Examiner report dissertation Pawel Kus

"Models of cancer genome evolution used to evaluate the role of selection and occurrence of new mutations"

Peter Van Loo, 18/08/2023

General remarks

Pawel Kus presented a dissertation on modeling cancer evolutionary parameters using multi-sample exome or genome sequencing data of four cancer types: acute myeloid leukemia (AML), breast cancer, laryngeal cancer and bladder cancer. Overall, the thesis is well-written and the candidate demonstrates a good knowledge of the relevant literature. The candidate leverages VAF (Variant Allele Frequency) distributions to model both subclonal expansions (binomial 'peaks') and neutral evolution ('neutral tails' showing a power-law shape). The candidate's work builds upon previous work by Mark Williams, Trevor Graham, Andrea Sottoriva and Giulio Caravagna on modelling neutral evolution from VAF distributions of mutations in cancer. Key innovations are: (i) the development of a novel R package 'cevomod' that fits a mixture of a neutral tail and or multiple subclones, which extends the functionality of the Mobster method at considerably improved computational efficiency; (ii) a detailed exploration of fitting the exponent of the neutral tail power law, allowing deviation from its theoretical value of 2; (iii) detailed application to breast, laryngeal cancer and AML in the context of progression, metastasis and relapse (2 samples per patient), as well as application to novel unique multi-sampling data of bladder cancer, capturing multiple stages of the disease (including precursor lesions).

Title of the thesis

The title of the thesis is appropriate and captures the work performed, which is development and application of approaches to model tumor evolution, and specifically neutral evolution vs. selection.

Assessment of the structure of the thesis and information on its various components

The thesis is clearly and logically structured. It starts with a brief introduction setting the stage and outlining the thesis. The second and third chapter provide a well-written, detailed and well-motivated introduction to the field of cancer genomics, demonstrating the candidate's knowledge of the relevant literature. The second chapter focusses on the biology, while the third chapter focusses on the bioinformatics and computational modelling of subclonal expansions and neutral tails. This is followed by a chapter outlining the data and methods used and developed. These are clearly described and well-motivated. The fifth chapter describes the results of the data analyses performed, in two parts: (i) analyses of two samples each for AML (whole-genome sequencing data), and breast and laryngeal cancers (whole-exome sequencing data), and (ii) analyses of multi-region (many regions) sequenced bladder cancer data, capturing multiple stages of the disease. The latter part was a collaboration with the group of Bogdan Czerniak with part of the candidate's work included in a published collaborative paper, and part unpublished. The results are biologically interesting and broadly make sense (though I have reservations about possible overcalling of mutations in the breast and laryngeal cancer data, which may confound some the results, as detailed below), and the chapter overall is well-structured and well-written. The thesis ends with a summary/discussion chapter, which is brief and to-the-point.

Assessment of the literature cited

The thesis contains an extensive reference list, which captures the field very well. The candidate has a good knowledge of the relevant literature.

The one paper I perhaps would have liked to see one or two lines of comments on, in the context of the bladder cancer data, is Lawson et al., Science 2020, PMID 33004514.

Evaluation of the purpose of the candidate's thesis paper

The objective of the thesis is well-defined and well-described. The work is clearly motivated.

Evaluation of the research methods used

The author uses cancer genomics data analysis approaches, including mutation calling, mutation signature analysis, and a novel subclonal reconstruction method (cevomod) to model VAF-distributions and thereby parameters of tumor evolution. The cevomod method is well thought-through and seems to work well. One key innovation of the method (as described above) is that it can model neutral tails with exponent (alpha) differing from 2. The methods used are appropriate for the questions asked.

As potential future work, it would be interesting to be able to formally statistically test the null hypothesis alpha=2. Knowing which deviations from the theoretically expected value are significant would give important insights that could help pinpoint biological and/or technical reasons for such potential deviations.

Evaluation of the candidate's discussion of his work

As detailed above, the discussion is brief and to-the-point. In addition to the discussion section, the candidate also includes appropriate discussion of his findings in the results chapter.

There is one major and a few minor points on which I would have liked to see a bit more discussion:

The major point is the rather surprising finding of absence of clonal mutation peaks in the breast and laryngeal cancer data. Multiple pan-cancer studies find such clonal peaks in virtually all cancers of all cancer types. Therefore, this result does not make sense and a critical discussion on what potential technical factors could contribute to this would be important to include in the thesis. I believe this ties in with two factors: (i) the very high mutation rates observed in these exome sequences, which are an order of magnitude higher than observed in other studies, and (ii) the modeling of VAF distributions rather than Cancer Cell Fraction (CCF; correcting for copy number changes). The candidate has good reasons to model VAF as copy number calling on exomes is challenging, yet the limitations of this should be critically discussed. It seems both a potentially very large number of false positive mutation calls and the 'blurring' of VAF

distributions due to copy number changes (not observed in AML as AML has very few CNAs) result in clonal peaks being merged with neutral peaks and (importantly) mutation artifacts.

Minor points:

- The mutation signature analysis on the bladder cancer data brings up a few unexpected signatures, including a signature of UV exposure. As bladder cells are not UV exposed, this result should be discussed critically. I also wonder if these patients have really been exposed to alkylating agents (which is a specific chemotherapy), and to aflatoxin and aristolochic acid (both are rare).
- Related to the major point above, I would have liked to see discussion on the influence of false-positive and false-negative mutation calls on the identification of neutral tails. While every tumor should have neutral tail mutations, on some tumors sequenced at limited coverage, these may not be detected. Vice-versa, false positive mutation calls may end up looking very much like neutral tails, potentially with somewhat different exponents (which could potentially be very interesting in the light of the candidate's main findings).

Practical application of the research results obtained

The cevomod package will be an important research tool for the scientific community, allowing further studies into tumor evolution.

Information on possible irregularities

No possible irregularities were found.

Assessment whether the dissertation provides an original solution to a scientific problem

As detailed above, the cevomod package is original and novel, and provides an important extension to the work of Caravagna and colleagues.

Assessment as to whether the doctoral dissertation demonstrates the candidate's general theoretical knowledge in the discipline and ability to conduct scientific work independently

As detailed above, the candidate displays a good knowledge of the relevant literature and has developed an interesting and scientifically relevant novel method. This testifies to his theoretical knowledge of the discipline and his ability to conduct scientific research independently.

Candidate's individual contribution to joint work

The candidate's method development work (chapter 4) and analysis of AML, breast cancer and laryngeal cancer (chapter 5 part 1) are entirely the candidate's own work. The bladder cancer work (chapter 5, part 2) contains both joint work with the group of Bogdan Czerniak (here, the candidate's contribution is well-defined and he focusses the work he describes on his own work), as well as further analyses performed by the candidate on the same data.

Minor comments and typos

- Gene names should be italicized throughout the thesis, as well as Latin terms (*de novo*, *in silico*, *et al.*, ...)
- (Somatic) copy number alterations are distinct from (germline) CNVs – I would use the term CNA for somatic events and avoid the term CNV, which usually refers to germline variants
- A few figures are not referred to in the main text (fig. 2.3; fig. 5.16; fig. 5.23)
- p. 2: "breast cancer (BRCA) and laryngel cancers (BRCA)," -> the last "BRCA" should say "LSCC"
- p. 7: "tumor suppressor genes are RB" -> "*RB1*"
- p. 8: "MYC, gene most often mutated in ovarian and uterine cancers" -> I would say "amplified" rather than "mutated" as *MYC* mutations are quite rare.

- p. 12 mentions chromothripsis in one sentence with the big bang theory, while chromothripsis events are punctuated events that are quite distinct from the big bang model. I would leave out the reference to chromothripsis, or research it more thoroughly.
- p. 18: "(HTS)[51]" -> add a space
- p. 19: "(WGS) ." -> remove space
- p. 19: "target sequencing" -> "targeted sequencing"
- p. 21: in the single-cell sequencing section, I would add a few sentences on the recent DLP/DLP+ and ACT methods
- p. 24: N_alt and N_ref: format analogously
- p. 24, eq. 3.4: CN_nut -> CN_mut, and one . should be + in the denominator
- p. 25: text below equation 3.8 describes M, b and c -> should probably be K, a and b.
- p. 29: lambda = b + d Cells -> add period after d.
- p. 48: I would be a bit more nuanced in the description of BRCA subtypes: basal-like breast cancer and triple-negative breast cancer are not synonym (though they are largely overlapping groups). The molecular subtypes in table 5.1 are transcriptionally defined and don't perfectly match with the hormone receptor status.
- p. 49: aliasing phenomena -> phenomenon
- p. 53: fig. 5.6 and 5.5 are referred to out of order
- p. 66: rate(Fig. 5.15) -> add space; particularly step -> steep
- p. 83: the distinguishing -> distinguishing

Conclusion

Pawel Kus has presented an overall well-written and well-rounded thesis. He has a good understanding of the literature and has produced original work. I highlighted a few minor and one somewhat more substantial concerns above, which can all be straightforwardly addressed in additional discussion. Overall, this is a deserving dissertation.