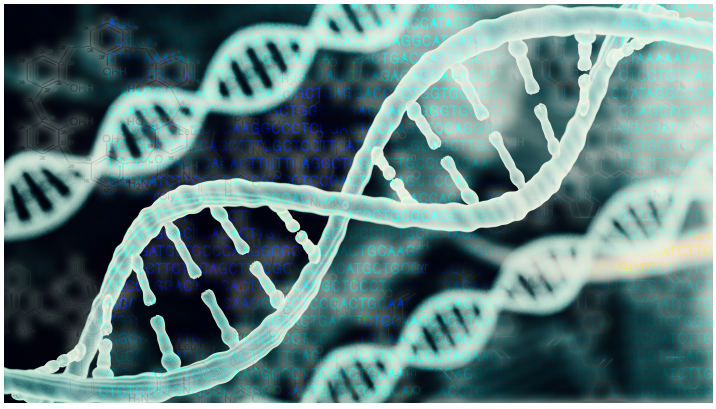
Source BioScience

Bioinformatics Report

Demo Report

Project SOUK000000



This report was generated by SR on November 17, 2019

Contents

Bcl2Fastq Conversion and Demultiplexing

Project Summary

Sequencing Workflow

Your samples have been successfully sequenced and processed. This means they passed all internal quality control (QC) steps. The full pipeline is shown in the flowcharts below.

Run Information

Sequencing Platform (Instr. ID)	MiSeq (762)
Run Setting	150bp Paired-end
Library Kit	TruSeq Stranded mRNA
Forward Adapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Reverse Adapter	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA
Total Number of Samples	26
PhiX Spike-in	0 %
Read Trimming	No trimming
Number of Lanes	1
Data Output Format[1]	FASTQ Phred+33 (Illumina 1.9)

R No data will be stored for longer than 12 weeks and after 12 weeks the data will be irretrievable by Source BioScience. The twelve week period begins when the data is written and not when the data is sent/posted to the customer. If multiple sets of data are delivered then the twelve week period will begin once the data is written for each individual data set. Extended data storage periods can be arranged upon request. Please contact us via email *technicalsales@sourcebioscience.com* for further information.

[1] More information about the FASTQ format can be found in the Appendix on page 14

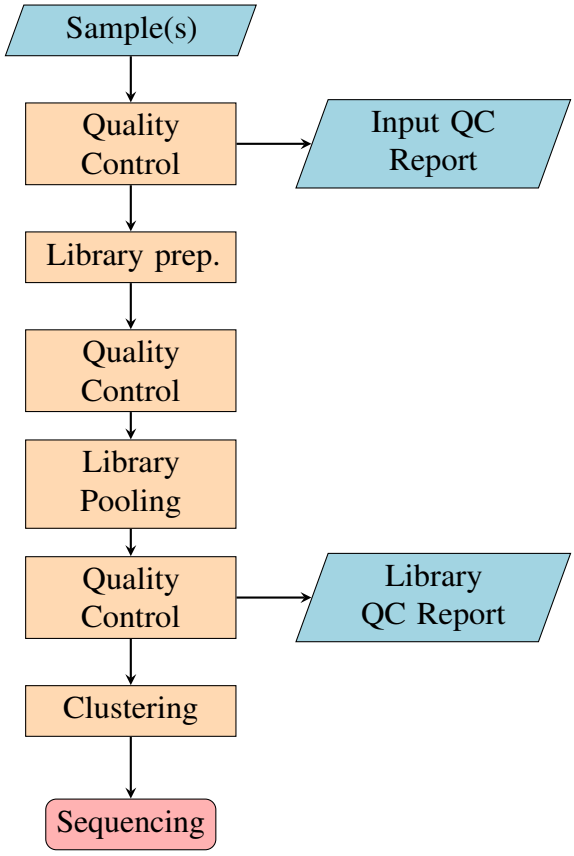


Figure 1.1: Laboratory workflow. Samples are processed and input material is carefully monitored throughout each step. You have received two QC reports reviewing concentration and quality of the material to predict sequencing performance as accurately as possible. Sequencing is not performed without customer approval in case minimum QC criteria are not met.

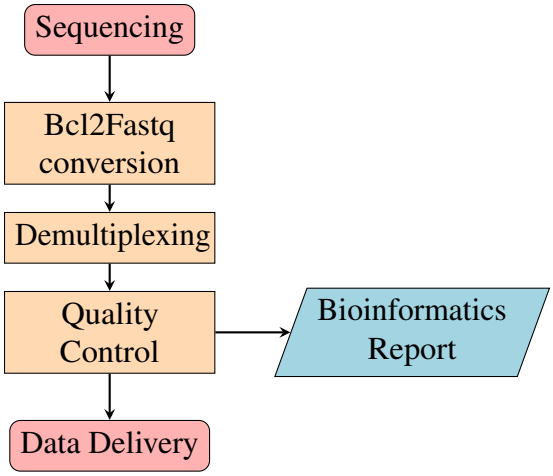


Figure 1.2: Sequencing and file conversion work flow.

Quality Report

The sequence data is provided as text files in FASTQ format[2], with one file per sample and respective sequencing primer used. Hence, for single end sequencing, you will see one file per sample with an extension "R1", containing forward primer derived reads. For paired end sequencing you will see an additional file with an extension "R2" per sample respectively. FASTQ files are named with the sample name and the sample number, which is a numeric assignment usually based on the order provided in the sample submission form. For example:

```
Sample1_S1_R1_001.fastq.gz
Sample1_S1_R2_001.fastq.gz
```

FASTQ is a format that provides per-base quality scores additionally to the called bases. These scores reflect the confidence for accuracy of a given base-call.

Flowcell 000000000-CR3CD

Lane	Sample	Barcode sequence	PF Clusters	Percent of the lane	Yield (Mbases)	Percent ≥ Q30 bases	Mean Quality Score
1	1	GGTTGCGAGG AATAGAG-CAA	100,289	0.69	30	94.96	36.51
1	10	GGTAACTCGC AAGTTG-GTGA	111,771	0.77	34	97.06	37.06
1	11	ACCGGCCGTA TG-GCAATATT	120,379	0.83	36	97.09	37.07

[2]Files are gzip compressed, an open source compression programme. For more information, see http://www.7-zip.org/ or http://www.gzip.org/

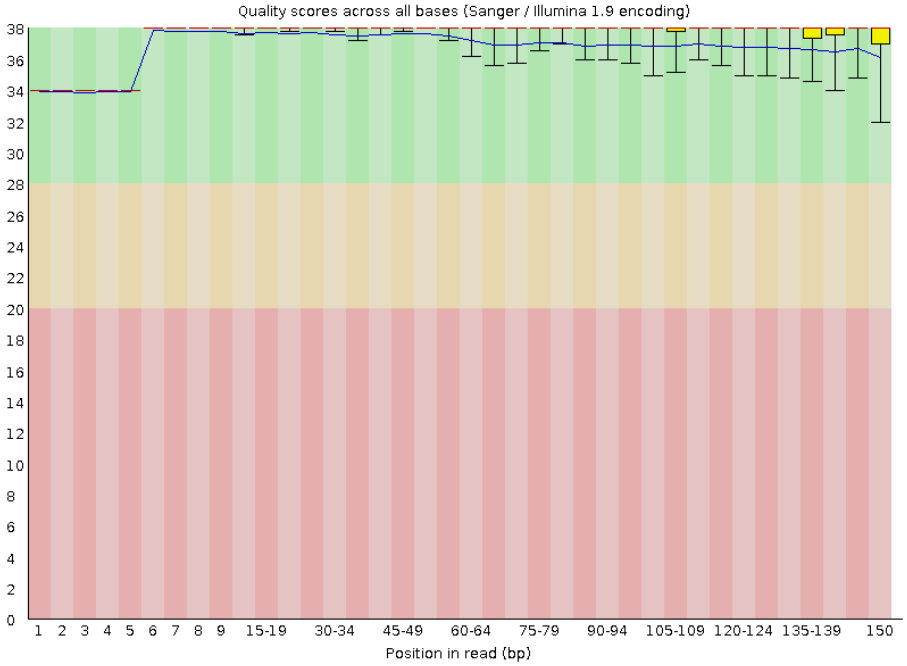
Lane	Sample	Barcode sequence	PF Clusters	Percent of the lane	Yield (Mbases)	Percent ≥ Q30 bases	Mean Quality Score
1	12	TGTAATCGAC-GATCAC-CGCG	96,352	0.66	29	96.20	36.83
1	13	GTGCAGACAG TACCATC-CGT	87,778	0.60	26	97.30	37.10
1	14	CAATCGGCTG GCTGTAG-GAA	108,710	0.75	33	97.85	37.26
1	15	TATGTAGTCA+ CGCAC-TAATG	105,360	0.72	32	97.99	37.29
1	16	ACTCGGCAAT· GACAACT-GAA	101,687	0.70	31	97.77	37.24
1	17	GTCTAATGGC-AGTG-GTCAGG	116,707	0.80	35	97.69	37.21
1	18	CCATCTCGCC-TTCTATG-GTT	91,780	0.63	28	96.97	37.01
1	19	CTGCGAGCCA AATCCG-GCCA	80,921	0.56	24	97.22	37.07
1	2	TAAGCATCCA-TCAATC-CATT	109,734	0.75	33	97.21	37.10
1	20	CGTTATTCTA+ CCATAAG-GTT	95,888	0.66	29	97.45	37.13
1	21	AGATCCATTA+ ATCTC-TACCA	105,713	0.73	32	97.76	37.24
1	22	GTCCTGGATA-CGGTGGC-GAA	106,477	0.73	32	97.56	37.17
1	23	CAGTGGCACT TAA-CAATAGG	107,165	0.74	32	98.02	37.30
1	24	AGTGTTGCAC· CTGGTA-CACG	109,827	0.75	33	97.83	37.26
1	25	GACACCATGT· TCAACGT-GTA	35,420	0.24	11	96.62	36.96
1	26	CCTGTCTGTC-ACT-GTTGTGA	52,182	0.36	16	96.26	36.87
1	3	ACCACGACAT· TCGTAT-GCGG	112,852	0.77	34	96.31	36.87
1	4	GCCGCACTCT· TCCGAC-CTCG	103,206	0.71	31	96.14	36.81
1	5	CCACCAGGCA CTTATG-GAAT	102,761	0.70	31	96.87	37.01
1	6	GTGACACGCA GCTTACG-GAC	87,174	0.60	26	95.22	36.55
1	7	ACAGTGTATG-GAACAT-ACGG	105,944	0.73	32	96.61	36.94

Lane	Sample	Barcode sequence	PF Clusters	Percent of the lane	Yield (Mbases)	Percent ≥ Q30 bases	Mean Quality Score
1	8	TGATTATACG+ GTCGAT-TACA	107,690	0.74	32	96.71	36.97
1	9	CAGCCGCGTA ACTAGC-CGTG	104,666	0.72	31	96.55	36.92

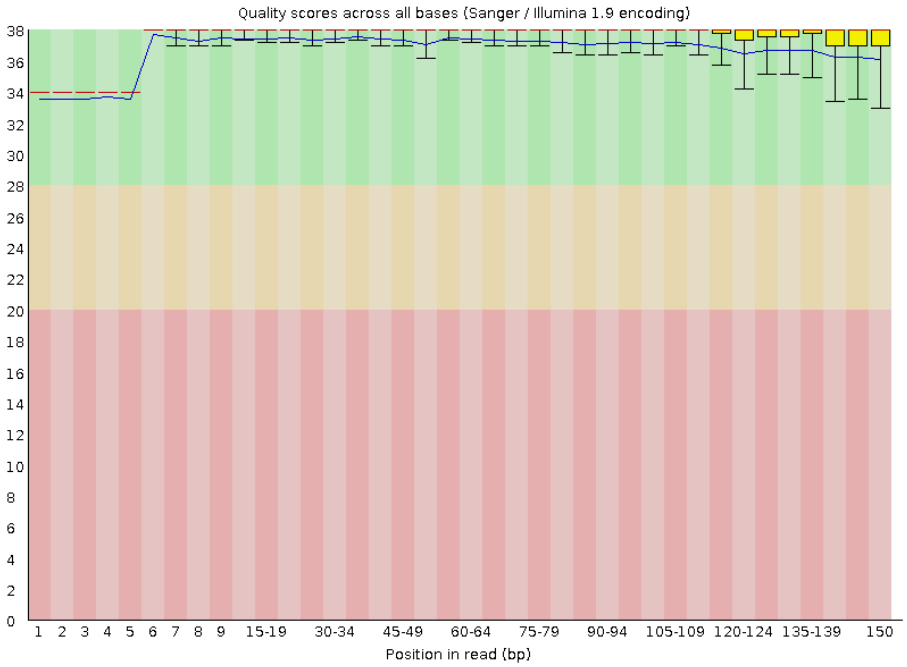
FASTQ Quality Plots

The entire sequencing run has been quality checked with FastQC[3], a tool for quality control of high-throughput sequence data. Hence, all reads per read direction were used to generate two summarising quality plots. Summary plots evaluating the quality scores across all bases of a run and the relative distribution of undetermined bases (N) are shown below.

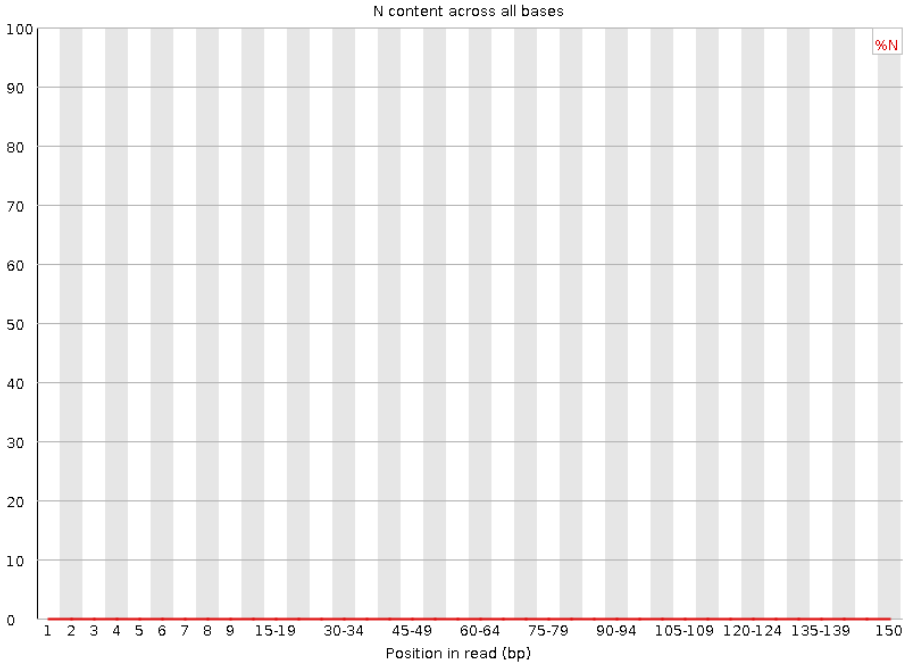
[3]More information on http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



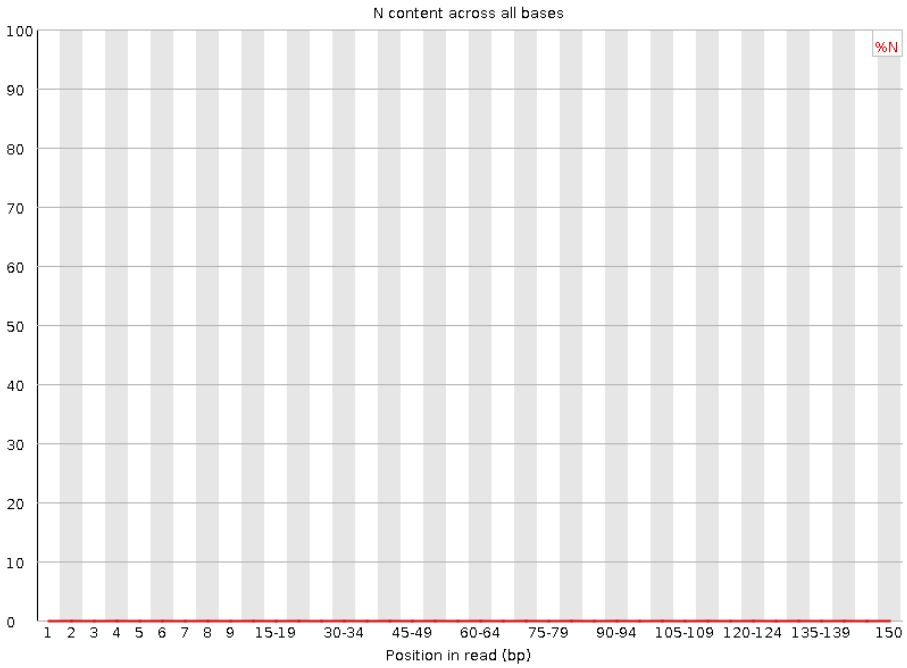
(A) Per base sequence quality plot generated for the **forward** read. Quality distribution is shown for every position in the read seperately. Illumina sequencing data shows a characteristic quality drop towards both ends. The longer the read length the lower the quality towards the 3' end.



(B) Per base sequence quality plot generated for the **reverse** read. Quality distribution is shown for every position in the read seperately. Illumina sequencing data shows a characteristic quality drop towards both ends. The longer the read length the lower the quality towards the 3' end.



(A) If the sequencing instrument was unable to determine a base it substitutes it with an N instead of a conventional base. The plot shows the relative number of Ns compared to the overall number of bases called along each position in the reads derived from the **forward** primer. The proportion of Ns is usually very low $< 5\%$. Higher numbers might indicate a problem with the sample/library or sequencing reagents.



(B) If the sequencing instrument was unable to determine a base it substitutes it with an N instead of a conventional base. The plot shows the relative number of Ns compared to the overall number of bases called along each position in the reads derived from the **reverse** primer. The proportion of Ns is usually very low $< 5\%$. Higher numbers might indicate a problem with the sample/library or sequencing reagents.

II Appendix

FASTQ File Format

Introduced by the Wellcome Trust Sanger Institute a FASTQ is a text file where nucleotide sequences are bundled together with estimated quality scores.

Definition 3.0.1 A quality score Q is defined as

$$Q = -10 \log_{10} p$$

where p is the probability that the called base is incorrect.

The relationship between the quality score and error probability can be read as follows:

Quality score	Probability that base is incorreect
10	0.1
20	0.01
30	0.001
40	0.0001

Since Illumina® pipeline version 1.8 the *Q*-score is encoded es ASCII 33 to 126 characters (ASCII+33). Officially described, however, are symbols only up to ASCII 73. Sequenced bases and qualities are bundled together with collected metainformation into four lines in a FASTQ file. Hence, a sequencing read is reported like this:

```
@NS500781:4:HCKH5BGXX:1:11101:1162:1050 1:N:0:13 Comment
CTGAGNAGCTGGGCTCCCGCTCTGGTGGGACACGCTGCCATCATTACTTTGATTAC
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEE
```

The header line starts always with '@' and contains colon- and space-separated metainformation created by the instrument, further described below. The second line is the actual read sequence. The third line usually startes with a +, but is empty otherwise. Line four contains the determined *Q*-scores in ASCII+33 encoding.

Field	Description
NS500781	sequencer ID
4	run ID
HCKH5BGXX	flow-cell ID
1	lane number
11101	tile within lane
1162	'*x*'-coordinate of cluster on tile
1050	'*y*'-coordinate of cluster on tile
1	member in a pair (forward: 1, reverse: 2)
N	Read filter (filtered: Y, pass: N)
0	control number (no control bits: 0, otherwise: even number)
13	index sequence or index id
Comment	Optional comment field. Usually empty.

FASTQ files are sorted according to their flow-cell coordinates. For paired-end runs it is ensured that both files have the same number of reads and the same order. If the other member in a pair, cannot be identified, a dummy read would be created, which is a read consisting only of Ns and *Q*-scores of 0. The following table demonstrates the relationship between the ASCII character and the *Q*-score:

Symbol	ASCII code	*Q*-Score	Symbol	ASCII code	*Q*-Score
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26
'	39	6	<	60	27
(	40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
,	44	11	A	65	32

-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			