Data Report

GLW-16

HiSeq run:

Instrument name	*SN365*
Slot	*B*
Instrument version	*“Hi-Seq 2000”*
Number of sequencing cycles	*2 x 50 + 7 (index)*
Sequencing kit	*TruSeq(TM) SBS v5*
Tiles	*32*
Data Analysis Pipeline	*HiSeq Control Soft. v. 1.3.8* *RTA 1.10.36* *CASAVA 1.7.0* *OLB-1.9.0*
Run code	*110415_SN365_B*

Libraries:

Sample preparation protocol:	*Genomic - mate-pair* *Paired-end* *Indexed*

Channel	Content
8	GLW-16 (+PhiX)

Data:

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.

Furthermore, a PhiX reference is spiked (small fraction added before sequencing and removed through its related index) in your channel to estimate the error rate for your sequences.

Index selection:

The sequences are attributed according to their index code (6 bases). Sequences are attributed with up to 1 mismatch allowed.

Therefore, the data set provided contains only sequences attributed without ambiguity to only one library.

Sample Name	Library Code	Index	File name	Reads	Bases
1	GLW-16	TGCGAC	110415_SN365_B_s_8_1_seq_GLW-16.txt	88'576'708	4'428'835'400
			110415_SN365_B_s_8_2_seq_GLW-16.txt	88'576'708	4'428'835'400
			Total:	177'153'416	8'857'670'800

Sequencing Quality control:

98.6% of the reads have been attributed to your library.

In addition to the error of the spiked PhiX, Illumina now evaluates the quality of the reads according to the percentage of bases having a base quality greater or equal than 30 (Q30).

The calculated error rates and Q30 are within specifications.

Channel	%Read		Error rate (PhiX)			Q30 (Channel)		
	Pass filter	Attributed	Read 1	Read 2	Average	Read 1	Read 2	Average
8	94.3%	98.6%	0.16%	0.36%	0.26%	95.8%	91.8%	93.8%

N.B.: We do not perform quality filters on sequence files that are within our quality specifications, besides the default filter done by the pipeline itself (i.e. “failed chastity”). We have observed that additional filters can alter the representation/variability of the sequence files (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.

Nomenclature:

Read: A sequence obtained after base calling. It's length is determined by the number of sequencing cycles.

Insert: Sample fragment that has been incorporated between two adapters during sample 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).

Mate-Pair libraries:

Duplicate and empty inserts removal for Mate-Pairs:

Mate pair libraries can have a relatively low total diversity (few millions independent constructs).

Therefore the dataset is screened for pair-read sharing exactly both reads on the first 30 bases. This can be expected to be due to PCR artifact and therefore only one copy of the pair is conserved (Unique pairs). The duplicate counts represent the total of sequences that were removed as they had already appeared in the data-set.

In parallel, the dataset is screened to eliminate the reads containing empty inserts (i.e. reads where the adapter sequence is found at the first base).

NB: Read 1, 3' adapter sequence used: “GATCGGAAGAGCGGTTCAGC”
Read 2, 3' adapter sequence used: “GATCGGAAGAGCGTCGTGTA”

The 3' adapter sequences are searched at the beginning of the reads, tolerating up to 2 mismatches in total.

The following table provide information about the ratio of duplicated reads and the number of unique empty inserts:

Fasteris code	Pair	Unique	% all reads	Duplicated	% all reads	Empty Inserts	% all reads
GLW-16	88'576'708	47'901'863	54.1%	40'674'845	45.9%	16'077	0.0%

The files have “RD30” (**r**emove **d**uplicate on **30** first bases on both ends) and “NotEmpty” (not empty insert construct) suffix added.

Linker search and removal:

The linker sequence was searched in the unique pairs, on both reads in both orientations, in 3 steps:

(NB: linker sequence used “ATAACTTCGTATAATGTATGCTATACGAAGTTAT”)

1) The complete sequence is searched without mismatch allowed.
2) If no match sequence is found, in successive steps the last base of the sequence is removed and it is searched at the end of the reads. The minimum size of 8 bases permits testing up to 42 bases.
3) Finally the remaining reads are search for not-exact matches of the linker. The first 5 bases are searched within the full reads sequences and at least 80% of the following bases must be identical to the sequence (max 44 bases).

NB: The reduce selectivity of 8 bases down to 6 when searching at the end can generate false positives.

After trimming of the linker sequence, the reads in which no match of the adapter sequences are found are retrieved. They are synchronized with the other end to provide only complete pair in order. The new files have a “NotLink” suffix added.

Library Code	File name	Reads	Bases
GLW-16	110415_SN365_B_s_8_1_seq_GLW-16.RD30.NotEmpty.NotLink.fastq	34'076'126	1'703'806'300
	110415_SN365_B_s_8_1_seq_GLW-16.RD30.NotEmpty.NotLink.fastq	34'076'126	1'703'806'300
	Total:	68'152'252	3'407'612'600

Files:

Index selected sequences e.g. 110415_SN365_B_s_8_1_seq_GLW-16.txt **Unique pairs without linker** e.g. 110415_SN365_B_s_8_1_seq_GLW-16.RD30.NotEmpty.fastq.NotLink
Format: **Illumina fast-q** *4 lines for each sequence:* *1- Unique identifier* *2- Sequence* *3- Orientation (always forward without mapping)* *4- Quality value for each base, corresponding to a Phred score*

The sequence files are sent on an external USB hard drive and on our secure server.

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