

The current mRNA pool in cells is not constant, it follows a natural dynamics. In addition, the composition of the transcriptome is strongly influenced by the rate of expression and degradation of gene products, as well as transcriptional responses to external and internal factors. These responses regulate the switch between the "on"- and "off"- state of genes. A single-cell RNA-sequencing data set containing normalized copy numbers of several mRNAs can be interpreted that each cell in the dataset is a time point (snapshot) in the dynamically fluctuating transcriptional processes in an as many dimensional space as the number of mRNAs revealed. The single-cell sequencing data are usually sparse matrices with several zero values. The common belief is that zero values are result of the gene transcription "off-state". To test this hypothesis we used our Patch-seq dataset sequenced to 22 million reads and recovered more than 19,000 transcripts from 84 cells of the frontal cortex in mice. Such an ultra-deep sequencing allows to investigate the complete transcriptome of the neurons. Neurons were divided to two sets, one was the Pyramidal cell group, and the other was the fast spiking interneurons group. By examining the presence and level of gene expression, we searched for groups of genes that help us understand the dynamics of the neuronal transcriptome, the pattern of "on"-state – "off"-state alteration of individual genes. Our results may lead to a better understanding of the phenotype of neurons at the transcriptome level and how changes in transcriptome dynamics lead to neurodegenerative symptoms.

Animals

Cell harvesting was made in acute slices, derived from male C57BL/6N mice aged between 27 and 40 days. All procedures were done under the EU Ethical Rules of Using Animals for Research Purposes (2010/63/EU revising Directive 86/609/EEC).

Brain slice preparation and patch clamp electrophysiology

Standard brain slicing technology was used to create 300µm coronal slices. Brains were removed from the isoflurane-anesthetized animals and placed in ice-cold artificial cerebrospinal fluid, cut by vibratome. Patch clamp electrophysiology measurements were performed using a standard setup built on Leica electrophysiology microscope DM6000 FS equipped DIC optics. The protocol started with targeting the selected cell and achieving GΩ seal resistance, followed by step-gradient depolarization method. We continued recording until the highest firing rate of the neurons, then the electrophysiological signals were amplified and the analogue data were digitized using a data acquisition device. Pyramidal cells were identified based on their apical dendrite, the larger triangular cell body and the regular firing of action potentials. To identify the fast-spiking interneurons, their smaller, spherical cell body and the high-frequency firing of narrow action potentials were used.

Amplification and sequencing

The cytoplasmic mRNAs of the investigated cells were isolated, individually reverse transcribed to cDNA, amplified by two-round aRNA amplification and performed into barcoded libraries for sequencing. Libraries were sequenced on HiSeq2500 to produce 100-base paired-end reads and also on NextSeq500 to produce 75-base paired-end reads. The data is available in GEO database under [GSE135060](#) accession number.

Comparison of data obtained by different sequencing methods based on the proportion of zero values

We had binarized the expression data, afterwards the summarized number of the zero values for each gene was divided by the total number of examined cells to get the proportion of zeros. We had excluded the genes that had only zeros, then put the remaining data in ascending order based on the values generated by the aforementioned process. The resulting data was divided to 10 equal sized groups and illustrated on separate boxplot for each dataset examined to demonstrate that our data show a general expression pattern and can be used for transcriptome analysis.

Examining the contribution of certain genes to the total detected copy number

Counts were normalized by the total copy number for each cell, plotted as a function of frequency.

Correlation analysis of gene expression pattern

The probability of co-expression for all possible gene pairs was calculated (using the formula below on the binarized data) to construct a similarity matrix:

$$p(x = k) = \binom{n}{k} q^k (1 - q)^{n-k}$$

where $0 \leq p \leq 1$, n : number of expression cases for gene 1, k : number of co-expression cases, q : number of expression cases for gene 2 / total number of cells. Based on the dendrogram generated from the matrix, we selected the 5 most representative groups including genes that show correlation in their expression pattern. The groups were subjected to GEO Enrichment analysis.

As shown in the figure, data obtained by Patch- and FACS- single-cell sequencing methods typically contain more zero values than bulk sequencing data. These zero values may indicate that genes are in off-state, in which case they are considered as biological zeros. This is the result of the fact that single-cell sequencing provides information about the expression pattern of the cell at a particular time point, while bulk sequencing works with pooled cell populations, therefore it shows an average expression level of the investigated genes. In other cases, the zero values can be non-biological zeros, suggesting that our detection method is inadequate to detect low-level gene expression. To exclude the possibility of poor quality of our measurement method, we compared our data with other single-cell sequencing data based on their zero distribution. The proportion of zero values for our data matches well with other Patch-seq data. Furthermore, the distribution is independent of the number of genes identified in the experiments.

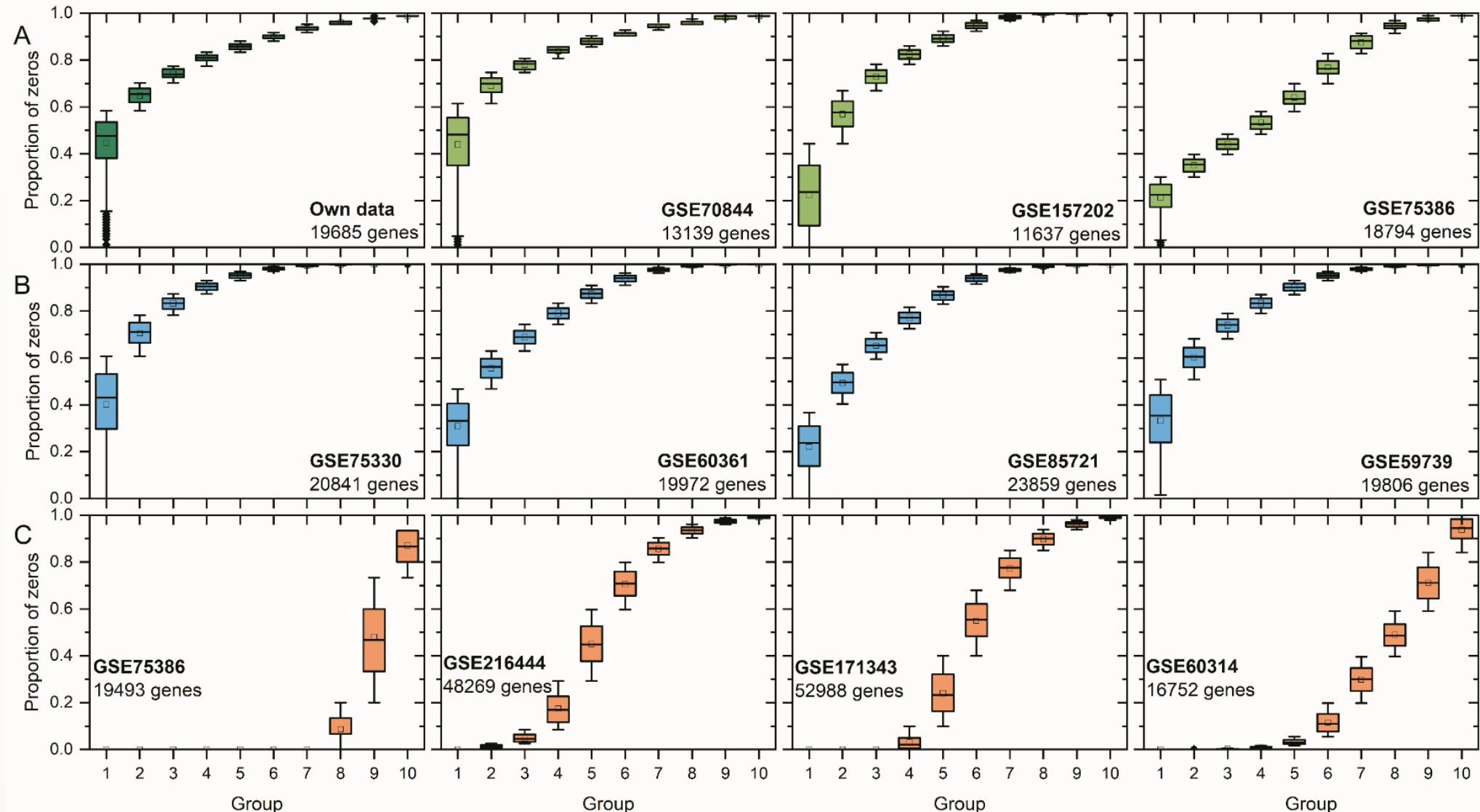
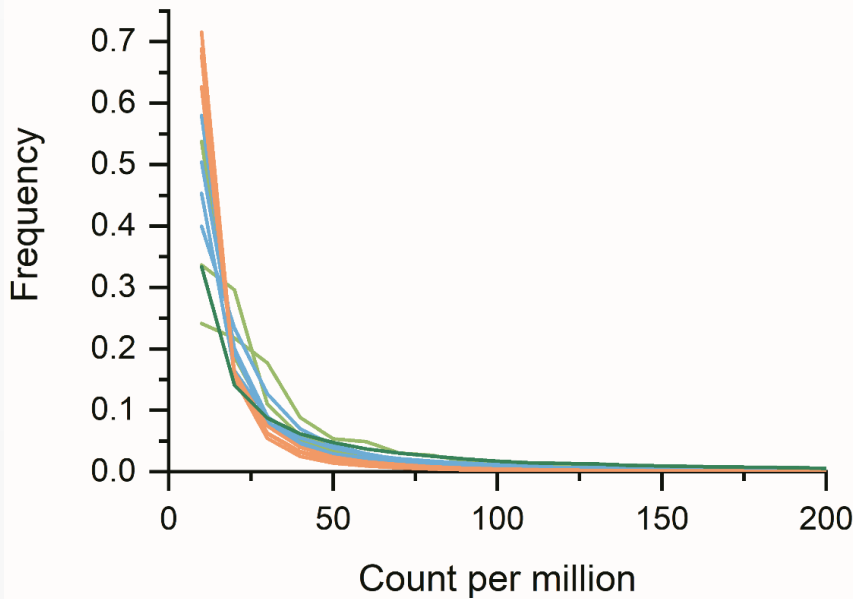


Figure 1. Comparison of zero value proportion in Patch-sequencing (A), FACS sorted-sequencing (B) and bulk sequencing (C) data broken down by 10 equal sized group. The boxplots show the distribution of genes that had values other than zero for each subsets. The number of genes tested in the experiments are also shown. The sequencing data obtained in mouse brain, except for GSE157202 and GSE60314 which are in *Drosophila melanogaster* and GSE 52988 which is in human.



As presented on the plot, the methods used to generate the tested datasets can detect gene products that occur in small copy numbers with similar sensitivity. The Patch-seq method we used fits the general pattern.

Figure 2. The relative frequencies were calculated as average distribution of normalized copy numbers. (The color codes are the same as in the previous figure.)

Figure 3. Results of GEO Enrichment analysis of groups based on expression pattern similarity indicated the most important /relevant biological process and its false discovery rate (FDR). The diagram of expression pattern and the network represented by String are also shown. Of the 19685 genes analyzed, 2787 genes were identified as part of one of the 5 networks.

Biological process	FDR
- Cellular aromatic compound metabolic process	0.00013
- Cellular nitrogen compound metabolic process	0.00062
- Base excision repair	0.00062
- DNA replication	0.00062

The proper function of DNA repair mechanisms in long-lived neurons is essential, because the absence of this process can lead to neurodevelopmental disorders and neurodegenerative diseases.

Biological process	FDR
- Mitotic cell cycle process	2.25e-11
- Cell cycle process	1.86e-10

Although neurons are post-mitotic cells, they also express different cell cycle-related genes. These genes are presumably involved in maintaining the post-mitotic state.

Biological process	FDR
- GPCR ligand binding	8.53e-09
- Response to stimulus	1.71e-07
- Signal transduction	4.51e-07
- Cell communication	1.11e-06
- Cytokine-cytokine receptor interaction	5.16e-05

Communication between neurons is mediated by various neurotransmitters and their receptors, including GPCRs. There is also an enrichment of genes involved in signal transduction and responses to stimulus.

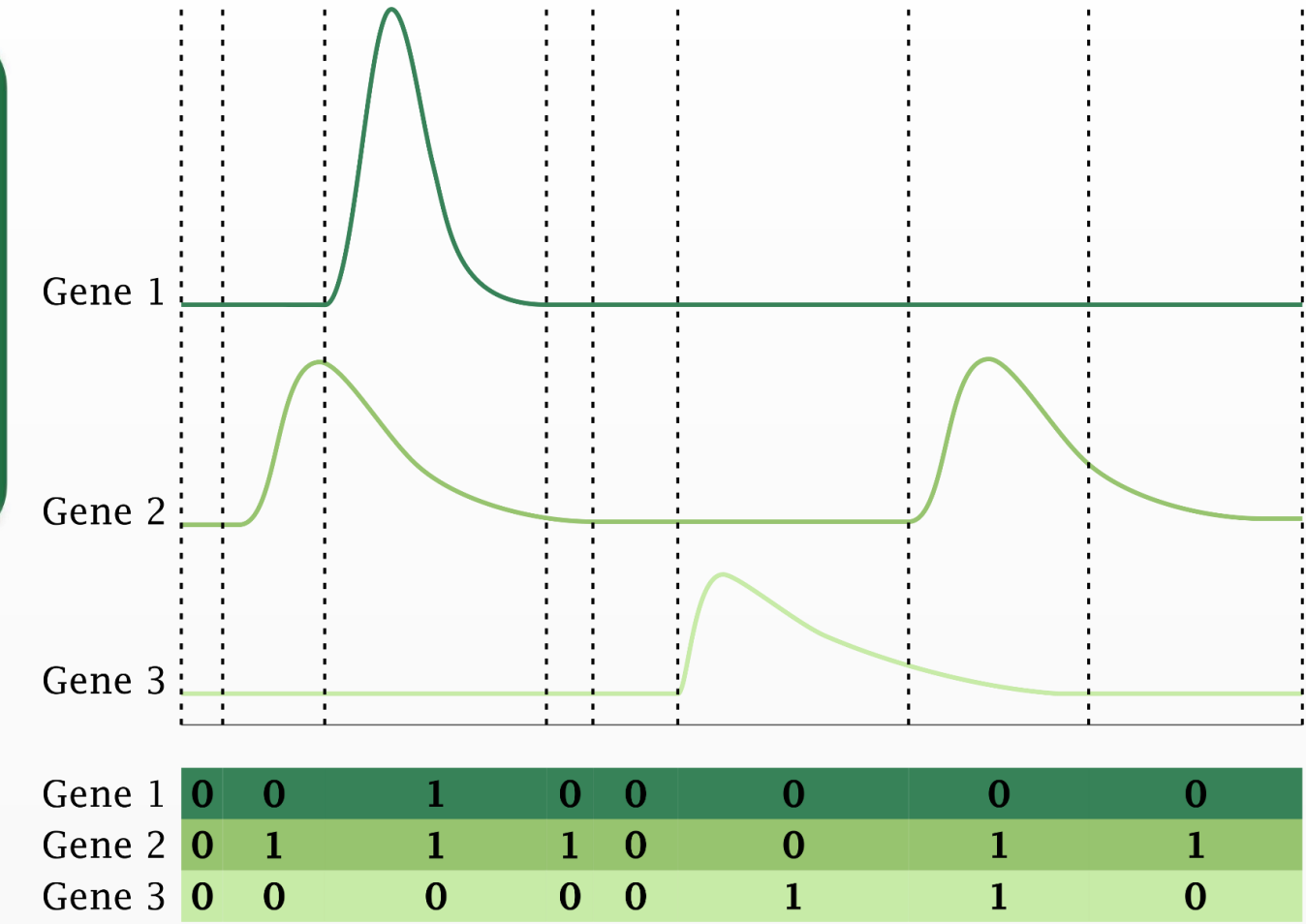
Biological process	FDR
- Inflammatory response	4.24e-08
- Cytokine activity	4.28e-06
- Receptor regulator activity	3.45e-05
- Response to organic substance	0.00034
- Response to external stimulus	0.0016

Because of the close cooperation between the nervous and immune system, enrichment is also observed for genes involved in the response to different stimuli.

Biological process	FDR
- Cytokine-cytokine receptor interaction	0.0029
- Organonitrogen compound metabolic process	0.0029
- Response to stimulus	0.0029

Cytokines, which play a key role in cell signaling, appear repeatedly in our GEO enrichment analysis, similar to genes associated with the immune response and DNA repair mechanisms.

- According to the examined datasets, it can be concluded that the Patch-seq and FACS sorted single-cell sequencing data are generally characterized by many zero values. Furthermore, that the zero values in the data are biological in origin, and indicate that the gene is switched off.
- Our Patch-seq method for detecting gene expression levels has similar high sensitivity to methods used in other studies for detecting gene products expressed at low copy numbers.
- Based on the results of the network analysis, genes with similar expression pattern form groups responsible for well-defined functions and not random.



In order to obtain information on the dynamics of gene expression, further studies will be needed. Bulk and single-cell sequencing data should be compared by interpreting the bulk data as an average, also assuming that the expression of all genes follows a lognormal distribution. Our aim is to determine the order of expression of genes, and a minimum time duration while the investigated gene is in off-state.

Figure 4. Schematic representation of a gene triplet that shows correlation in their expression pattern. Top: lognormal distributions model fitted for Patch-seq data. Bottom: Binary on-state, off-state matrix. The minimum length of a gene off-state can be determined for regions which show that at least an other gene is in its on-state.

Ravas, L. et al. *Cerebral cortex* 31: 731-745 (2021). Fuzik, J. et al. *Nat Biotechnol.* 34:175-183 (2016). Ravenscroft, T. *J Neurosci* 40:7999-8024 (2020). Földy, C. et al. *Proc Natl Acad Sci USA* 113:E5222-31 (2016). Marques, S. et al. *Science* 352:1326-1329 (2016). Zeisel, A. et al. *Science* 347:1138-42(2015). Habibi, N. et al. *Science* 353:925-8 (2016). Usoskin, D. et al. *Nat Neurosci* 18:145-53 (2015). Barry, A. et al. unpublished data (GSE21644). Bowles, K. et al. *Cell* 184:4547-4563.e17 (2021). Lin, Y. et al. *BMC Genomics* 17:28 (2016).

Acknowledgements: This work was supported by the National Brain Research Program NAP 2.0 and 3.0 of the Hungarian Academy of Sciences (NAP2022-1-3/2022; NKP-2017-00002) and the National Research, Development and Innovation Fund of Hungary (PD135510, K138937, FIEK_16-1-2016-0005), and InnoScience Ltd.