HMP WGS Read Processing

[1]Broad Institute of MIT and Harvard
[2]The Genome Institute, Washington University School of Medicine
Author: Sarah Young[1], John Martin[2], Karthik Kota[2], Makedonka Mitreva[2]
Version: 1.0c
Effective Date:

1 Abstract

2 Introduction

This SOP describes the procedure used to process HMP WGS reads.

3 Requirements

4 Procedure

To process WGS reads, we followed a series of steps to ensure quality and privacy of the datasets. The main steps in the process were: identify and mask human reads, remove duplicated reads, and trim low quality bases. Here we describe each of these steps in detail.

4.1 Identify and mask human reads

- Raw trace data were first submitted to NCBI's Sequence Read Archive by the sequencing centers.
- There, human bases in the reads were identified and masked using BMTagger (unpublished: Rotmistrovsky, K. and Agarwala, R.).
- The processing centers downloaded full SRA-formatted files from SRA ftps for each SRR id. To assure download integrity, md5 sums were also downloaded and confirmed upon download completion.
- Once downloaded, fastq-dump v1.2.0 (part of the SRA Toolkit) was used to extract fastq files using the following options: fastq-dump -E -A $srr_id -D $srr_id/ -DB '@$sn/$ri' -DQ '+$sn/$ri' -O $sample/ >& $sample/$srr_id.fastq-dump. The –E option assured that all bases were written to the fastq file, while the –DB and –DQ options assured proper read naming.

4.2 Remove duplicated reads

- Once fastq files were created for each run, they were aggregated into a single fastq by sample (SRS id). This aggregated fastq file was then converted into BAM format using FastqToSam (Li et al.)
- Duplicate reads were marked and removed using a modified version of EstimateLibraryComplexity, part of the Picard tool package for manipulating SAM and BAM formatted data (Fennel, T.). This tool employs a method for identifying duplicate

Author: Sarah Young[1], John Martin[2], Karthik Kota[2], Makedonka Mitreva[2]

reads which are artifacts of the sequencing process without requiring alignment to a reference.

4.3 Trim low quality bases

- Lastly, we trimmed low quality sequence using a modified version of trimBWAstyle.pl (Fass. J), that worked directly from BAM files.
- This script was used to trim bases off the ends of reads which had a quality value of 2 or lower. This threshold was chosen to remove all bases of uncertain quality as defined by Illumina's EAMMS (End Anchored Max Scoring Segments) filter, and which are marked with quality values of 2.
- Reads trimmed to less than 60bp were removed, and their partners, if longer than 60bp were placed in a separate singletons file. This script produces three fastq files, one for each read pair end, and a third for singletons left after trimming.

5 Implementation

These processed read files were then uploaded to the DACC, where they are currently available under the http://hmpdacc.org/HMIWGS section.

6 Discussion

7 Related Documents & References

- Li,H. et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078–2079.
- Fass, J. trimBWAstyle.pl. Unpublished. The Bioinformatics Core at UC Davis Genome Center.
- Fennel, T. et al (2009) Picard. http://picard.sourceforge.net/index.shtml.
- Rotmistrovsky, K. and Agarwala, R. BMTagger: Best Match Tagger for removing human reads from metagenomics datasets. Unpublished.

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.0	Sarah Young, John Martin, Karthik Kota,		Establish SOP

Author: Sarah Young[1], John Martin[2], Karthik Kota[2], Makedonka Mitreva[2]
Version: 1.0c
Effective Date:

	Makedonka Mitreva		
1.0c		09/20/2011	Converted to standard template