Read Mapping

Read Mapping Exercise

Download fastq files from server

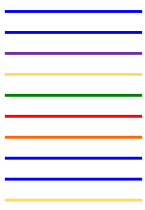
Quality assessment using FastQC

Map sequences to reference genome using bowtie

Visualize data in IGV

Read vs Sequence

10 reads, 6 sequences



What is the difference between a read and a sequence?

Quality Control

Assessing read quality

Phred quality score: a measure of the quality of base calling:
Q = -10 log(P) where P is the error probability

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality Control

Index sequence

Read 1
1 @D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
2 CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
3 +
4 CCCFFFFFHHHHHJIJGHIJJJIIJJJJJGGGFFFFEADDHHHFHFF@@DC>

Read 2
1 @D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
2 TCAATATTTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJJGFHIJJJJJJJJJJJJJJJFHHIIJJHGHJFGHJJJ:

Read 3
1 @D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
2 GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJJJJJJJJJIJJJJJJJJJJHIJJJJJGIIJFGIJJJ

Metzker, M.L. (2010) NRG

Line 1: sequence ID, description, and index; begins with @

Line 2: sequence; contains only A, C, T, G, and N

Line 3: optional sequence ID; begins with +

Line 4: signal quality of each base, cryptic code, ASCII_base 33 or 64

Quality Control

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

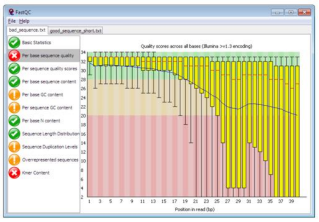
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

Quality Control

FastQC: a quality control tool for high-throughput sequencing data



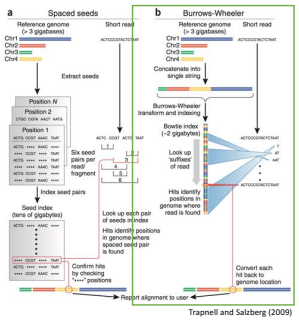
Mapping Reads

Table 1 A selection of short-read analysis software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

Trapnell and Salzberg (2009)

Bowtie



Trapnell and Salzberg (2009)

SAM/BAM Format

SAMtools: a software package for mining NextGen sequencing data after alignment.

SAM (Sequence Alignment Map): A widely use format for storing alignment data for high-throughput sequencing reads.

BAM (binary SAM): Compressed SAM (binary format).

The file is broken down into two sections:

1. **Header section (optional):** contains general information about the data such as alignment software used, reference genome aligned againsts, etc. Header lines start with @.
2. **Alignment section:** contains much of the same information as a fastq file, such as sequence and base quality scores, as well as information about alignment to reference sequence.

For a more complete description, see https://genome.sph.umich.edu/wiki/SAM

SAM/BAM Format

11 fields + optional 12th TAGs field (not shown)

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

Visualizing Genomics Data in a Genome Browser

Integrative Genomics Viewer (IGV)