**PhD Thesis Examiner's Report**

**Examiner:** **Zemin Ning**

**The Wellcome Sanger Institute, UK**

**Candidate:** **Son Hoang Nguyen**

**The University of Queensland, Australia**

**Thesis Title:** **Real-time analysis for Nanopore sequencing data**

========

**Summary:**

========

By directly targeting single DNA molecules, the DNA sequencers like MinION developed by Oxford Nanopore Technologies (ONT) offer real-time sequencing, where reads are available for analysis as soon as they have passed through the sequencer. To cope with the time-based data production, algorithms and software tools are needed for various applications when the platform is under operation. Followed by a detailed review on sequencing technologies and genome assembly, this thesis presents a comprehensive study on ONT analysis including data processing, scaffolding assemblies from other sequencing platforms, barcode sequencing, assembly graph and viral genome analysis. During the PhD period of time, the genome scaffolding tool npScaff has been published in Bioinformatics (Nguyen et al., 2017) and application on scaffolding and data analysis have been presented in a paper in Nature Communications (Cao, M. D.,Nguyen, S. H. et al., 2017). It is likely that more paper will come out in the near future as a result for this fruitful PhD thesis.

==========================

**Conclusion and Recommendation**

==========================

The merits of this PhD thesis are the algorithms designed and software tools developed for data processing and genome assembly based on ONT sequencing data. The research work conducted by the candidate is significant and original. As a result, two papers have already been published in good journals, Nature Communications and Bioinformatics. Now the third generation long read sequencing is playing a dominant role in de novo assemblies and genomics community will benefit from these useful pipelines. Data analyses carried out are comprehensive and in most cases there are comparisons against alternative methods or tools. The format and literary presentation of the thesis are satisfactory. Writing of the document is of professional standard and I enjoyed reading the thesis.

Based on these observations, I recommend that the candidate be awarded the degree of Doctor of Philosophy after minor corrections.

=================

**Review of each chapter**

=================

Chapter 1 presents a concise literature review including sequencing technologies, genome assembly and prospects of data analysis in real-time. With an introduction to genome sequencing, a focus has been put on Oxford Nanopore sequencing MinION. Coping with high levels of sequencing errors, particularly homopolymer errors is the biggest challenge in the community. In the section of genome assembly, general methods are introduced in long and noisy reads and this paves a way for future use in the next chapters. Nanopore sequencing has the feature of real time data production and this requires a prompt response from software to process all the related information.

Chapter 2 describes a genome scaffolding pipeline npScarf, a real-time system using MinION long reads. The methods implemented allow assembly completion directly using Nanopore sequencing data for bacterial genomes. Five different datasets were used to verify the pipeline and it demonstrates that a certain level of read coverage is needed in order to achieve a good level of scaffold continuity. For example, the K. pneumoniae ATCC 13883 could be improved to four scaffolds with 18-fold coverage of the MinION data. If read coverage is below 10X, the scaffolding continuity is very much limited. A comprehensive analysis has been conducted to compare other existing tools for genome scaffolding, including popular tools such as SPAdes and SSPACE. For the tested cases, npScarf performs well both in N50 and assembly accuracy.

Chapter 3 discusses real-time demultiplexing in Nanopore sequencing and the developed tool npBarcode. High throughout and multiple samples at the same time are the common features for recent third generation sequencing technologies. In the case study, the author showed the pipeline allowed simultaneously scaffolding for seven bacterial samples while the sequencing was still in progress. After 16 hours of sequencing and with about 80Mbp of Nanopore reads, one K. pneumoniae isolate was completed.

Chapter 4 presents practical solutions using assembly graph in real-time by long reads with npGraph. Assembly graph can be related to a number of applications, inclusing contig construction, repeat analysis, variant calling and haplotype determination. Methods as well as software tools have been developed in using assembly graphs various cases. The npGraph pipeline takes inputs from Illumina assembly graphs together with long read graphs of ONT or PacBio and paths from short or long reads are traversed to visualise genome features as well as assembly error detection. Hybrid assemblies have also been discussed.

In the last chapter, Chapter 5 introduces MinION sequencing applications for bacterial genomes, plasmids or viruses. In bioinformatics analysis, it described a reference-based method to detect concatemers.

================

**Suggested corrections**

=================

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.

2. Contents in Chapter 1 on ONT sequencing could be improved by providing more information about devices, procedures of data processing and future developments.

3. Table 4.2 The top solid line is missing.

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

5. Equation index: make sure every equation has an index

Y = X + Z (2.1)

Some equations in Chapter 2 have, but some in Chapter 4 don't have.