Michał Dąbrowski Ph.D. D.Sc.
Laboratory of Molecular Neurobiology
Nencki Institute of Experimental Biology
Polish Academy of Sciences

Review of the doctoral dissertation

Title: Cell-type resolution data analysis using different transcriptomics approaches for improvement of current methods

Author: Elyas Mohammadi, M.Sc.

The doctoral dissertation was prepared at the 3P – Medicine Laboratory of the Medical University of Gdańsk. The supervisor of the doctoral dissertation is Dr. Jakub Mieczkowski Ph.D., D.Sc.

The doctoral dissertation submitted for evaluation has the form of a collection of published and thematically related scientific articles, accompanied by summaries in Polish and English, and a self-report in English, in accordance with Art. 187 of the Act of July 20, 2018. Law on Higher Education and Science (Journal of Laws of 2018, item 1668).

The doctoral dissertation consists of three scientific articles: Mohammadi et at. "Size matters: the impact of nucleus size on results from spatial transcriptomics.", *J Transl Med.* 2023; Swatler et al. "4-1BBL-containing leukemic extracellular vesicles promote immunosuppressive effector regulatory T cells.", *Blood Adv.* 2022; and Mohammadi et al. "Improvement of the performance of anticancer peptides using a drug repositioning pipeline. *Biotechnol J.* 2022. These articles have a common theme of cell-type resolved analysis of trancriptomics data; namely data with spatial and single-cell resultion from the brain (paper I), bulk trancriptomics data from sorted primary immune cells (paper II), and public trancriptomics data from cancer cell-lines treated with hundreds of potential drugs (paper III).

Following Art. 186 of the above-mentioned Act, regarding multi-author publications, each of the works included in the dissertation is accompanied by statements of co-authors, which describe their individual contributions to the publication. In two of the three papers included in the dissertation, the doctoral student is their first author, which clearly indicates his leading role in these publications.

In the publication (Mohammadi et al. 2023) the doctoral candidate conducted the analysis of spatial transcriptomic data obtained from two regions of the human brain, namely: orbitofrontal neocortex and temporal neocortex, with each region examined on frozen sections of brain tissue from two deceased adult male donors. The specimens were provided by Harvard University and

Massachusets Alzheimer's Disease Research Center and the research was conducted in accordance with the consent of the local Bioethics Committee at Medical University of Gdańsk. The two brain donors were classified as "healthy", meaning that they were not diagnosed with Alzheimer's disease. The spatial transcriptomics data analyzed by the doctoral candidate were obtained in experiments performed by his collaborators, who used 10x Genomics Visium platform – a widely used spatial transcriptomics method. In this method, a tissue section is placed on a specially designed Visium matrix, which is a microscope slide with a dense hexagonal arrangement of spots of oligonucleotides printed on the slide, these oligonucleotides encoding in their sequence information about the location of a given spot on the slide. These oligonucleotides are used to prime the reverse transcription of mRNA, which is happening on the slide, thereby capturing spatial positions of the transcripts in the tissue section. The tissue sections are then dissociated from the arrays, and the cDNA is pooled and amplified. The resulting cDNA libraries are then sequenced, including the sequences of the primers, using next-generation sequencing technology, yielding a spatially-resolved gene expression profiles of every spot. In the Visium technology spots are 55 μm in diameter placed at 100 μm center-to-center distance. These sizes are larger than a typical cell nucleus/cell body. Therefore, one spot typically overlaps several neighboring cells, which can be of different cell-types. At the same time, spatial transcriptomics data, similarly to single-cell or single-nucleus RNA-seq data, contain missing data for many transcripts. For both these reasons, optimally assigning an expression profile of a Visium spot to cell-type(s) is not trivial.

For the research described in the work (Mohammadi et al. 2023), it was crucial that in the spatial transcriptomics experiment each of the examined brain regions was represented by two consecutive frozen sections. The doctoral candidate analyzed the data obtained separately from each of these sections and made an interesting observation that the assignment of spots to cell-types was much worse for neurons than for the two remaining analyzed cell-types, namely astrocytes and oligodendrocytes. Interestingly, no such difference was reported before for the spatial transcriptomics data obtained on the same Visium platform for the mouse brain. The doctoral candidate then investigated the effects of using the standard integration method available in the Seurat software for integration of data from two consecutive slices, which was named Consecutive Slices Data Integration (CSDI). The doctoral candidate demonstrated that performing CSDI resulted in a large improvement of spot assignment to neurons, and a better agreement of the spot clustering results with the histological architecture of the cortex. The Authors hypothesized that the observed difference may be related to the

fact that in human (but not in mouse) the average size (about 20 μm) of a nucleus from a gray-matter neuron is much larger than the thickness of the tissue sections (10-12 μm). Consequently, the RNAs from a single neuronal nucleus are split between two consecutive slices. The Authors suggest that this may lead to missing rare neuron-specific transcripts in single slice data, which becomes corrected by Consecutive Slices Data Integration. As rightly indicated in the publication (Mohammadi et al. 2023), the use of standard integration method available in the Seurat software for Consecutive Slide Data Integration was reported before by the creators of this software. Out of reviewer's duty I checked that the previously reported integration was performed on data from a different spatial transcriptomics platform, namely STARmap, and for a different purpose, namely data aggregation, and that it was not followed by assessment of effects of the integration on cell-type assignments or clustering, which underscores the novelty and importance of the results obtained by the doctoral candidate.

In the publication (Swatler et al. 2022), the doctoral candidate performed an analysis of bulk RNA-seq data from sorted populations of human T regulatory lymphocytes (Tregs) treated with extracellular vesicles, followed by analysis of overrepresented transcription factor binding sites motifs in the promoter regions of differentially expressed genes. The results of these analyses showcased the usefulness of bulk RNA-seq combined with cell sorting for analysis of cell populations defined by their cell-surface antigens.

In the publication (Mohammadi et al. 2021), the doctoral candidate designed and implemented a data analysis pipeline for drug repositioning, which is leveraging on the massive public transcriptomics data generated by the LINCS L1000 NIH project. This project probed the response of about a hundred cancer cell lines to over 900 perturbing agents, including many approved drugs. The motivation for developing this pipeline was the desire to identify approved drugs that could be repositioned (repurposed) for a combination therapy together with with anti-cancer peptides. In their publication, the Authors pointed to the results of previous studies indicating that negatively charged heparan sulfates and chondroitin sulfate on the cell surface can bind anticancer peptides and in this way inhibit their cytotoxic effect on cancer cells. Hence, drugs affecting cell surface levels of these glycosaminoglycans may have a potential to enhance the effects of anticancer peptides. To identify such drugs, the doctoral candidate, using the data for four arbitrarily chosen cancer cell lines and 900 perturbing agents, first, identified genes whose expression was highly correlated expression with 32 genes known to affect the metabolism or trafficing of heparan sulphate or chondroitin sulphate. As an earlier preparatory step, the doctoral candidate investigated, which of the three correlation coefficient (Pearson's, Spearman's, or

Kendall's) should be used to yield most functionally relevant correlations. Having identified the genes that are highly correlated with the known genes involved in the metabolism of heparan sulphate or chondroitin sulphate, the the doctoral candidate analyzed the effects of 900 perturbing agents on expression of these genes, which enabled him to identify six drugs that consistently down-regulated the expression of these genes in all four studied cancer cell lines.

The correctness and originality of the papers constituting the dissertation were confirmed by the reviewers at the stage of their publication, therefore my comments concern mainly the self-report.

The self-report is written in English, but inclues mandatory abstracts in both Polish and English. The main part of the self-report has a typical structure. It starts with an "Introduction" (13 pages), followed by: "Aims" (one page), "Materials and Methods" (5 pages), "Results" (5 pages), "Conclusion" (one page), "References" (54 items on 4 pages), and ends with the section "Publications" containing the papers themselves. Each of the above sections is divided into subsections corresponding to the three publications included in the doctoral dissertation. Every paper includes a section "Author contributions" or "Authorship" that precisely describes the contribution of every author to a given publication. The self-report adequately describes the rationale for the research and the content of the three articles forming the doctoral dissertation. Nonetheless, as a reviewer, I have the following comments:

The Aims on page 25 are formulated in a language that is too technical. Their shorter formulation in the Abstract is equally informative and easier to understand.

In the "Introduction", the subsections: "Immortalized cell lines" and "Primary cell sorting" are not necessary, as they were not in the focus of the doctoral candidate's work for the dissertation. The subsection entitled "Challenges and solutions" is particularly interesting and well-written. At the same time, I am missing in the "Introduction" the part(s) that would describe the analytical methods used in the dissertation. In particular, I am missing a description or an illustration of the mathematical steps of the integration method implemented in the Seurat software used by the doctoral candidate. This deficit could potentially be corrected (time-permitting) during the defense.

The section "Materials and Methods" is well-written and provides sufficient details. Notably, two papers (Mohammadi et al. 2022, 2023) are accompanied by the code that was used to conduct the analysis provided as open-source in a public repository. In the "Results" section the doctoral candidate presents, and also carefully discusses the results of each paper, wheras in the "Conclusion" he links them back to the aims of the dissertation. The language of the self-report is clear and precise and it is

carefully prepared in editorial terms – I have found few cases of incorrect or clumsy wording (e.g. "Normal cell lines" on page 11, "sequencing of treated and untreated cell lines" on page 13), or typographic mistakes (e.g. "nucelli", instead of "nuclei", on page 9).

I have the following questions to doctoral candidate:

1) After integrating data from two consecutive slices, is it possible to check the spatial location of the corresponding spots (anchors)? If so, did the doctoral candidate attempt to check whether they map to the corresponding positions of each slice? (I am not asking for the results of such an analysis, unless it has already been performed).

2) The observation that integration improves spot assignment to neurons specifically in humans is very intriguing. The possible explanation put forward by the Authors impies either a fluctuating detection of rare transcripts (a technical issue) or non-homogenous distributions of some transcripts within the nucleus (a biological reason). Is there evidence for either, or both, in the literature?

The aforementioned comments do not affect my assessment of the substantive results presented in the dissertation, which I classify as satisfactory. I am therefore pleased to announce that: The doctoral dissertation of Elyas Mohammadi submitted for evaluation meets the conditions set out in Art. 187 of the Act of July 20, 2018. Law on Higher Education and Science (Journal of Laws of 2018, item 1668)[1]. Therefore, I request the Pharmeceutical Sciences Council of the Medical University of Gdańsk to admit Elyas Mohammadi to further stages of the procedure for conferring the degree of doctor of medical sciences and health sciences in the discipline of pharmaceutical sciences.

Michał Dąbrowski

Warsaw, 9.01.2024

1 Przedstawiona do oceny rozprawa doktorska Elyasa Mohammadi spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668).