ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Robust cell subsets decomposition from tissue expression profiles for biomarker identification**
Author's name:	Evžen Šírek
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Petra Hrubá
Reviewer's department:	Transplant laboratory, Institute for Clinical and Experimental Medicine

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **extraordinarily challenging**

Evaluation of thesis difficulty of assignment.

This thesis aims to apply a rather new approach of deconvolution of gene expression profile of globin depleted peripheral blood samples and used them in DEG analysis to reveal biomarkers of operational tolerance in kidney transplantation. It is quite challenging as most methods were developed mainly on microarray data, and in this master thesis it was necessary to use them on data from RNA sequencing which requires different normalization. Furthermore, data were obtained by RNAseq of whole blood samples and not on isolated peripheral blood mononuclear cells. Author used 10 publicly available methods of deconvolution, compared their performance and used them to perform DEG analysis.

Satisfaction of assignment **fulfilled**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

Author managed to fulfill submitted assignments. He described all available methods of deconvolution, selected 10 most suitable for RNAseq data from whole peripheral blood and applied them on provided data. Besides original aim to perform deconvolution analysis, the author performed complete data alignment as provided data matrix did not contain information relating to the alignment to genome and checked quality of mapping. Deconvolution results (fractions of B cells, CD4 and CD8 T cells and NK cells) were used as another input for DEG analysis whereas 4 methods of DEseq2 were compared**.** The number of identical genes among top 100 ranked DEG between those 4 methods ranged from 38-61%. Identified top-ranked genes were further compared by gene set enrichment analysis. Significant top ranked GO terms were similar between different models. Adding fractions of particular type cells did not change top ranked GO terms, however, led to p-value increase.

Method of conception **outstanding**

Assess that student has chosen correct approach or solution methods.

Author selected the most suitable publicly available methods of deconvolution originally developed either on microarray or RNAseq data and applicable to tissues, PBMC or blood and applied their performance on provided data from RNAseq of whole blood. He checked the correctness of selected methods on several levels: checked quality of mapping, the number of top identically differentially expressed transcripts and genes after aggregation of transcripts to genes and the number of top differentially expressed genes and top GO terms after taking into account the results of deconvolution.

Technical level **A - excellent.**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

In this master thesis, author performed thorough literature research and applied ten of the deconvolution methods on RNAseq data of peripheral blood. He performed complete data alignment to Human genome. He compared all used deconvolution methods and also performed DEG analysis which incorporated the deconvolution results. All these analyses require substantial knowledge of using bioinformatics tools in R software which author of this thesis managed at high technical level.

REVIEWER'S OPINION OF FINAL THESIS

Formal and language level, scope of thesis	**B - very good.**
The presented thesis is written in comprehensible English, typographically well-organized with only minor typographical mistakes.	

Selection of sources, citation correctness	**A - excellent.**
Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.	
References are properly cited, student used appropriate sources of literature regarding introduction to molecular biology and RNAseq methods, as well as for selection of used bioinformatics methods.	

Additional commentary and evaluation
Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

The research undertaken is clearly described, the analyses and conclusions drawn from the research are well-justified and accurately referenced. The implications and limitations of the research are fully discussed. In conclusion, author of this master thesis compared different methods for deconvolution of gene expression applied for whole blood RNAseq data and showed that results are quite robust especially in the terms of gene set enrichment analysis.

Questions for defense:

1/ All deconvolution methods needs count matrix on gene level and not on transcript level, it was necessary to somehow aggregate all transcripts coded by one gene. Author correctly tried to find the best method and also checked the overlap between top deregulated transcripts and genes. He found that the overlap was quite low which may raise the question whether the method of transcript aggregation to gene level is biologically appropriate. Some genes have up to 27 transcripts variants, some have only two of them. Another problem is that not each transcript represents a fully biologically protein coding sequence. Therefore, the method of transcript aggregation is of great importance and author did not use any simple method (as for example sum of all transcripts for one gene) but used tximport tool of R library that takes into account effective length of transcript. Can you, please, describe in more detail how tximport tool deal with transcript aggregation?

2/Do you think it would be possible to aggregate only protein coding transcripts to genes?

3/ In provided dataset deconvolution did not reveal differences in particular cell types proportion between OT and CR group. Do you think that in case of significant differences in particular cell type, this will significantly affect results of DEG where deconvolution result is used as one input variable?

I evaluate handed thesis with classification grade **A - excellent.**

Date: **1.6.2020** Signature: