Thesis amendment

Son Hoang Nguyen – s4399652

The University of Queensland

I. Responding to Examiner A

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.”

The thesis has been subjected to another round of proof-reading and editing for better readability.

Figure 1.5: change “Amplification” to “Library prep” as suggested by examiner.

Page 14: “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: 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: change “Abruijin” to “Flye”, adding citation “Kolmogorov et al 2019”

Page 23: change “450Kbp per second” to “450bp per second”

Page 32: change to “miniasm/racon pipeline [48]”

Page 35: change to “metallo-beta-lactamase”

Page 42: 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: change to “P. aeruginosa” in Fig 3.2g caption

Chapter 3: An exhaustive benchmarking for each strain is depicted in Appendix Figure B.2 as mentioned in page 58.

page 83: 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: the terms had been used earlier in npScarf (page 49).

Section 4.3: 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.

Page 104: change to "plant infectious family"

Page 112: change line 13 in Algorithm 3 to "SS(S,M)"

II. Responding to Examiner B

1. 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 figures.

2. 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. Adding top line to Table 4.2

4. Change Table 5.1 layout.

5. Numbering all Equations in Chapter 4 (page 83,84)