HMP Hybrid Assembly
Human Genome Sequencing Center
Baylor College of Medicine

Author: Michael E. Holder
Version: 1.0c
Effective Date: 04/04/2011

1 Abstract

2 Introduction

This SOP describes the procedure for hybrid metagenomic assembly using Illumina fragment paired end and 454 fragment sequence data derived from samples collected by the Human Microbiome Project.

3 Requirements

3.1 Software requirements

The assembler used was the 454/Roche GS Mapper/Assembler, April 19, 2010 pre-release version.

4 Procedure

4.1 Process Illumina FastQ

- Screen for Human sequence
- Remove duplicates, trimmed for quality by the SRA and the HMP's data processing working group and converted to a format which can be used by the 454/Roche GS Mapper/Assembler). For a given sample, 3 files (read 1, read 2, and singletons) are supplied and subsequently merged into a pair of FASTA and Phred quality files. If both read 1 and read 2 exists for a given fragment they are considered to be a pair and the FASTA header of each is annotated such that read 1 is forward and read 2 is reverse.
- Also, each read's header is given tags to indicate the common library and common template. Reads that do not pair are considered singletons and no special action is taken other than to include them in the FASTA and quality files.

4.2 Trim data for quality

454 SFF data which has been screened for human sequence by SRA is trimmed for quality by using the assembler's inherent quality trimming mechanism. This process is accomplished by handling one SFF file at a time using the following steps:

a. Setup and start the SFF assembly using the -noa flag which will only read and analyze the input and produce a 454TrimStatus.txt file.
b. Parse the 454TrimStatus.txt file to gather the left and right trim points for each read plus the resulting trimmed length.
c. Using the caveat that reads be 60bp or greater in length, taken from the HMP's PGA assembly requirements, a list of valid read names and a list of valid read names their trim points are compiled

d. Using the SFF file command, trimmed SFF files are produced by importing the original SFF file and selecting the valid reads and their trim points, thus excluding unwanted reads.

4.3 Produce sample assembly

The sample assembly is produced using the outputs of items *4.1* and *4.2* above. In setting up the assembly, the 454 SFF data is input first and followed by the Illumina data which is designated as paired end. Since all data has been trimmed, and to suppress any additional trimming during assembly, the -notrim option is used.
Other options used for assembly are -cpu 32, -large, -nobig, and -rip.

5 Implementation

6 Discussion

7 Related Documents & References

This work was supported by a 2010 IBM Shared University Research (SUR) Award on IBM's Power7 high performance cluster (BlueBioU) to Rice University as part of IBM's Smarter Planet Initiatives in Life Science/Healthcare and in collaboration with the Texas Medical Center partners, with additional contributions from IBM, CISCO, Qlogic and Adaptive Computing.
<http://www.ibm.com/developerworks/university/sur/index.html>
<http://www.ibm.com/developerworks/university/sur/index.html>
[www.ibm.com]

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.00	Michael E. Holder	04/04/2011	Establish SOP
1.0c		09/20/2011	Converted to standard template