UNIVERSITY OF MINNESOTA

Duluth Campus | **_Department of Biomedical Sciences_**
Medical School Duluth

217 Med
1035 University Drive
Duluth, MN 55812-3031

Office: 218-726-7911
Fax: 218-726-7937
http://www.med.umn.edu/duluth
Email: biomed@d.umn.edu

May 2015

Review of Vojtěch David Master Thesis, University of South Bohemia

Overall organization: This strong Master's thesis centers around techniques for sequencing and organizing reads from a U-indel edited mitochondrial transcriptome. The Introduction focuses exclusively on a description of U-indel editing and in the last paragraph leads one to understand that Parts I and II will serve as a guide to methods the candidate designed to adapt existing transcriptome technologies and analysis software to U-indel editing transcriptomes.

The project's main objective is rather ambiguous. The term "analysis of U-indel editing" is used several times. The reader is left to guess exactly what that means. Does this mean determining the complete complement of editing sites in the encoded transcripts? Determining the relative efficiency of editing of the various gene products? Determining if products can be alternatively edited to code for different products? All of the above? Spelling this out may make it easier to understand the application of the modified and new software solutions developed to the two very different projects described (*T. brucei* and *Perkinsela* projects).

As the Introduction relates to U-indel editing, concerns are fairly specific (see below). However, the real topic of the thesis is development of a technology. Therefore, an introduction to exactly what next-generation (Illumina) sequencing is and what steps are involved a standard analysis workflow is needed. This is especially true if the reader is to understand why transcriptomes containing reads not present in the original genome pose an analytical challenge.

From the title, it was not apparent that in Part 1 two separate projects would be presented, one involving an organism that is not *Perkinsela*. The treatment of software development for U-indel modified transcriptomes using two different datasets is awkward and confusing as written. Topic appears to waffle between *Perkinsela* sequencing, iCLIP sequencing, and concepts to consider in general. Hard to determine when a sentence refers to either of the projects specifically or in general to adaptation of these technologies to U-indel trancriptomes. Maybe a different organizational structure may work better?

Part 1 would be easier to understand if the RNA editing analysis flowchart was presented right at the beginning so the reader understood the steps that would he/she will be walked through in the course of that chapter.

The Discussion is not an evaluation of the work and its implications for the field. Rather, it is a description of future directions, namely a proposal for a formal workflow for a U-indel editing solver. This solver proposes to replace a modified Bowtie2 analysis used in the actual thesis work with seeding and alignment using T-less reads. If this section is to remain a proposal for a future workflow rather than a true discussion, it should be identified as such. Also, Figure 7 would then benefit greatly from an indication of what boxes in the flowchart have been worked out in the two projects described, and which ones are new or should be improved.

Overall experimental: Since the Part 2 manuscript will very likely be the first published study to organize Illumina reads to make conclusions about transcriptome-wide degree and variation in U-indel editing, this work is very important. It also appears in general rigorous.

The main concern is the generalizations and the lack of specificity of reported results in Part 1. For example, how much shorter are iCLIP reads? In each case, what is the "reference sequence" for seeding? Is it absolutely necessary to have a reference sequence? What is meant by read loss "counting twice" for heavily edited U-indel data? When two solutions to the seeding problem were presented, which was utilized and why? What is meant by Bowtie working in "the traditional way" and "respecting" read length? How was rejection threshold ultimately determined in Bowtie, could not understand it. What exactly is a "pre-defined" seeding region for T. aligner? Exactly how small are read sets generated by iCLIP? What does it mean for a fully edited sequence to be "loaded" in T-aligner. Etc. Most importantly, what are THE CONCLUSIONS you are taking from this analysis you just described? What were the endpoints? Especially ambiguous for the iCLIP analysis.

Specific concerns:

- Mention of anchor regions in gRNAs would be good context for Introduction.
- Introduction paragraph 2 re: transcripts not edited over their entire length, are you referring to RNAs that are in the process of being edited/have experienced aborted editing, or are you referring to RNAs that have only a small edited region? If it is the latter, then it is not the region that it typically edited. *T. brucei* is the best studied model system and the MURF2 and CYb genes have their editing regions almost at the very beginning (5' end) of the mRNA. Furthermore, when referring to the early-branching kinetoplastids in that paragraph, it might be informative here to say "including *Perkinsela*".
- iCLIP, Bowtie, reference sequence, and multiple terms in Fig 7: define.
- Discuss limitations of poly(A) selected total RNA as input for studies of kinetoplast mt transcriptomes.
- Figure 2: What are the blue wavy lines in the far left image in Fig2? Each image should have a letter associated with it that is referred to in the legend text.
- It is very confusing that the concept of a "junction" region between the pre-edited and edited portions of the gene, and its potential origins, is not directly introduced either in the thesis or the paper. Throughout Part 1 and the manuscript, it appears that the definition of "alternatively edited" that is used is different from what others in the field may be defining it as. Maybe a few sentences in the Introduction would clarify this.
- A correction to the manuscript: Introduction: 9S and12S transcription is NOT developmentally regulated. The differences in 9S and 12S stage-specific RNA abundances were found to be likely a result of differential stability, not differential transcription, in Michelotti et al 1992 (same lab as Adler 1991).
- Have someone proofread again both for English usage, grammar (consistent tenses) and typos.

Conclusion:

The study performed is of high quality. Once the points stated above have been attended to, it shall certainly fulfill requirements for a master thesis, comparable to or exceeding what is expected for our Integrated Biosciences Master degree at the University of Minnesota.

Sincerely,

Sara Zimmer

Sara Zimmer, Ph.D.
Assistant Professor
University of MN Medical School Duluth
Department of Biomedical Sciences

Review of the Master thesis "**High-throughput analysis of uridine insertion and deletion RNA editing in *Perkinsela***" by Vojtěch David

In the presented thesis, author tries to explain peculiarities of a high-throughput analysis of the RNA-edited deep sequencing data. It consists of two parts introduced by maybe a bit too short general introduction. I'd definitely like to read more about the unique host-symbiont system and overall biology of *Perkinsela*. Author just refers to an attached manuscript, but the information there is pretty scarce too, restricted to a single paragraph. The first part is written by Vojta only, with a detailed description of the analysis algorithm and the workflow. The manuscript submitted to an unspecified scientific journal, describing RNA editing in the mitochondrion of basal kinetoplastid *Perkinsela*, makes up the second part. It is obviously more general, by far more readable and of better overall quality. But given the composition of co-author team, this is hardly surprising. It is apparent Vojta did a great job and tremendous amount of work, which easily justifies awarding him the title. I have no problem with the second part at all and have thoroughly enjoyed reading it. Unfortunately, this can not be said/written about the first part, and I would like to focus my review mainly there.

As already stated, the first part consists of the detailed description of algorithms and pitfalls of the correct assembly of the NGS data of the RNA-edited system. Now, I absolutely believe this is a rather advanced topic hard to explain. Actually, after reading the first part of the thesis carefully several times, I am sure about one thing: I will try to avoid working on the bioinformatics analysis of this kind of data (and will try to delegate it to Vojta). On the other hand, I am also positive it could have been written in a more comprehensible manner. And I have a good reason for it: the respective part of the attached manuscript. There, with the help of seasoned co-authors, Vojta managed to describe the procedure in simple-yet-effective way in the comparable (and definitely sufficient) level of detail.

The problem of the first part is both in the structure of the text and the language used. I miss at least one paragraph describing general outline and overview of the analysis. Vojta goes to the detail right from beginning of the respective parts and often leaves it on the reader to guess the reasons and consequences. The suggested introductory part should also contain the flowcharts of the analysis, actually quite nice and comprehensible, but somewhat buried at the end. Then there is a language part of the problem. The whole thesis is written in English, yet the first part seems to be written in some meta-English. The words are definitely right, but sometimes one has to work really hard to get the meaning. Looks like Vojta is deeply emerged in the topic and could not be bothered to explain it also to 'outsiders', or had a severe pressure (or maybe the combination of both?). The first part of thesis is concluded by Discussion, which really is not the discussion in the usual sense as it doesn't contain any reflection of the work in the context of existing literature (maybe because there is nothing relevant available? Just guessing). Instead, Vojta here suggests improved algorithm and the possible future direction, which is actually nice and somewhat improves his score. It definitely confirms Vojta's insight and expertise.

I have one general comment/question, which I'd like to hear the answer during the thesis defense: The experimental design was tested on two sets of real-life data: RNA-seq run of *Paraomeba – Perkinsela* system and iCLIP read sof *Trypanosoma brucei* . Do you see any benefits of testing of the performance of improved Bowtie2 and T-aligner code also on the simulated RNA-editing dataset (which preparation could not be that complicated)?

Finally, reviewers usually provide list of typos/minor errors to prove they did they job correctly. Here are my top picks:

Annotation: bachelor instead of master! Looks like an outcome of the reckless copy-paste event.

page 1 - additional bracket in reference

page 3 – 'Because' instead of 'Although' or use of 'even' after 'Although'

- Fig 2 legend 'amplified wit' instead of 'with'
- the reference to Fig 2 and 3 in a text is switched

page 4 - looks like there is a verb missing in the first sentence of paragraph

- last paragraph "If a seeding region is carefully chosen..." Does that mean the regions were targeted to some (never-edited) region of the sequence? Could this affect the outcome also in negative way?

page 5 - last paragraph suggests T-aligner has been used as an alternative sw for alignment of the seeded reads. However, in the previous chapter about seeding, author describes setting the seeding parameters also for T-aligner. Please explain if possible.

I hope that from my review is apparent I find the thesis interesting and scientifically relevant and valuable (especially the second part). In spite of my objections to the first part I have no doubts it meets all the requirements for successful defense, and I wish the author luck in his future scientific career.

Aleš Horák
Biology Centre ASCR, Institute of Parasitology
České Budějovice