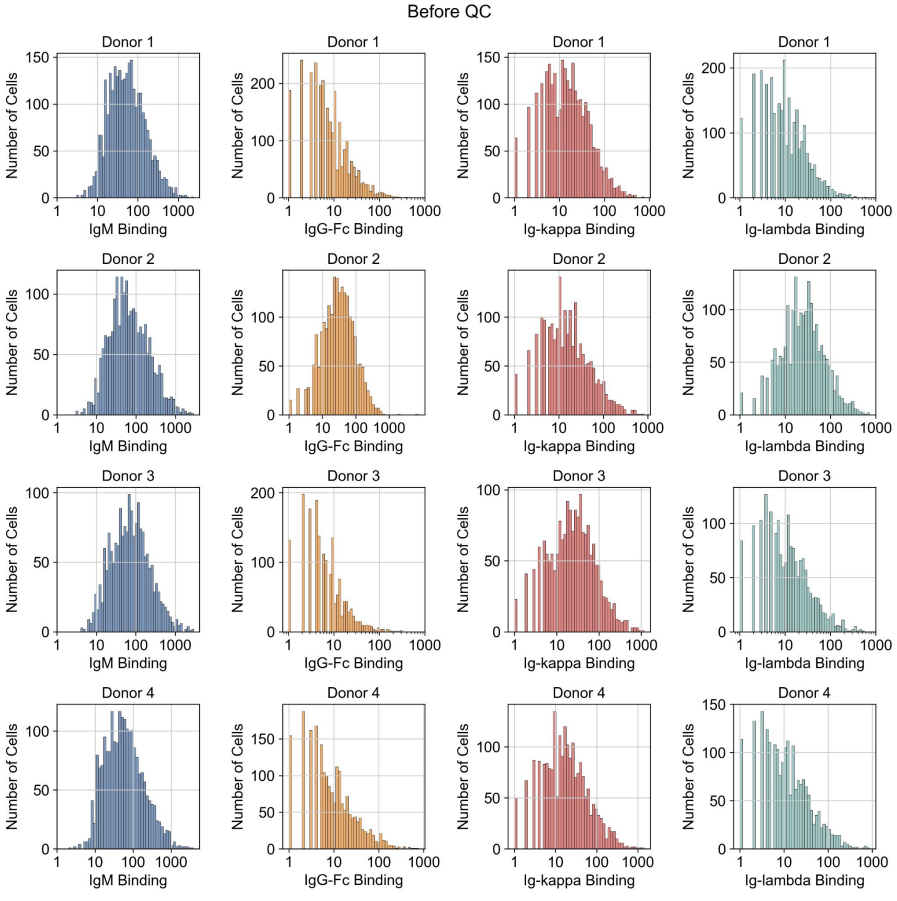
We remove outliers in our data, defined as values differing by 5 MADs (median ± 5 MAD), for the following metrics:

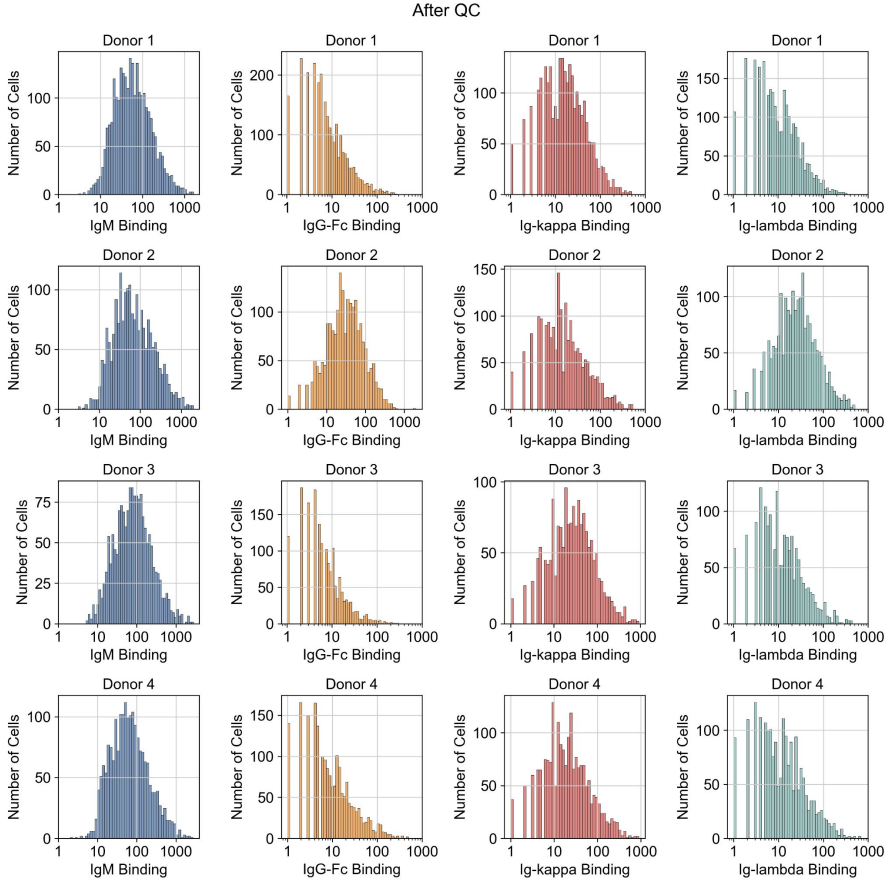
- Logarithm of the number of antibodies with at least 1 count in each cell.
- Logarithm of the total counts in each cell.

Normalization

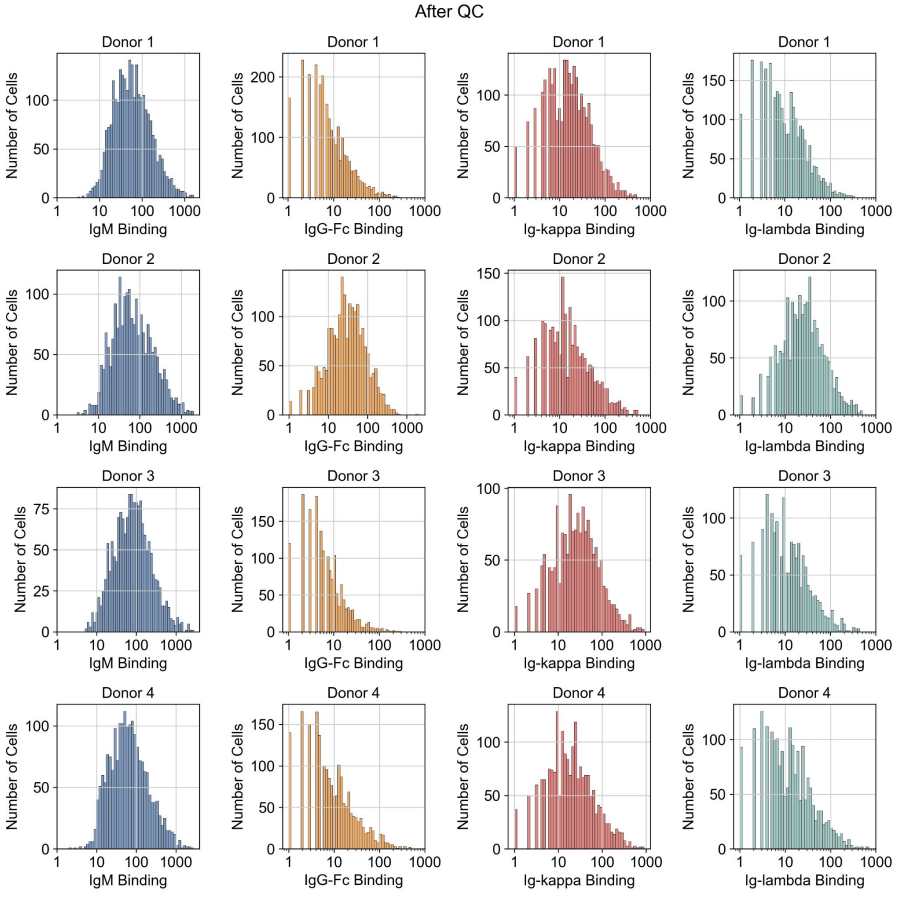
We normalize the data using the shifted log normalization with a scaling parameter of 10.

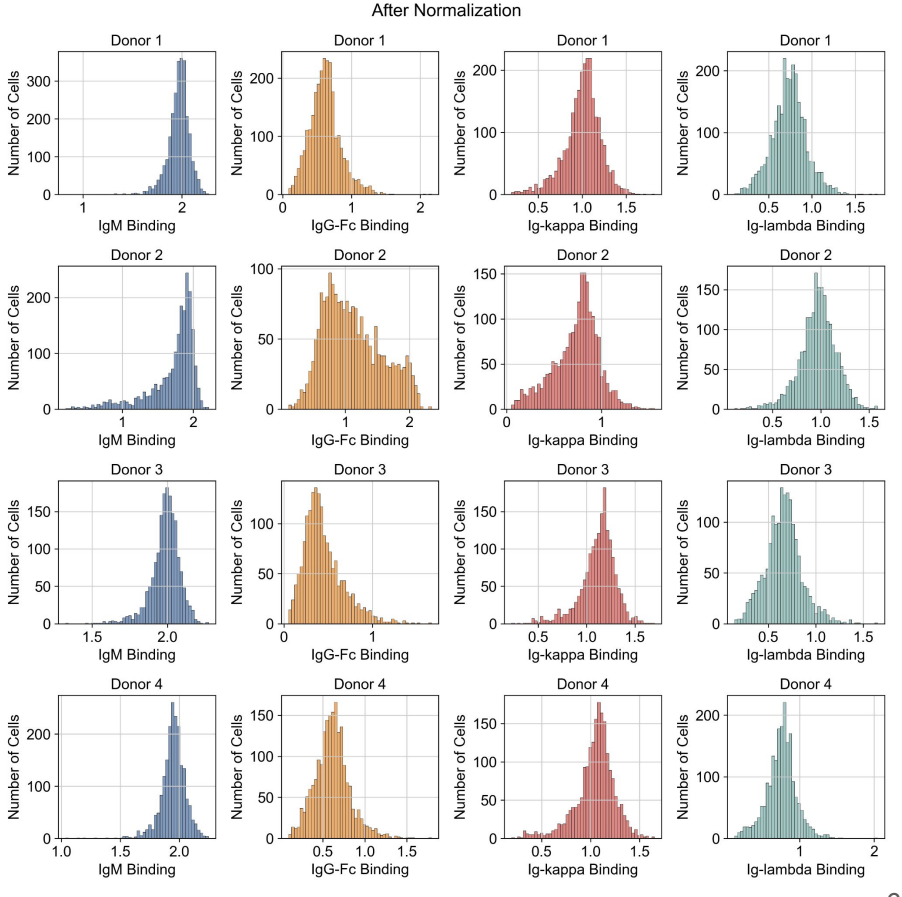
Antibodies Binding





Antibodies Binding





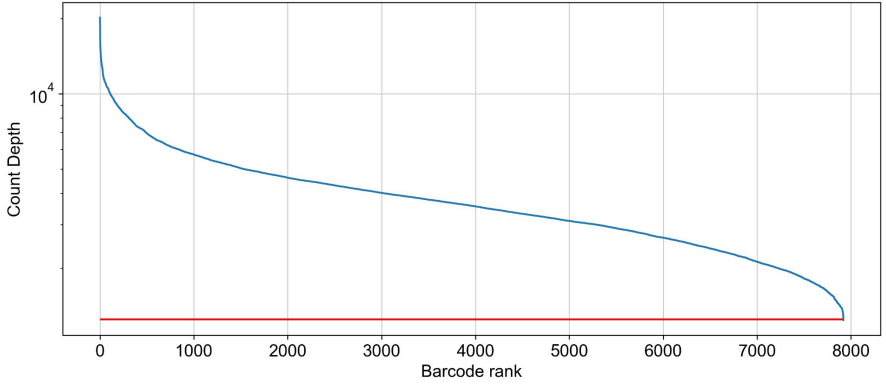
scRNA-seq Data Processing

After Cell Ranger data processing and the antibody quality control, the **scRNA dataset** consists of **7,920 cells and 36,601 genes**.

Gene Preselection

We are interested in working only with protein-coding genes, so a gene preselection was done. The list of genes is from **Ensembl BioMart** database.

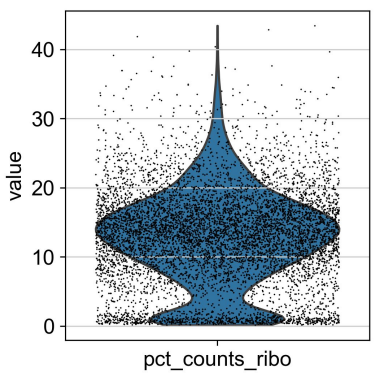
After this stage we have **19,253** genes in the dataset. Ambient RNA Correction was already done in Cell Ranger, the count depth was approximately 1250. The minimum number of genes detected in a cell is 1000 (on the next slide).

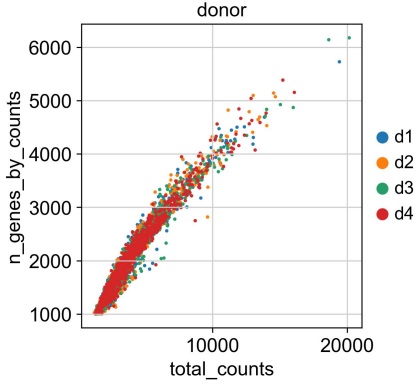


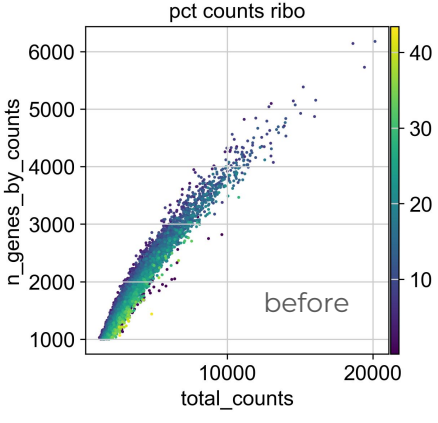
Damaged Cells Filtration

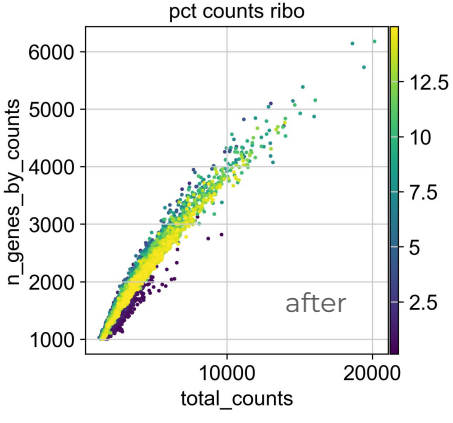
To filter out cells with broken membranes, which usually have a high amount of mitochondrial counts, few detected genes and a low count depth (the number of counts per barcode). *To ensure accurate representation of informative mRNA transcripts, it is necessary to exclude cells with high ribosomal RNA reads from further analysis (?).* Ribosomal gene names start with "RPS" and "RPL"; mitochondrial with “MT-”

- Outliers - cells which have a high proportion of mitochondrial/ribosomal gene counts (differ by 5 MAD of the median) - are filtered out.
- Also, cells with a percentage of mitochondrial 8 % and ribosomal counts 15 % exceeding are filtered out.
- Number of cells after filtering of low quality cells: 4,914







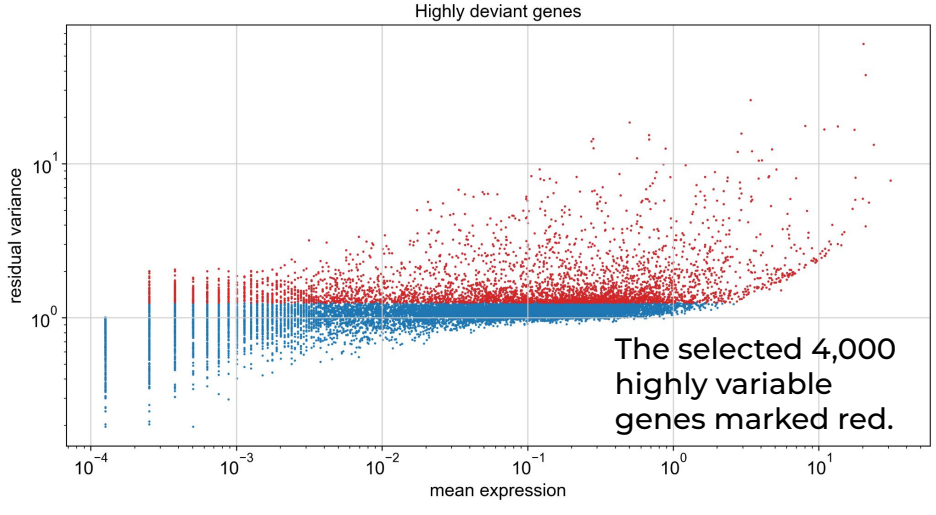


Gene Selection

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].

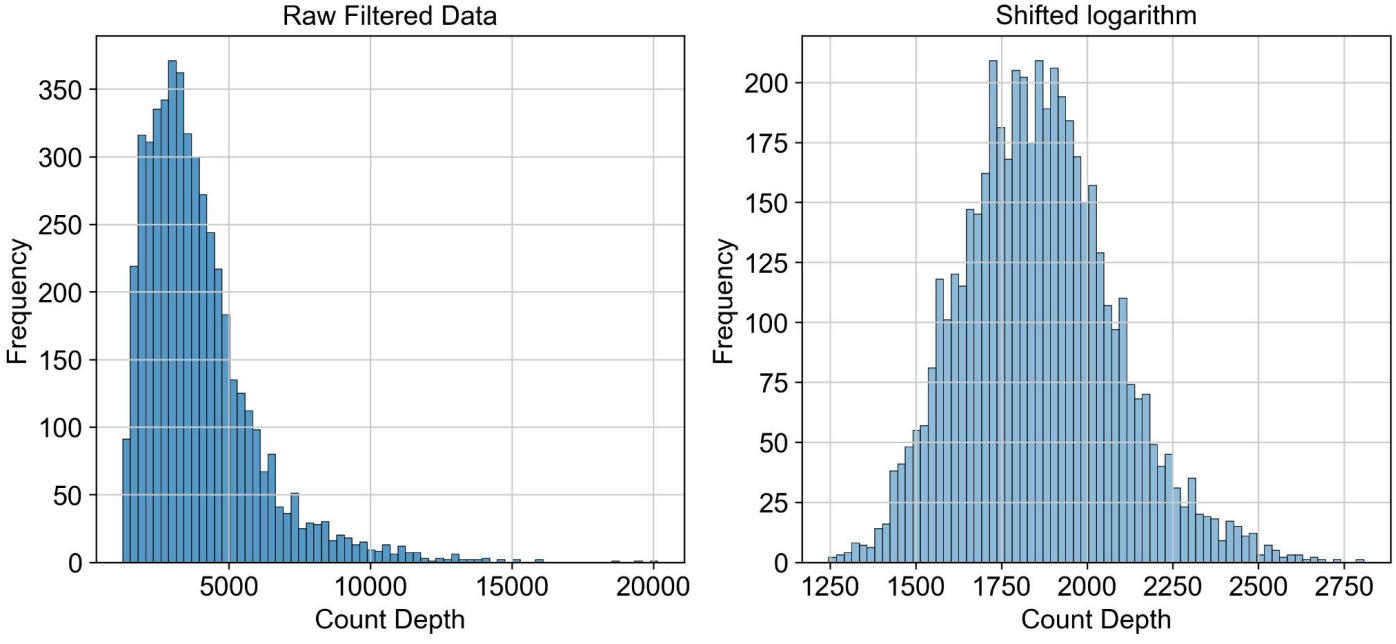
```
Top-20 highly variable genes
----------------------------------------

gene_name
CRABP1        0.0
ERBB4         1.0
HIST1H4C      2.0
SST           3.0
NXPH1         4.0
LSAMP         5.0
DLGAP1        6.0
NRXN3         7.0
NEGR1         8.0
VIM           9.0
GALNTL6      10.0
NKAIN2       11.0
DCC          12.0
TMSB4X       13.0
RBFOX1       14.0
CNTNAP2      15.0
ROBO1        16.0
CCSER1       17.0
PDE1A        18.0
PCDH9        19.0
```



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

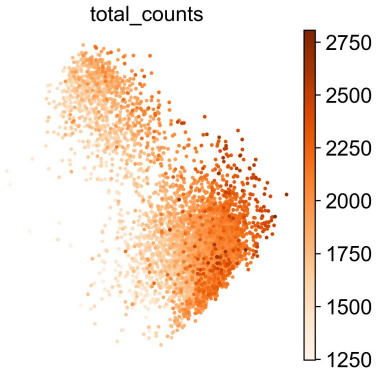
Normalization of RNA Data

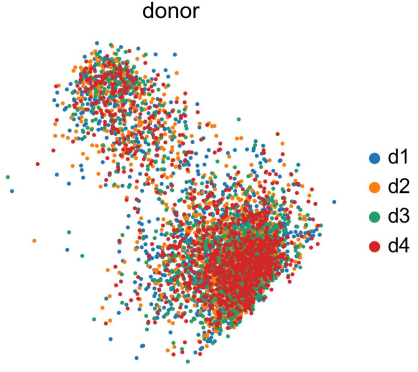
Shifted logarithm normalization aims to make the variances across the dataset more similar. It used for subsequent dimensionality reduction and identification of differentially expressed genes.



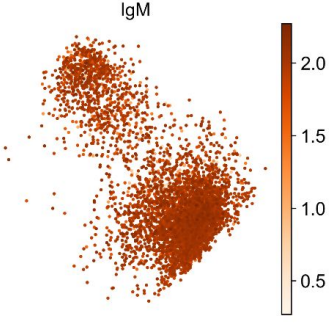
Dimensionality Reduction

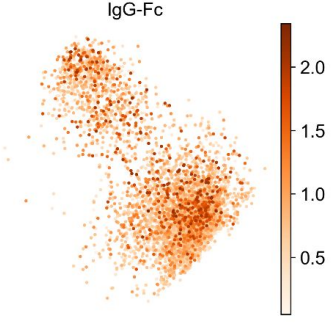
PCA

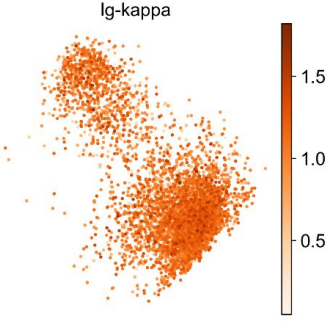


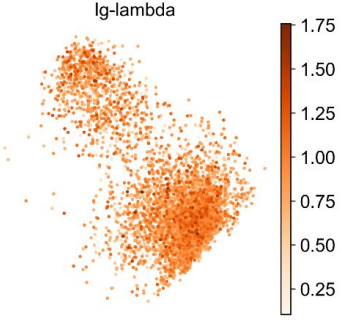


- Normalized counts



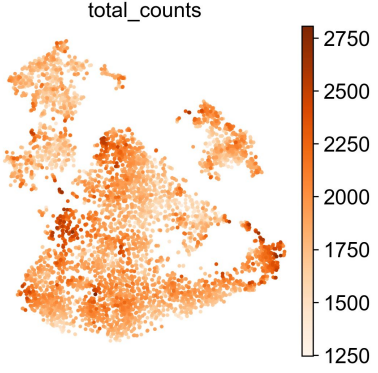




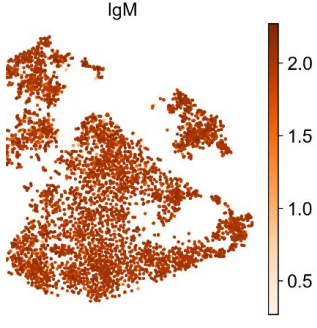


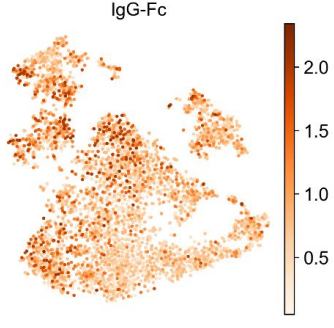
Dimensionality Reduction

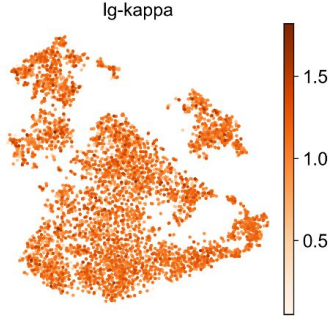
t-SNE

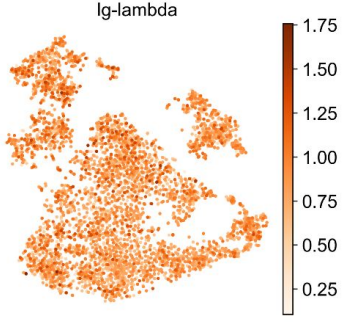


- Normalized counts







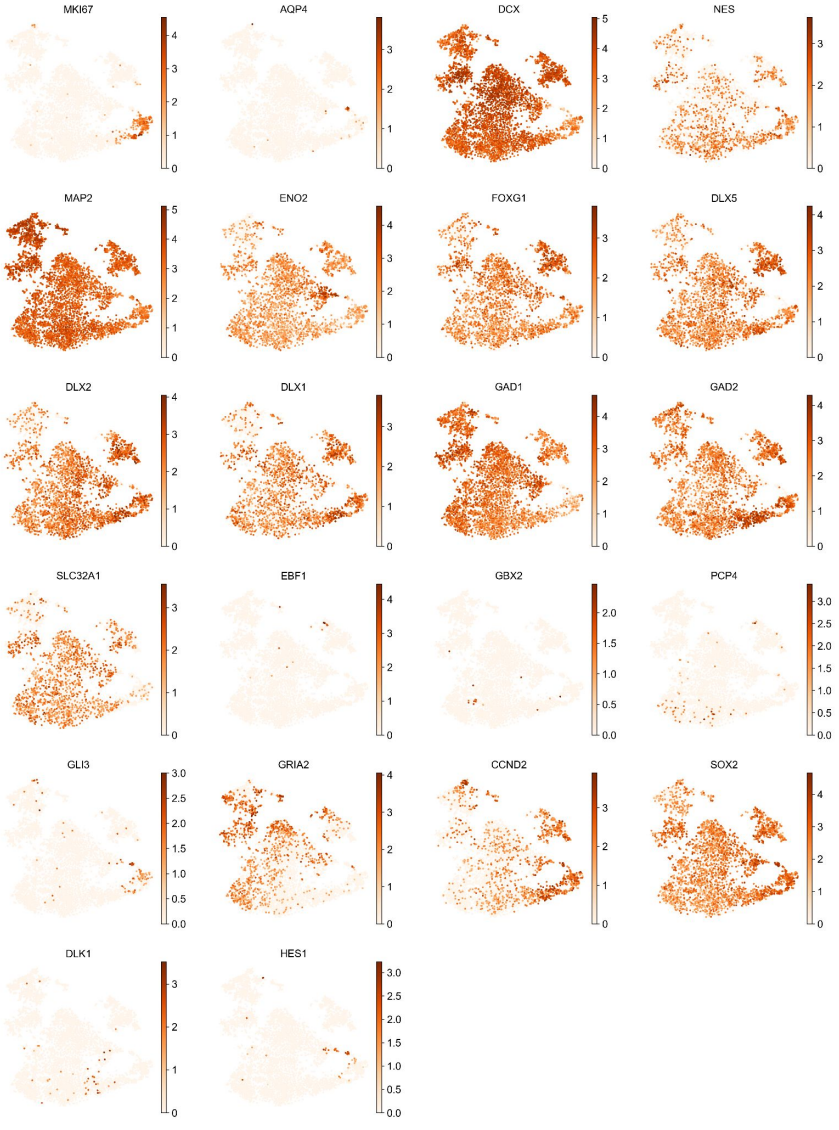


Brain Organoid Genes

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 22 (from a list of 41 genes).

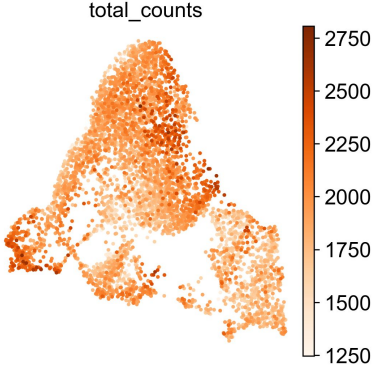
Their expression is shown on the right **t-SNE** figures.



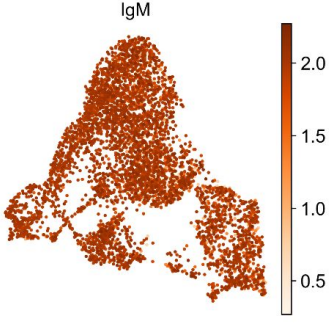
[1] Kanton, Sabina, et al. "Organoid single-cell genomic atlas uncovers human-specific features of brain development." Nature 574.7778 (2019): 418-422.

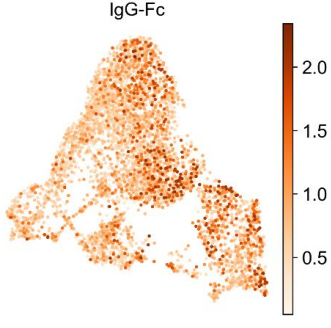
Dimensionality Reduction

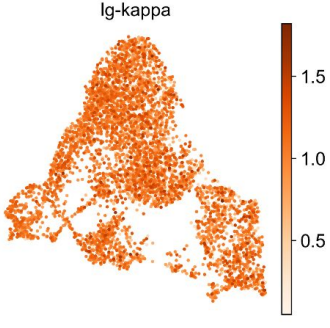
UMAP

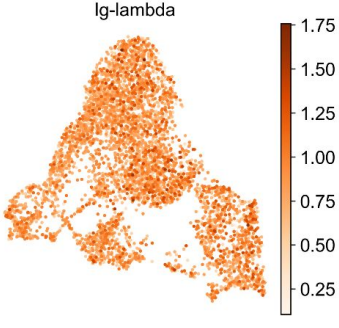


- Normalized counts

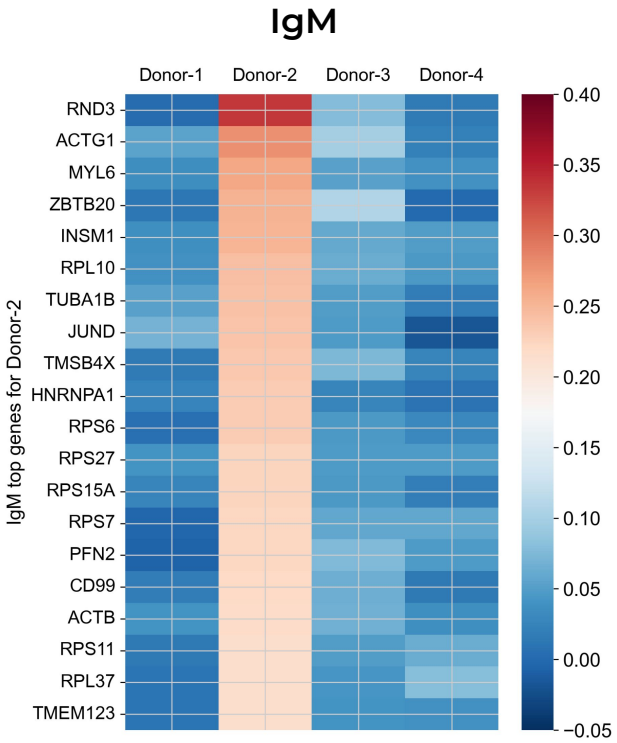


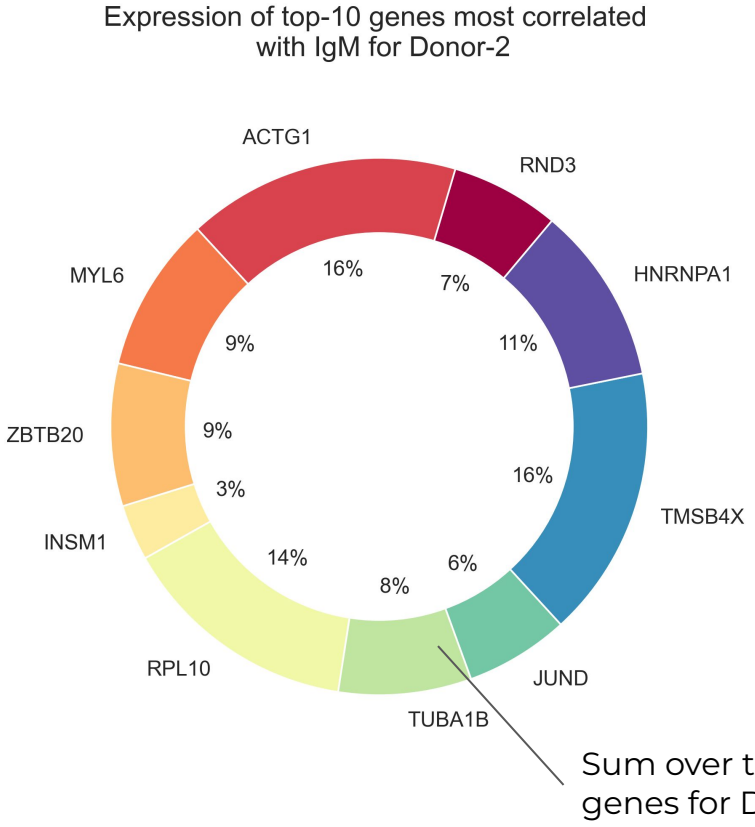






Spearman's Rank Correlation

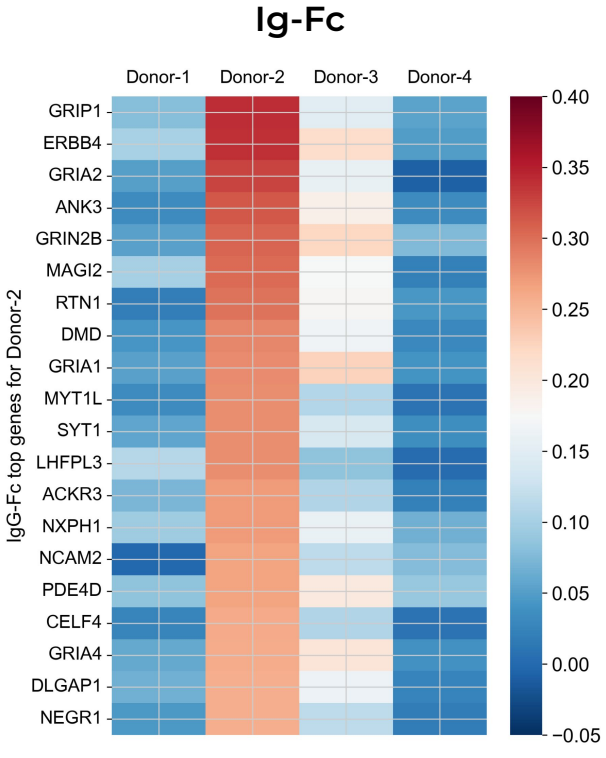


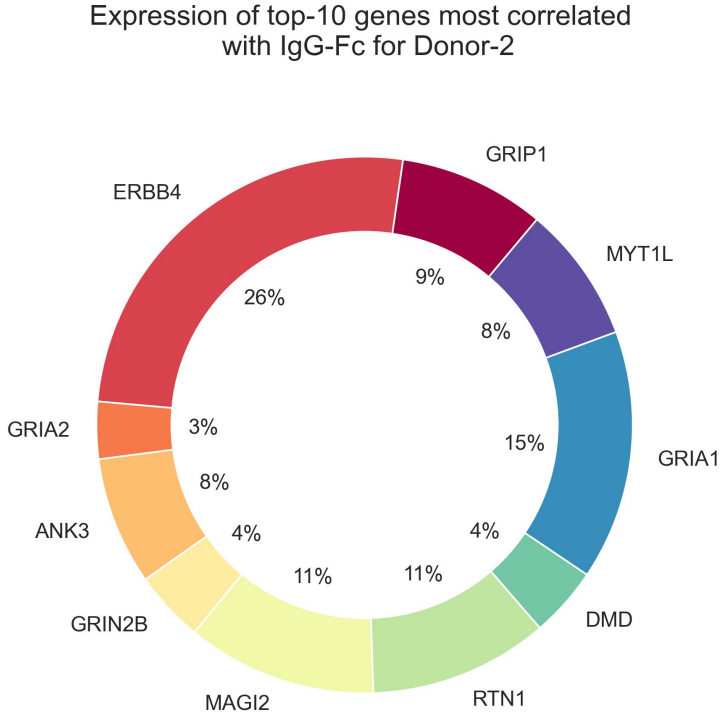


The top 20 **positively** correlated genes were selected for Donor-2, and for this list correlation coefficients were computed for the rest of the donors.

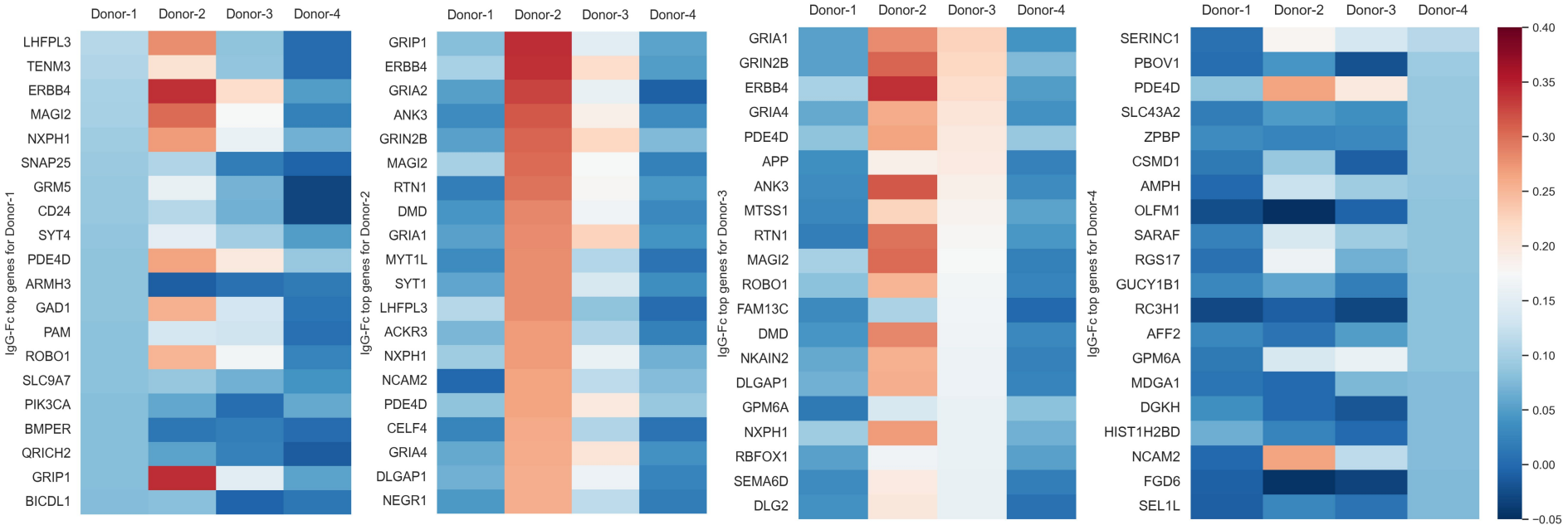
Sum over the top correlated genes for Donor-2

Spearman’s Rank Correlation

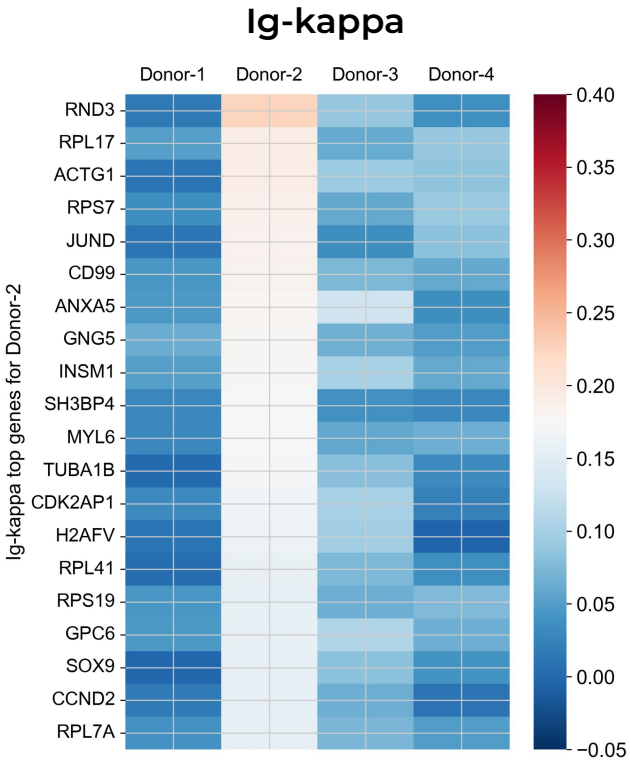


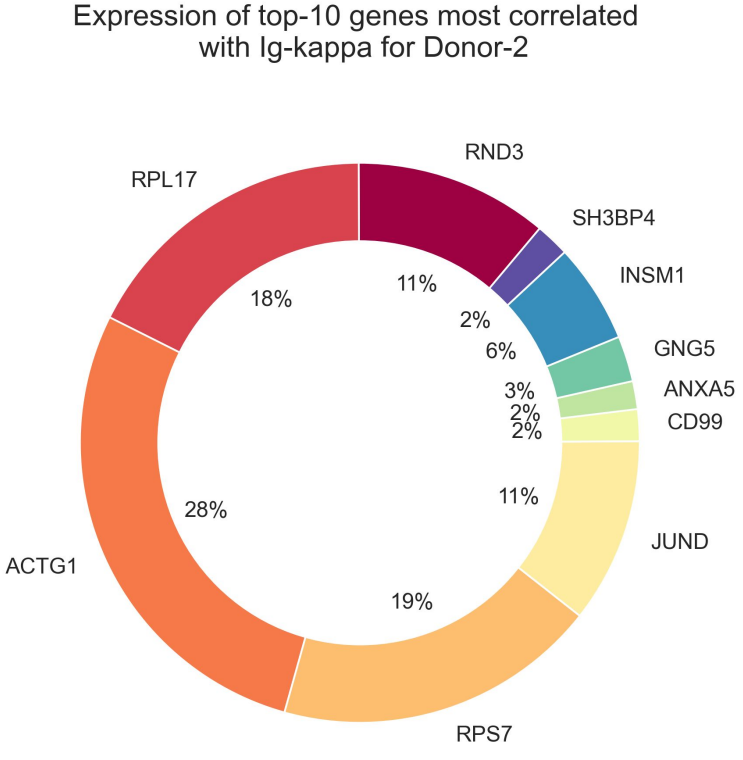


Correlations From Other Donors

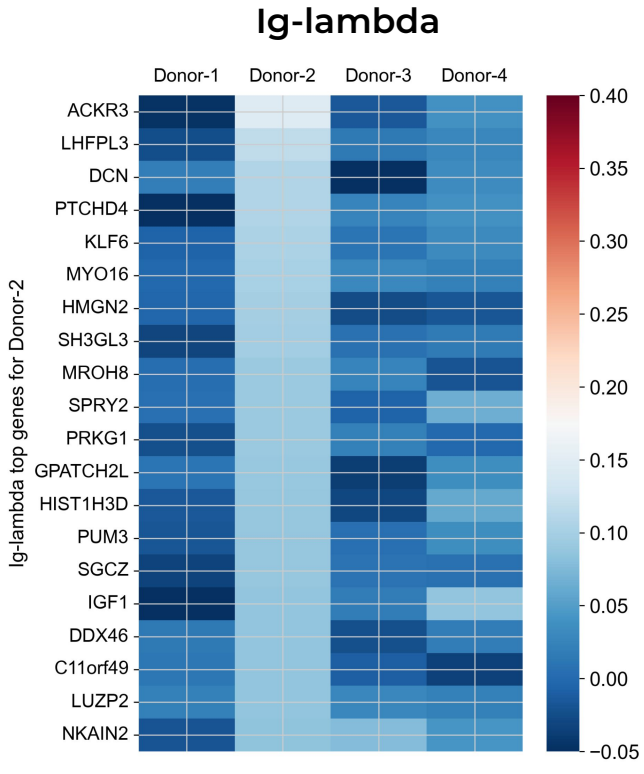


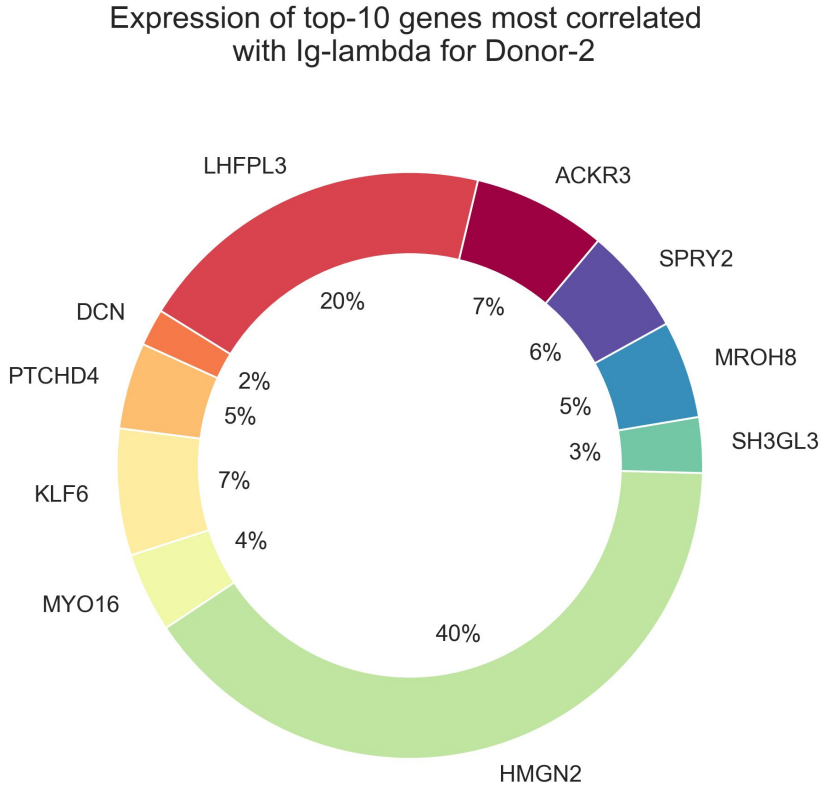
Spearman’s Rank Correlation



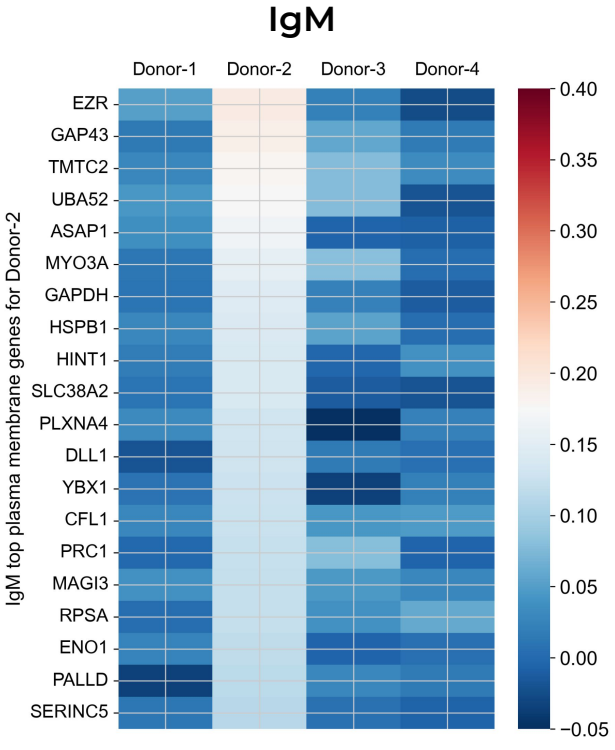


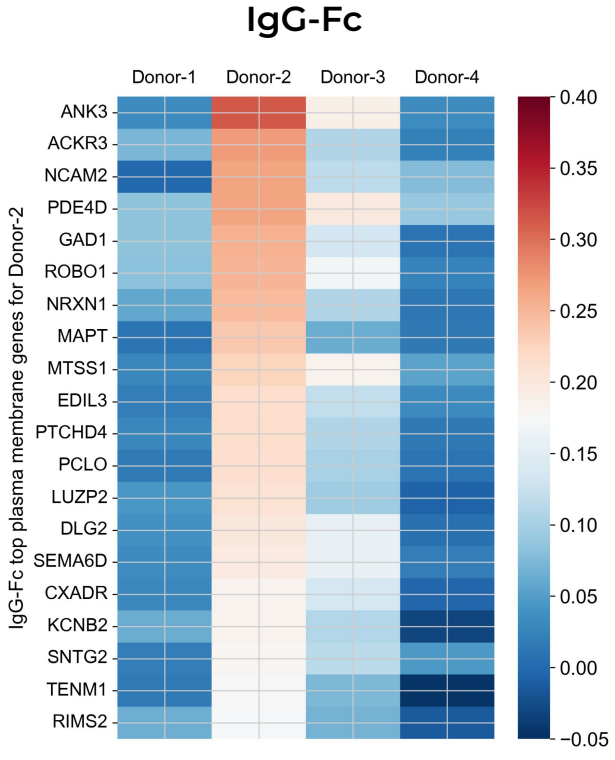
Spearman’s Rank Correlation





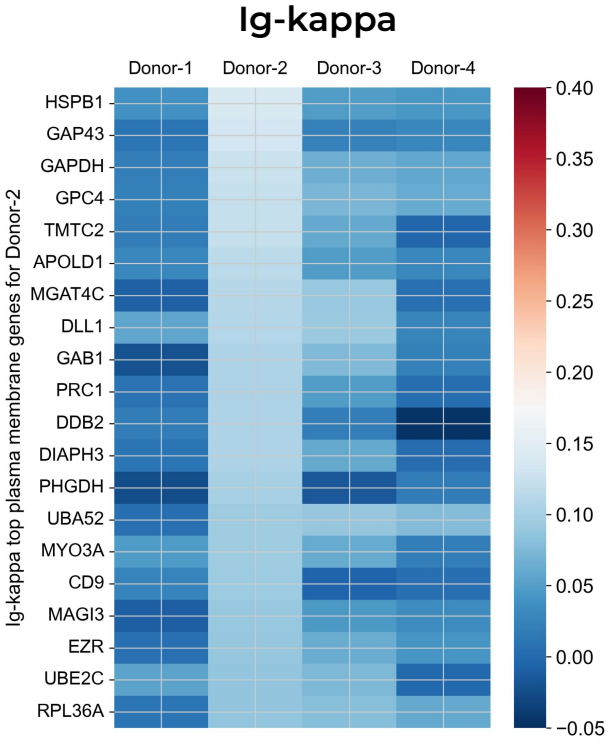
Plasma Membrane Gene Selection

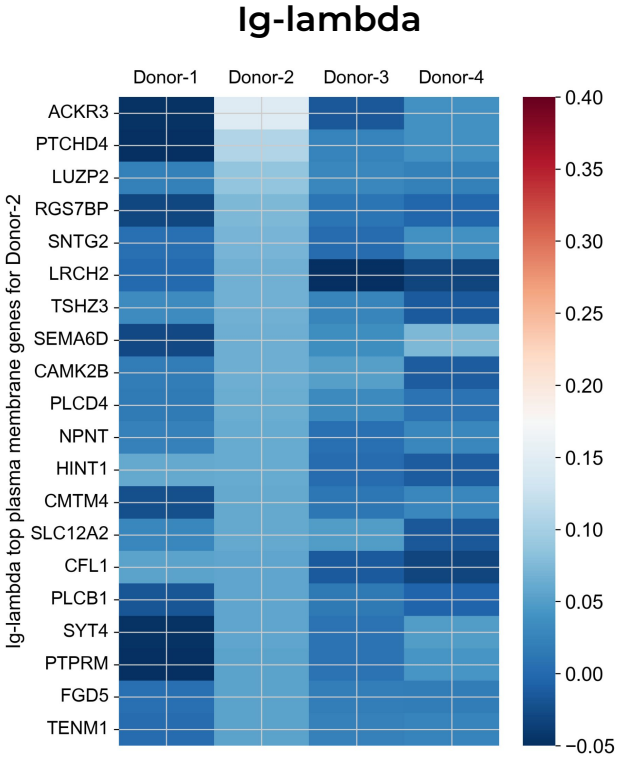




The top 20 positively correlated genes were selected for Donor-2, and for this list correlation coefficients were computed for the rest of the donors.

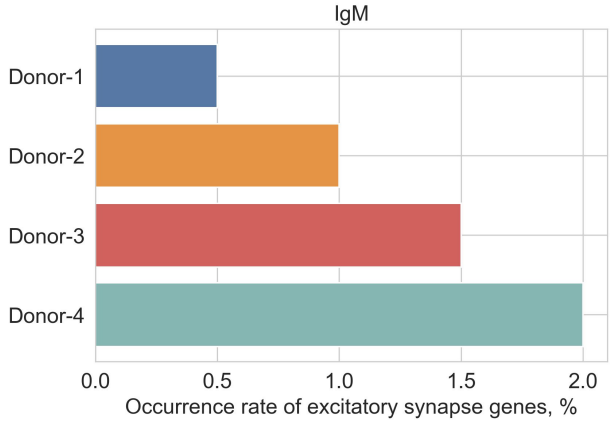
Plasma Membrane Gene Selection

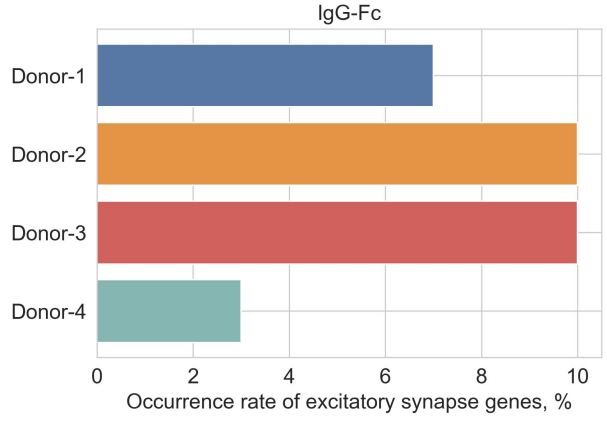


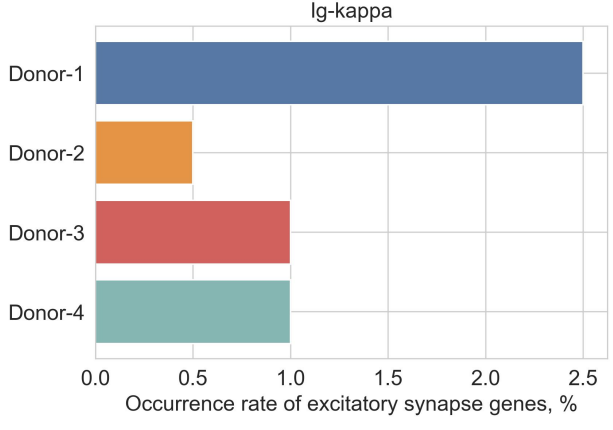


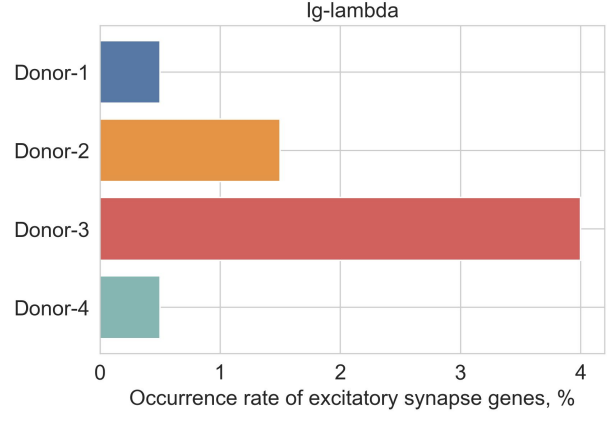
The top 20 positively correlated genes were selected for Donor-2, and for this list correlation coefficients were computed for the rest of the donors.

Antibody Binding and Gene Expression









* From the top-200 correlated genes

Thank you for your attention!