HiSeq Services
Data Report

JVF-13-24

Customer		Fasteris	
Institution	"Dpt. of Biology,	Project	JVF-13-24
Contact	Tomaz Skrbinsek	Date / ref.	DR_2017-07-24_JFV-13-24
e-mail	tomaz.skrbinsek@gmail.com	Author	Manuele Castelnuovo

Summary

Sequenced libraries	**JFV-13-24**
Number of libraries	**12**
Sample type	**Amplicon**
Run ID	**170721_SND405_A**
Number of cycle	**2x150+7**
Instrument	**HiSeq**
Yield(GB)	**66.4**
Quality control	**Within specifications**

Author's comments:

1 Introduction

Next-Generation DNA sequencing (NGS) using Illumina technology:

DNA or RNA samples are processed in vitro to generate a library of short inserts.

Having passed quality control, the library is sequenced on Illumina HiSeq or MiSeq instruments. HiSeq can run in High-Output V4 mode (1 or 2 flow-cells, each with 8 independent lanes of 220-300 million pass filter clusters per lane) or Rapid-Run (1 or 2 flow-cells, 150-300 million pass filter clusters per flow-cell). MiSeq produces 15-25 million pass filter clusters per run (per flow-cell).

All reads have the same length. When using a forward sequencing primer, the sequences are called single-reads. The inserts can also be sequenced from both ends using a forward and a reverse sequencing primer, generating paired-reads.

In 2014, the typical yields are 28-38 Gb for a 1x125 bp lane on HiSeq High-Output; 45-90 Gb for a 2x150 bp flow-cell on HiSeq Rapid-Run; and 9-15 Gb for a 2x300 bp flow-cell on MiSeq.

Upon request, our bioinformatics team is happy to help you to extract the biologically useful information from your data.

NB: Although the illumina sequencing instruments deliver high-quality data, it is still a young technology and you must obtain independent validation of the data.

illumina CSPro CERTIFIED SERVICE PROVIDER

Fasteris is the first facility certified by illumina for sequencing applications (2008)

2 Samples

Sample name	Sample type	Library ID	Library type	Index	Insert min	Insert max
DAB13	Amplicon	JFV-13	MetaFast	ATCACG (Illumina)	50	150
DAB14	Amplicon	JFV-14	MetaFast	CGATGT (Illumina)	50	150
DAB15	Amplicon	JFV-15	MetaFast	TTAGGC (Illumina)	50	150
DAB16	Amplicon	JFV-16	MetaFast	TGACCA (Illumina)	50	150
DAB17	Amplicon	JFV-17	MetaFast	ACAGTG (Illumina)	50	150
DAB18	Amplicon	JFV-18	MetaFast	GCCAAT (Illumina)	50	150
DAB19	Amplicon	JFV-19	MetaFast	CAGATC (Illumina)	50	150
DAB20	Amplicon	JFV-20	MetaFast	ACTTGA (Illumina)	50	150
DAB21	Amplicon	JFV-21	MetaFast	GATCAG (Illumina)	50	150
DAB22	Amplicon	JFV-22	MetaFast	TAGCTT (Illumina)	50	150
DAB23	Amplicon	JFV-23	MetaFast	GGCTAC (Illumina)	50	150
DAB24	Amplicon	JFV-24	MetaFast	CTTGTA (Illumina)	50	150

Table 1: Description of the samples

3 Sequencing specifications

3.1 Instrument

Serial number	**D00405**
Manufacturer	Illumina
Version	HiSeq 2500
Slot used	A
Basecalling pipeline	- HiSeq Control Software 2.2.58 - RTA 1.18.64.0 - CASAVA-1.8.2

3.2 Run

Run ID	**170721_D00405_A**
Mode	HiSeq Rapid Run (RR)
Number of cycles	2x150+7
Number of lanes	2
Flow cell ID	HTKK7BCXY
Flow cell version	HiSeq Rapid Flow Cell v2
Kit version	HiSeq Rapid SBS Kit v2

3.3 Fasteris specifications[1]

Service	**Full flow-cell**
Yield (minimal expected, in Gb)	60
Q30	75 %

[1] Fasteris specifications for the run described in points above

4 Basecalling summary

4.1 Parameters

Lane	Expected project yield (Mb)	Number of mismatch in the index selection[1]	Phix spiked[2]	Other multiplexed libraries[3]
1	60'000	1	Yes	No
2	60'000	1	Yes	No

Table 2: Basecalling parameters
[1] According to Fasteris specifications, 1 mismatch is only authorized when all the indexes differ by at least 3 bases.
[2] Indicates if a PhiX reference is spiked (i.e. a low concentration of a PhiX library is added before sequencing and selectively retrieved through its related index) in your lane to estimate the error rate for your sequences.
[3] Indicates whether libraries from other projects are present in the lane

The base calling pipeline proceeds to the demultiplexing prior to the generation of fast-q sequence files, *i.e.* by separating the libraries according to their indexes. When different libraries are multiplexed in the same lane, a very low proportion of cross-talk may happen (reads sorted to wrong index)

4.2 Results

The sequences are sorted according to their index code (6 bases).

Lane	Expected yield (Mb)	Library ID	Yield (Mb)	%PF[1]	Cluster (PF)	Q30[2]	Mean qual. (PF)
1	5'000	JFV-13	3'082	74.19	10'274'076	85.36	34.84
2	5'000	JFV-13	3'090	73.49	10'301'539	85.51	34.87

Lane	Expected yield (Mb)	Library ID	Yield (Mb)	%PF[1]	Cluster (PF)	Q30[2]	Mean qual. (PF)
1	5'000	JFV-14	2'263	71.17	7'541'749	84.04	34.47
2	5'000	JFV-14	2'266	70.52	7'552'459	84.55	34.60
1	5'000	JFV-15	2'724	69.67	9'081'213	83.19	34.14
2	5'000	JFV-15	2'710	69.17	9'033'344	83.69	34.28
1	5'000	JFV-16	3'120	72.44	10'399'727	85.76	35.03
2	5'000	JFV-16	3'142	71.82	10'472'855	86.13	35.12
1	5'000	JFV-17	3'271	71.60	10'903'329	85.07	34.79
2	5'000	JFV-17	3'285	70.92	10'950'830	85.51	34.91
1	5'000	JFV-18	2'749	65.93	9'164'964	80.40	33.26
2	5'000	JFV-18	2'752	65.16	9'172'906	81.19	33.49
1	5'000	JFV-19	2'681	61.64	8'935'297	75.98	31.79
2	5'000	JFV-19	2'668	60.73	8'894'297	76.97	32.08
1	5'000	JFV-20	3'229	73.48	10'761'914	87.01	35.45
2	5'000	JFV-20	3'252	72.93	10'840'739	87.30	35.52
1	5'000	JFV-21	2'757	68.48	9'189'753	80.30	33.17
2	5'000	JFV-21	2'756	67.76	9'186'106	80.93	33.35
1	5'000	JFV-22	2'794	70.59	9'313'094	84.36	34.57
2	5'000	JFV-22	2'809	69.88	9'362'090	84.83	34.69
1	5'000	JFV-23	1'670	68.93	5'565'031	82.28	33.88
2	5'000	JFV-23	1'655	68.20	5'516'124	82.78	34.03
1	5'000	JFV-24	2'848	67.03	9'492'165	81.09	33.51
2	5'000	JFV-24	2'853	66.28	9'510'878	81.78	33.71

[1] PF stands for 'passed filter' i.e. clusters that fulfill the default Illumina quality criteria
[2] % of bases (PF) with a quality score greater or equal to 30

The yield of some of the libraries is below the expected yield.

4.3 Quality control

Spiked-PhiX:
Fasteris developed an “in-lane” control spike in each lane of the flow-cell. These spiked control reads are mapped on the PhiX reference genome. (Details can be found in the document “HiSeq_MiSeq_QCandSpecifications.pdf”).

Error rate (only for information):
The error rate is then measured by the number of mismatches over the total number of bases of mapped PhiX reads. (c.f. : “HiSeq_MiSeq_QCandSpecifications.pdf”).

Q30:
The illumina pipelines estimates the reads quality according to the percentage of bases having a base quality value greater or equal to 30 (Q30), ie less than 1 error in 1000 bases. (c.f. : “HiSeq_MiSeq_QCandSpecifications.pdf”).

Lane	Error rate Read1 %	Q30 Read1 %	Error rate Read2 %	Q30 Read2 %	Error rate Average %	Q30 Average %	Within specifications
1	0.57	93.42	0.88	88.15	0.725	90.785	Yes
2	0.57	93.18	0.73	88.58	0.650	90.880	Yes

N.B.: We do not perform quality filters on sequence files that are within our quality specifications, besides the default 'failed chastity' filter done by the pipeline itself. We have observed that additional filters can alter the representation/variability of the sequences (e.g. specific removal of sequences with secondary structures).

The base-calling pipeline can sometimes call bases as blanks ('N'). Blanks correspond to unattributed bases. If the blank rate is lower than 0.05%, no further information is provided.

5 Nomenclature

Library:	The DNA or RNA samples are processed into short fragments of 20-1000 bp, depending on the protocols (e.g. genomic shotgun, transcriptome, ChIP-SEQ, small RNA, etc..) and cloned *in vitro* between the 3' and 5' adapters. Most libraries are amplified by PCR to generate enough DNA for precise measurement of its concentration, a key factor for maximizing the yield (too low concentration will mean not enough DNA clusters and too high will results in too many overlapping DNA clusters that will be eliminated at quality filter step during base-calling)
Insert:	Sample fragment that has been incorporated between two adapters during library preparation after fragmentation and size selection.
Adapter:	3' and 5' sequences added during library preparation (used for PCR amplification, DNA cluster generation on the flow cell and sequencing)
Index:	A 6-to-8 bases DNA sequence tag found in adapter sequence to uniquely identify each library. The index is read separately from the inserts and its sequence is used at demultiplexing step.
DNA Cluster:	A DNA Colony generated on the flow-cell from a single DNA molecule of the library. It is perceived as a single sequencing unit by the base-calling pipeline, even if 2 DNA clusters are overlapping. In such case, the double sequence produced is eliminated at filtering step.
Read:	A sequence obtained after base calling. Its length is determined by the number of sequencing cycles. All the reads have the same length. Runs done using only forward sequencing primer, will generate one **single-read** per DNA cluster. Runs done using the forward and the reverse sequencing primers will produce two **paired-reads** (one pair) per DNA cluster.
PF Clusters:	Pass filter clusters. The illumina pipeline uses the chastity filter *(c.f. : "HiSeq_MiSeq_QCandSpecifications.pdf")* to remove sequences produced from clusters with low signal to noise ratio (e.g. overlapping DNA clusters).
PhiX Spike:	A Fasteris-developped quality control to measure the real error rate in the lane. About 0.5% of a PhiX library is added in each lane.
Demultiplexing:	Sorting the reads according to the indexes of each library

6 File format

Index selected sequences e.g. 170721_SND405_A_L001_JFV-13_R1.fastq.gz 170721_SND405_A_L001_JFV-13_R2.fastq.gz
File extension: fastq.gz
Format: Illumina fast-q format. A text file listing the sequences and their quality. Each group of 4 lines describe one sequence: *Line 1: is the sequence name (a unique identifier of the sequence)* *Line 2: the base sequence itself* *Line 3- orientation of the sequence (always + to indicate forward strand)* *Line 4- quality value for each base, encoded as a Phred score*

The sequence files are available to download on our secured server.

N.B. The sequence files are compressed as .gz archives. The archives can be uncompressed on linux OS using a gzip -d command. We cannot guarantee, due to their large size, that they can be uncompressed on Windows or MacOS systems.