A comprehensive evaluation of assembly tools for viral genome assembly.

Shashidhar Ravishankar Eldin Talunzic Christian Olsen
Fredrik Vannberg

May 10, 2016

Abstract

Abstract will go here

Introduction

Genome assembly has been one of the most difficult problems in bioinformatics. With the advent of second and third generation sequencing methodlogies, sequencing new genomes has become ever more easy. Current development of accurate assembly strategies are of at most importance,in order to avoid misassemblies and gaps in the assembled genomes.

Previous studies such as the GAGE and Assemblathon2 projects have attempted to evaluate the efficiency and accuracy of current genome assembly tools. From the results of these studies it is very clear that, using just one assembly methodology, does not ensure the best assembly. Moreover, the use of a particular assembler must be motivated by the dataset being assembled. Though these studies give a clear picture of the current status of genome assembly, and suggest which assembly strategy would be the optimal strategy for a wide range of organisms. It is necessary to understand that a strategy that might work for bacterial genome, may not produce the best assembly for a viral genome. Thus it becomes important to evaluate the different assembly strategies and identify the best methodlogy based on the complexity of the dataset.

Plasmodium malariae is one of the five malaria parasites known to infect simians and humans. The genomes of closely related species such as *P.falciparum*, *P.knowlesi*, *P.vivax* have been sequenced completely. The AT rich genomes of the *plasmodium* species make the task of genome assembly very difficult. The *plasmodium* genome is known to have AT rich repeat sequences which have been previously hard to assemble over due to the limitation of assemblers in accurately assembling repeat regions longer than the read length. This is a known drawback of short read sequencing technologies. To overcome this issues, we obtained paired end Illumina MiSeq data 250*bp* as well as PacBio RSII long reads. By utilizing data generated from these two strategies we hoped to overcome the limitations of short read sequences to assemble over long repeats and the high error accompanied with long reads from PacBio SMRT sequencing.

Here we present the strategy we implemented in the assembly of the *P.malariae* genome. We will describe the various assembly strategies we tried, the pros and cons of each method, and describe a assembly pipeline we believe to be the best for the large viral genomes such as *P.malariae*. In order to ensure that the results described here are replicable, all the parameters for the assembly as well as the complete pipeline is made pubically avaiable.