KINGHORN CENTRE FOR CLINICAL GENOMICS | GARVAN INSTITUTE

Garvan Institute of Medical Research
384 Victoria Street
Darlinghurst, NSW 2010
Australia
T: + 61 (0)2 9355 5846
E: kccgseq@garvan.org.au

Overview

This document details the general file structure of the deliverables for your Illumina HiSeq X sequencing data.

Included in the hard drive you will find FastQ files, MD5 checksums, FastQC Reports and a Project Summary Report based on Picard WgsMetrics (whole genome) tools. Each of these files will be further discussed in this document.

Project Deliverables

The sequencing data is found in a project folder which holds the same name as the project manifest.

```
* Project Folder/
    + Run Folder(s)/ - a project may contain one or more run folders.
    + SequencingDeliverablesReport.pdf - This document.
    + SequencingProjectReport.pdf - Report on the sequencing data.
```

The sequencing data is separated based on run folders.

Run folders are folders grouping the sequenced lanes in a single flow cell.

```
* Run Folder/
    + inputFastq/  - this folder contains FastQ files with their MD5 checksums for the samples in this run folder.
    + inputFastq.rawFastQC - / This folder contains the FastQC reports for the samples in this run folder.
```

The run folders are named following this pattern:

RunDate_InstrumentID_RunNumber_InstrumentSideAndFlowCellID

Item	Description
RunDate	The date the run began.
InstrumentID	The barcode of the sequencer machine.
RunNumber	The run number of the sequencer machine.
InstrumentSideAndFlowCellID	Instrument side refers to the side the flowcell was loaded on the instrument (A or B). FlowCell ID refers to the barcode on the flowcells.

For example, if we had a project R_MyProject with two separate run folders, the folder manifest name would be R_MyProject_M001 and the structure could look like this:

```
* R_MyProject_M001/
    + 141107_ST-E00199_0018_AH0ANUALXX/
    + 141107_ST-E00106_0121_BH078KALXX/
    + SequencingDeliverablesReport.pdf
    + SequencingProjectReport.pdf
```

The two run folders:

```
* 141107_ST-E00199_0018_AH0ANUALXX/
    + This run started on the 7th of November in 2014. This run was done on the A side of the sequencer with barcode ST-E0(
* 141114_ST-E00106_0121_BH078KALXX/
    + This run started on the 14th of November in 2014. This run took place on the B side of the sequencer with barcode ST-
```

FastQ

The FastQ files contain the sequencing data as a string along with its corresponding quality scores. The quality scores use the Sanger encoding format.

The FastQ files are generated using Illumina's Bcl2fastq 2.15.0.4.

The FastQ files come in pairs of a Read1 and Read2 which is signified by the R1 and R2 in the file name. Each paired end read of FastQ files is generated from a single lane in the flow cell. However, a single sample may have multiple lanes, resulting in having more than one pair of FastQ files per sample. The sample is run on multiple lanes if the coverage requested is larger than 30X or if the single lane did not result in the 30x coverage.

For further information about FastQ format please read this article and visit this website.

Please note that the FastQ files are raw sequence files and will contain the Standard Illumina Adapter AGATCGGAAGAGC*. For Runs completed on October 2014 onwards(e.g. 141014_ST-E00106_0121_BH078KALXX), the adapters have been masked for Read1 and Read2 using the adapter sequence. i.e. the MaskAdapter and MaskAdapterRead2 is set to the Standard Illumina Adapter above during Illumina's Bcl2fastq2 Conversion software.

A spike of PhiX control at 1% is added in each lane. The library is made up of small well-characterised PhiX genome which offers sequencing and alignment benefits. For further details on the product used, please visit the Illumina product website

Each FastQ file is named using the following pattern:

FlowCellID_Lane_SampleID_Species_Index_Manifest_Read.fastq.gz

Item	Description
FlowCellID	The ID of the flow cell used for sequencing
Lane	The lane on the flow cell used for the sample
SampleID	The sample ID
Species	The Species the sample was taken from (default is human)
Index	Index sequence for multiplexing (optional)
Manifest	Manifest consisting of the project name followed by the manifest id. E.g R_MyProject_M001
Read	The Read number (R1 or R2)

For example, if we had a FastQ file named H0ANUALXX*5_4736337_Human*_R_MyProject_M001_R1.fastq.gz where the

> H0ANUALXX is the flow cell ID
> 5 is the lane number used on the flow cell.
> 4736337 is the sample ID.
> Human is the sample reference.
> R_MyProject_M001 is the manifest name (Project name is R_MyProject).
> R1 indicates the read number.
> fastq.gz is the file extension.

MD5 Checksums

The transfer of large files, such as the FastQ files, is prone to errors resulting in incomplete or corrupt files. To ensure that the files have transferred over safely onto this hard drive we have generated a MD5 checksum file for each FastQ file.

MD5 checksum is created by an algorithm that examines the original file and creates a checksum data string out of the original file's characteristics.

After transfer of the FastQ file and the MD5 file is complete, we run a checksum on the transferred FastQ file to check that it is identical to the original file.

The same MD5 files can be used to ensure that the files are transferred onto your machine without error. For a guide on how to do this please visit this website:

MD5 Checking Guide using GNU Parallel

FastQC

We have run FastQC across these FastQ files and provided the summary reports. FastQC is a quality control tool for high-throughput sequence data.

Sequencing Project Report

The Sequencing Project Report outlines the project samples and the quality of their sequencing.

To ensure that for each sample we have generated the requested coverage, we align the sequence generated in each run using Illumina's Isaac aligner (Raczy et al, 2013; link to whitepaper), then to determine the coverage, we calculate the mean coverage by using the following equation as described by the manufacturer:

((MEAN_COVERAGE) x (1 - PCT_EXC_DUPE - PCT_EXC_OVERLAP))/(1 - PCT_EXC_TOTAL)

If a sample does not meet the requested coverage, the sample is queued again for additional sequencing.

The total of the mean coverage from different runs is added up to determine if the sequence meets the required coverage.

The report name is SequencingProjectReport.pdf

Software Tools

Software tools used and their versions

```
* Illumina  bcl2fastq 2.15.0.4
* FastQC 0.10.1
* Illumina Isaac aligner 01.14.07.17
* Picard CollectWgsMetrics v1.119
```

Further Information

For more information please visit the KCCG Public Wiki.

Garvan Institute of Medical Research
384 Victoria Street, Darlinghurst, NSW 2010
T: + 61 (0)2 9355 5846 | E: kccgseq@garvan.org.au
http://www.garvan.org.au/research/clinical-genomics
https://ccg.garvan.org.au/confluence/display/KP/KCCG+Public+Wiki