omics.data.edu.au

Genomics Requirements Summary

This document is a high level summary of the genomics area requirements and information captured during the discoverability phase.

Summary

Genomics is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes

At the Ramaciotti Centre for Genomics, next generation DNA sequencing data is collected from a Pacific Biosciences (PacBio) RS-II machine.

For DNA from bacterial samples collected at the Ramaciotti on the PacBio RS-II, staff use the Pacific Biosciences SMRT Portal (specifically the HGAP (Hierarchical Genome Assembly Process) assembly algorithm), to perform a *de novo* genome assembly for the client.

DNA sequencing data is also captured from an Illumina (MiSeq or NextSeq) machine. These shorter read Illumina data are used as an input into the Snippy tool (https://github.com/tseemann/snippy) in order to correct assembly errors that can arise from PacBio data assembled using SMRT HGAP.

For the BPA-funded SEPSIS project, these data will be sent to the BPA data portal (managed by the Centre for Comparative Genomics, Perth) by staff at the Ramaciotti using Datasend (where the recipient is sent an email and presented with a URL for data download).

It is intended that genome annotation (i.e. defining which genes may be present) will also be undertaken by the bioinformatics staff at the Ramaciotti using the command-line driven tool Prokka (https://github.com/tseemann/prokka).

Groups Consulted

Ramaciotti Centre for Genomics

- Marc WIlkins - Director
- Nandan Deshpande - Bioinformatician
- Tonia Russell - PacBio technician

Victorian Life Sciences Computation Initiative

- Torsten Seeman

Work Pattern

- Provider of a DNA sample registers with the Ramaciotti.
- Physical DNA sample is sent it to the Ramaciotti centre with a local ID on tube.
- Ramaciotti staff perform 3xQC tests on sample (nanodrop, Qubit and a pulsed-field gel). Results are emailed to DNA provider.
- PacBio DNA sequencing performed.
- Raw data is stored coming off machine on Ramaciotti server
- PacBio data is assembled using SMRT HGAP (software installed on Ramaciotti server)
- Raw and assembled genome files are sent to sample submitter through a local (Ramaciotti) deployment of Datasend (www.filesender.org).
- Note that for the SEPSIS project bioinformaticians at Ramaciotti will also undertake:
    - a correction of potential errors of the PacBio assembly by using Illumina data that is also being generated (this uses the Snippy tool), and
    - genome annotation will also be undertaken by the bioinformatics staff at the Ramaciotti using the Prokka tool.

Data Overview

PacBio 'raw' data

- Scale, volume: ~12-14 Gb per run.
- Types, formats:

    15 "analysis results" files come off a PacBio machine for each run, but Ramaciotti only only send a subset of these to customers
    i.e.: fasta, fastq, Bax.h5

PacBio assembled genome data

- Scale, volume: ~ 1Gb per run
- Types, formats:

    data-aligned_reads.bam
    data-aligned_reads.bam.bai
    data-corrected.fasta
    data-corrected.fastq
    data-polished assembly.fasta
    data-polished-assembly.fastq
    results-polished_coverage_vs_quality.csv

Snippy corrected genome data

- Scale, volume: modest
- Types, formats: When calling SNPs, Snippy produces 17 files all with a common prefix:

    .tab A simple tab-separated summary of all the variants
    .csv A comma-separated version of the .tab file

.html A HTML version of the .tab file
.vcf The final annotated variants in VCF format
.vcf.gz Compressed .vcf file via BGZIP
.vcf.gz.tbi Index for the .vcf.gz via TABIX
.bed The variants in BED format
.gff The variants in GFF3 format
.bam The alignments in BAM format. Note that multi-mapping and unmapped reads are not present.
.bam.bai Index for the .bam file
.raw.vcf The unfiltered variant calls from Freebayes
.filt.vcf The filtered variant calls from Freebayes
.log A log file with the commands run and their outputs
.consensus.fa A version of the reference genome with all variants instantiated
.aligned.fa A version of the reference but with - for unaligned and N for depth < --minfrac (does not have variants)
.depth.gz Output of samtools depth for the .bam file
.depth.gz.tbi Index for the .depth.gz (currently unused)

Prokka genome annotation data

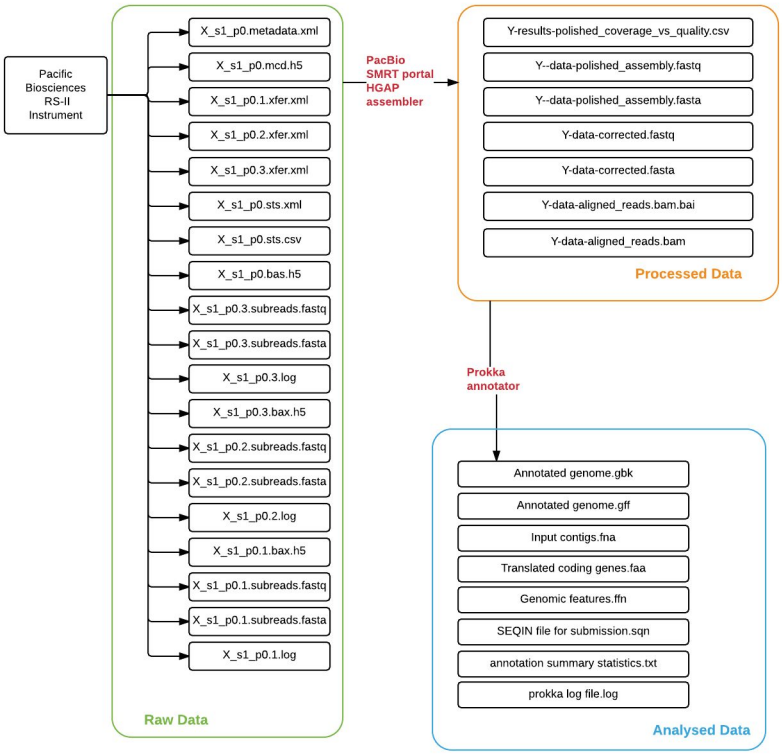
- Scale, volume: modest
- Types, formats: Prokka produces 10 files all with a common prefix:

.fna FASTA file of original input contigs (nucleotide)
.faa FASTA file of translated coding genes (protein)
.ffn FASTA file of all genomic features (nucleotide)
.fsa Contig sequences for submission (nucleotide)
.tbl Feature table for submission
.sqn Sequin editable file for submission
.gbk Genbank file containing sequences and annotations
.gff GFF v3 file containing sequences and annotations
.log Log file of Prokka processing output
.txt Annotation summary statistics

See also Figure below:

omics.data.edu.au

Ramaciotti PacBio Data/work flow



Software/Applications used

For *de novo* genome Assembly:

- SMRT portal HGAP (HGAP consists of pre-assembly, de novo assembly with Celera® Assembler, and assembly polishing with Quiver - see http://files.pacb.com/software/smrtanalysis/2.3.0/doc/smrtportal/help/Webhelp/SMRT_Portal.htm).

For genome assembly correction:

- Snippy (Finds SNPs between a haploid reference genome and NGS sequence reads. It assists correctling the assembly errors (aligns the reads backs to the contigs to check for discrepancies). - see https://github.com/tseemann/snippy#correcting-assembly-errors).

For genome annotation:

- Prokka (a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files. - see https://github.com/tseemann/prokka).

omics.data.edu.au

Deficiencies in processes and tools

None identified

Relationship with other Streams

Genome:Transcriptome:Proteome

Prof Marc Wilkins, Director of the NSW Systems Biology Initiative and the Ramaciotti Centre at UNSW is an expert in genomic, transcriptomic and proteomics science (and is credited with originally coining the term ‘proteomics’).

Whilst the existence of potential genes identified through annotation of *de novo* assembled genomes is usually validated using transcriptomic data, a novel complementary approach pioneered by Prof Wilkins is to validate the gene predictions in genome assemblies by comparison to proteomics data from the same species.

An open-source tool developed by Prof WIlkins for this purpose is the PG-Nexus: see https://projects.ands.org.au/id/AP11