List of changes

Son Hoang Nguyen – s4399652

The University of Queensland

I. Responding to Examiner A

First, throughout the thesis, various datasets are used to evaluate the algorithms that have been developed. Along the path of development, structural decisions get made and default parameter values get chosen. It is not uncommon for these decisions to be made in a way that fits them to the testing data that is in use during development. The result, can, in some circumstances, lead to unintentional overfitting of the resulting algorithm to the data that was available during development. Ideally, to get a candid and accurate estimate of how a new method performs, the performance as compared to other methods would be evaluated on a totally new data that was not available at the time of development. While I realise that it is probably not practical to generate and analyse new data as part of the revisions process, I do think it is important to add text that states clearly whether or not the testing data was held aside during the development process.

Response: Regarding the concern of overfitting due to testing data, in fact, only part of the data reported in this thesis were used for development. In more details, E.coli K12 and K.pneumoniae genomes were used for npScarf testing, synthetic data were used for npGraph and , the only 7-concatemer read in chapter 5. Applications of tools to other data sets (e.g. 4 XDR strains in chapter 3, real data in chapter 4, all other concatemerics reads in chapter 5) were unsupervised and reported accordingly as-it-is.
Of course, there was an inevitable limitations from the early stage of the development, however, as stated in the Contributions section (page 27), the tools are continuosly improved by processing feedback and reported bugs from another sources:
“The source code of the whole project is made available and applied in numerous use cases internally as well as from community. Feedback from users and external data sources are treated with care to further improve the performance and functionality of the developing software.”

Second, there are a large number of typographical and grammatical issues throughout the document. While I was able to understand the meaning of the text in spite of these, a reader who is less familiar with the topic or perhaps less familiar with English than myself may struggle. I don't think it's essential to correct these as they are mostly cosmetic but I do want to highlight the fact that the readability could be improved.
Response: The thesis has been subjected to another round of proof-reading and editing for better readability.

Figure 1.5: I don't think "Amplification" is a universal step in sequencing. Perhaps "Library prep" should be used instead.
Response: change “Amplification” to “Library prep” as suggested by examiner.

page 14: "DNA sample is normally sheared by restriction enzymes" some people do this but isn't "normal". There are many ways to shear DNA. Sonication, Tagmentation, and more.

Response: “normally” is for the fact of shearing DNA into fragments before sequencing, not for the method of shearing using restriction enzyme. The sentence is modified to “...is normally sheared, e.g. by restriction enzymes, into short fragments...” to clarify the meaning better.

page 14: "The rationale is to reduce the chance of missing pieces and also to cover the inevitable errors" this doesn't make sense to me. If I take it literally it sounds incorrect but maybe it's just not written in a way that I understand.
Response: change to “The rationale is to have enough coverage to compensate the stochastic errors which are inevitable during library preparation and sequencing process.”

page 20: I think "ABruijn" has been renamed as Flye?
Response: change “Abruijin” to “Flye”, adding citation “Kolmogorov et al 2019”

page 23: "each molecule progressing through the pore at 450Kbp per second" I think you mean 450bp / s, not kbp?
Response: change “450Kbp per second” to “450bp per second”

page 32 bottom "Miniasm [42]" this is not a hybrid assembler. miniasm has no consensus module nor error correction.
Response: change to “miniasm/racon pipeline [48]”

page 35: "mellalo-beta-lactamase" -> metallo-beta-lactamase
Response: change to “metallo-beta-lactamase”

page 42: what happened to the Canu+Pilon and Miniasm+Pilon assemblies for E. coli. With 67x coverage there should be enough to get asssemblies with these methods.
Response: No change. Canu and Miniasm did not generate decent skeleton assembly for E.coli as discussed in the last paragraph of section Results (page 46)

page 59: Fig 3.2g "K. aeruginosa" -> "P. aeruginosa"
Response: change to “P. aeruginosa” in Figure 3.2g caption

Chapter 3: the design is sensible, supporting streaming to other applications. The benchmarking could be much better though. The problem with only measuring accuracy on the Staph aureus genome is that some barcode sequences might be much easier to resolve from nanopore data than other barcode sequences. By only evaluating accuracy of one we get a very limited view of the accuracy. I appreciate that the reason for doing this was that the other genomes are fairly similar to each other and so it is more difficult to establish ground truth assignments for the reads. However, due to the nature of bacterial pangenomes, each of these genomes is likely to have a large number of genes that are specific and unique to that genome. Reads containing these pangenome genes should be unambiguously assignable to that genome. While there may not be many strain-specific genes among the four K. pneumo strains, the other genomes surely have hundreds if not a thousand unique genes. This would allow at least some reads to be assigned to these other genomes and accuracy could be computed on this subset of assignable reads.
Response: An exhaustive benchmarking for each strain is depicted in Appendix Figure B.2 as mentioned in page 58.

page 83: KL on Poisson distributions is a nice idea for use with clustering. any idea how well this works?
Response: KL distance (not KL divergence) had been used as one in many options for a distance function. It provides an estimation of error if two distributions are merged and this approach has been used in several clustering applications for mixture models fitting.

page 85: "Multiplicity estimation" some of this sounds similar to the methods implemented in Unicycler. Please cite if appropriate.
Response: the terms had been used earlier in npScarf (page 49).

Section 4.3: how are tandem repeats handled? Djikstra's shortest path algorithm is going to give an acyclic path, but the true genome could have a repeat contig occur multiple times in between two unique contigs. It's fine if this isn't handled, but I do think the text should be explicit about whether or not it's handled.
Response: Djikstra's shortest path finding algorithm had just used in npGraph to reduce to search space for candidate paths. Tandem repeats, if present, would have longer distance than the lower bound and are surely included in the candidates list. The best path can be detected by highest likelihood in terms of alignments, distance, coverage among the list of candidates.

p104: "plan infectious family" -> "plant infectious family"
Response: change to "plant infectious family"

p112: algorithm 3, line 13, is this supposed to be SS(S,M)?
Response: change line 13 in Algorithm 3 to "SS(S,M)"

II. Responding to Examiner B

1. The major scope of the thesis is about genome assembly. I would suggest a flowchart to describe the whole process of de novo assembly including data input, contig construction, genome scaffolding and base polishing.
Response: flowcharts for de novo assembly are already presented separately for short read and real-time long read data in Figure 1.5 and Figure 2.1. Another assembly flowchart for summary is a good idea but mostly will overlap the available content.

2. Contents in Chapter 1 on ONT sequencing could be improved by providing more information about devices, procedures of data processing and future developments.
Response: adding more information for Oxford Nanopore sequencing section (pages 6-9) about devices development and applications.
Procedures of data processing are described in section 1.2.2

3. Table 4.2 The top solid line is missing.
Response: adding top line to Table 4.2

4. Make sure all the tables the sane layout, but Table 5.1 is different.
Response: change Table 5.1 layout.

5. Equation index: make sure every equation has an index
$Y = X + Z \quad (2.1)$
Some equations in Chapter 2 have, but some in Chapter 4 don't have.
Response: numbering all Equations in Chapter 4 (page 83,84)