A Thesis for the Degree of Ph.D. in Science

Computational pipelines for assembly, analysis, and evaluation of genome sequences

January 2018

Graduate School of Science and Technology
Keio University

Vasanthan Jayakumar

別表6
（3）

Keio University

Thesis Abstract

No. _____

Registration Number	■ “KOU” □ “OTSU” No. *Office use only	Name	Jayakumar Vasanthan

Thesis Title

Computational pipelines for assembly, analysis, and evaluation of genome sequences

Thesis Summary

DNA sequencing has enabled the determination of genome sequences of a plenty of organisms. From Sanger of the first-generation sequencing, through second-generation sequencing, DNA sequencing has come a long way foraging more recently into third-generation, single molecule sequencing. Although, second-generation sequencing helped resolve the genomes of a numerous organisms, the assembled genomes were mostly fragmented with long unresolved bases termed as gaps. A major caveat of the second-generation sequencers is the short read length, which is incapable of resolving repetitive genomic sequences, leading to fragmented genome assemblies. In general, if a repeat sequence is longer than that of the sequenced read, it would be impossible to assemble the genomic region enclosing such repetitive regions. The biggest advantage offered by third-generation sequencing is the exponential increase in average read lengths, which are generally longer than most genomic repeats. With the third-generation sequencing on the rise, the development of a number of long-read based *de novo* assembly tools also came into the scene.

In this dissertation, several computational methods and pipelines for genome assembly of third-generation sequencing data were developed, along with the annotation of genomic features such as genes and repeats, and the evaluation of the quality of assembled genomes using third-generation sequencing. In the first study, the design and development of computational pipelines using third-generation sequencing data was applied to assemble the genome of *Ipomoea nil*, a popular flowering plant from Japan, producing various mutational patterns. Several parameters were needed to be adjusted to obtain the desirable assembly. Hence in the second study, the focus was on tuning parameters in the pipeline and applying the same to the genome assembly of a variety of organisms, and in the process evaluating all the available long-read assemblers. As no such evaluation study was attempted using third-generation sequencing reads, the study was designed in such a way to guide researchers on which assembler to choose for their respective assembly projects.

Chapter 1 introduces the concepts of sequencing and *de novo* assembly.

Chapter 2 discusses the pipeline constructed in this thesis for long-read assemblers.

Chapter 3 reports the assembly of a pseudo-chromosomal level draft genome of *I. nil*. The draft genome has a scaffold N50 of 2.8 Mb, and a contig N50 of 1.8 Mb, and hence the quality of the assembly is comparable to those achieved using Sanger sequencing reads. The draft genome has enabled the identification and the cataloguing of the *Tpn1* family transposons, known as the major mutagen of *I. nil*, and analysing the dwarf gene, *CONTRACTED*, located on the genetic map published in 1956. Comparative genomics analysis has suggested that a whole genome duplication in *I. nil*’s Convolvulaceae family, distinct from the most recent Solanaceae event, has occurred after the divergence of the two sister families.

Chapter 4 reports the evaluation of ten long-read assemblers using a variety of metrics on PacBio datasets from different taxonomic categories, with considerable differences in genome sizes. The evaluation also serves as a guide on efficiently tuning parameters for a genome assembly. The results helped narrow down the list to a few assemblers that can be effectively applied to eukaryotic assembly projects. Moreover, it is demonstrated how best to use limited genomic resources for evaluating the genome assemblies of non-model organisms.

Chapter 5 presents an overall summary of the thesis. Also, future research projects stemming from this research are discussed.