OJGP
Ontario Joint Genomics Program

Data processing

This document covers the primary considerations surrounding data processing for the Ontario Joint Genomics Program (OJGP) member centres.

Please contact inquiries@ojgp.ca if you need more information.

Data storage and retention

Sequencing on the Illumina platform generates a large amount of data with each run. The NovaSeq 6000 system can generate as much as three terabases per flowcell while the NovaSeqX Plus system can generate as much as eight terabases on a 25B flowcell. Once converted to FASTQ format, depending on the level of compression, a terabase of sequence occupies, on average, 450 GiB of disk space. Another way to think of this is that a 30X whole genome requires about 60GiB of disk space to store the raw sequence in FASTQ format.

Analysis of this data generally involves alignment to a reference sequence and then assessing variation on the aligned sequence, which may increase the data footprint by 2-3 times. The first consideration is having sufficient disk space to store this data and for how long that data will need to be retained in that space. Developing a process to efficiently process sequences and quickly remove data following the release of any deliverables will reduce storage costs. Options for storage include robust local disk (redundant, fault-tolerant, high capacity) or cloud storage solutions (Google, Amazon, Illumina BaseSpace). Archiving solutions for data that needs to be maintained but does not need to be readily available include tape-storage solutions (e.g., Iron Mountain) and lower cost cloud archiving (e.g., Amazon Glacier). For OJGP assays with clinical reports, there is an accreditation requirement to maintain the raw sequence data and any files that are used as direct input to the clinical report (FASTQ for two years, report for 10 years). For RUO projects, we recommend a best practice of retaining FASTQ files for 60 days (or less) before moving to a long-term storage solution for two years using one of the archiving solutions suggested above in case data needs to be re-examined. Note that long term storage solutions are relatively cheap for data that likely will never need to be retrieved.

Analysis

There are a variety of approaches to managing analysis of the sequencing data. Analysis generally includes running workflows for generation of quality control metrics and processing through analysis pipelines to generate output appropriate for the library types and experimental design. Analysis can be run on local systems, and in general is done on a high-performance cluster with computational nodes having sufficient power and memory to process data efficiently. Analysis can also be run in a variety of cloud environments. Analysis solutions may use open-source software, which is generally free and supported by the research community, or proprietary solutions with cost for both use and support. Automation of analysis pipelines will reduce hands-on time and reduce the overall turnaround-time to completion. A number of workflow management systems exist including Nextflow, Cromwell or Snakemake. Depending on the capabilities and size of a centre, it can consider running analysis using the on-board capabilities of the NovaSeq X (Dragon processor) for your sequence capture and processing needs. There are a variety of analysis workflows and pipelines available through different groups including the OICR Genome Sequence Informatics (GSI) team, which can be incorporated into your own analysis pipelines. GSI also has a bring-your-own-data service to process data using its pipelines on OICR servers (or on the cloud) on a cost recovery basis.

Data transmission

Given the large size of the data being generated and needing to be moved around, access to high-speed networks with a large bandwidth for transmission is required.

Release of data

As a sequencing core, data must be available to clients or customers who have provided the samples and who own and are ultimately responsible for maintaining the data being generated. This includes the raw sequence data, generally in compressed FASTQ format, and any analysis deliverables. Data is generally distributed through secure FTP sites, but can also be made available through cloud services, or through deposition to various sequence archives (e.g., EGA, dbGap, GEO). Efficient methods for data distribution will reduce the need to retain the data at the sequencing core.

Glossary

Terabases – Equivalent to $1x10^{12}$ bases

FASTQ - a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores

OJGP – The Ontario Joint Genomics Program

OICR – Ontario Institute for Cancer Research

RUO – Research Use Only

GSI – Genome Sequencing Informatics

FTP – File Transfer Protocol