Systems Biology Laboratory
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
Phone: (+420) 549 494 476
Fax: (+420) 549 491 820

From: David Šafránek, Ph.D.
Assistant Professor

e-mail: safranek@fi.muni.cz

Review of PhD Thesis of Ing. Ivan Vogel

Herewith I provide the review of PhD thesis "Single Cell Genotyping" of the doctorate candidate Ing. Ivan Vogel (BUT FIT). The review is structured with respect to the individual evaluated aspects that lead to the final assessment presented at the end of the review.

Thesis Topic

The thesis of the PhD candidate targets a very important and relevant bioinformatics topic of DNA genotyping at the single cell level. This topic is practically motivated by the needs of explaining the rapidly increasing number of discoveries on living organisms. Finding satisfactory answers to many of arising biological questions cannot be done without zooming into the structure and behaviour of a single cell. Exploring the cell information basis contained in DNA is a fundamental procedure that will enable further steps toward comprehensive understanding of living cells behaviour. Significance of this research targets, e.g., development of anomalously behaving cells such as cancer cells.

The goal of the reported research was to develop a quite comprehensive framework supporting the pipeline of genotyping starting from single nucleotide polymorphisms (SNP) array samples through several levels of data processing to automatised cross-validation of the obtained results. On the one hand, coverage of this sequence of goals was necessary in order to succeed with bringing a novel contribution to single cell genotyping. On the other hand, the need to cover all of these steps naturally displays the dissertability of the non-trivial research question to be solved.

In general, the topic of the thesis is definitely very interesting and in-line with current needs of bioinformatics. To the best of my knowledge, there are not yet many working solutions addressing the problem. The most accepted existing (not optimal) solution is mentioned in the thesis as the related work and compared against the new results of the candidate. In my opinion, the topic of this dissertation, especially the uniqueness of the developed method, confirms originality of this dissertation.

Research Contribution

The main research contribution of the dissertation under review is an algorithmic framework for genotyping from SNP array data. In general, the novelty is in bringing a working solution to the problem of single cell genotyping for which robust solutions do not yet exist. The candidate succeeds with his approach by demonstrating his algorithms working on two different datasets.

The contribution has been published in a highly impacted journal Bioinformatics (2019). The candidate acts as the first author of the paper with 90% contribution thus proving his work is fundamental for this research. It is worth noting that the framework has been recently applied in a strong case study bringing new insights to chromosome errors in human eggs (a paper published in Science – Scientific Reports). The candidate acts as a co-author of that paper with 10% contribution.

In conclusion, there are four papers published by the candidate until now, all of them are journal papers that have very high impact (IF>3.7) thus proving the quality of his contribution to the research community.

Thesis Text

The text of the thesis is organised as a self-contained complete description of the research published in the Bioinformatics article extended with some additional work. The text is structured in nine chapters.

Chapter 2 describes briefly the biological context of genomics with a special emphasis on heterogeneity observed at the single cell level. Additionally, the general technical workflow of processing the genomics data is described. My only issues with this part are that some of the shortcuts are presented before they are introduced (e.g., WGA). It would be of great help to see an overall list of all shortcuts used in the thesis.

Chapter 3 is dedicated to processing of the genomics data. Since the thesis works with SNP and NGS technologies both are described in sufficient detail in this technical background section.

Chapter 4 targets the background for the computational part of the thesis. In particular, relevant machine learning techniques are described here. I have found the description of the techniques very relevant and sufficient not only for the purpose of the thesis but also for getting a general overview. Final sections are devoted to existing genotyping algorithms giving the state-of-the-art of the topic. The chapter is closed with an interesting comparative study of the existing techniques giving thus clear motivation for a novel contribution.

Chapter 5 describes the original work published in Bioinformatics. The main result is a novel algorithm for filtration of the noise from the genotypes obtained from single-cell data. The method is unique in combining several machine learning techniques. In particular, the method adapts Random Forests to identify the noise in the SNP data and Gaussian mixtures method is then used to classify the noise regions. The description is mainly devoted to the methodology, a minor part makes a description of the algorithm. My only issue with this chapter is the fact that the flowchart at page 52 is hardly readable (it could have been typeset in landscape orientation).

Chapter 6 describes precise validation experiments. The content of this chapter is also published in the Bioinformatics paper. In addition, there is a section describing the set of experiments showing the sensitivity of the parameters and quality of the training data used. This section has not been published yet, however, it supports the practical utility of the developed method.

Chapter 7 describes the steps necessary for interpretation of the results achieved with the previously defined pipeline. In particular, the candidate presents several algorithms that are optimisations and adaptations of existing algorithms. The goal is to apply the techniques to data from a specific domain – meiosis of female reproductive cells. This chapter thus gives a necessary technical basis for application of the method.

Chapter 8 is devoted to application of the pipeline to real data and validation of the results with NGS. Although the chapter is rather short in comparison with previous chapters it gives a good evidence of practicability of the developed method.

Chapter 9 provides conclusions and several comments on future work.

In general, the text is well-organised and gives enough information allowing the reader to systematically understand the contribution of the candidate. It can be considered as an extended summary of the research contribution. The content is sound and complete. The language quality is acceptable, however, I would suggest some improvements if it is still possible (a careful reading by a native speaker would help in that case). There are several typos and typographical issues that could have been avoided if a better software such as latex have been used for typesetting.

Overall Evaluation

To sum up, the research presented in the thesis displays the following attributes: a non-trivial problem very relevant for the current needs; a rigorous and precise work of the candidate giving an interesting and original solution to the problem; well-described research outcomes published in highly-impacted journal papers; crucial role of the candidate in conducting this research. The thesis meets the standard requirements imposed on a dissertation thesis in the field of bioinformatics. Based on the facts mentioned in this review, I am pleased to suggest to the committee to accept the thesis with the grade A and nominate the candidate for the title Ph.D.

Brno, 20th January 2020 RNDr. David Šafránek, Ph.D.

Questions

1. How the pipeline performs (in terms of computing time) on a typical set of real data?
2. Identify the potential bottlenecks in the computational efficiency. Is the technology already prepared for high-performance hardware or what needs to be done to improve scalability if that is an issue?
3. Where do you see places for further improvements in both quality of the results and efficiency of the computation?
4. What was your exact contribution to the Science paper? Can you comment on experiences you have learned?