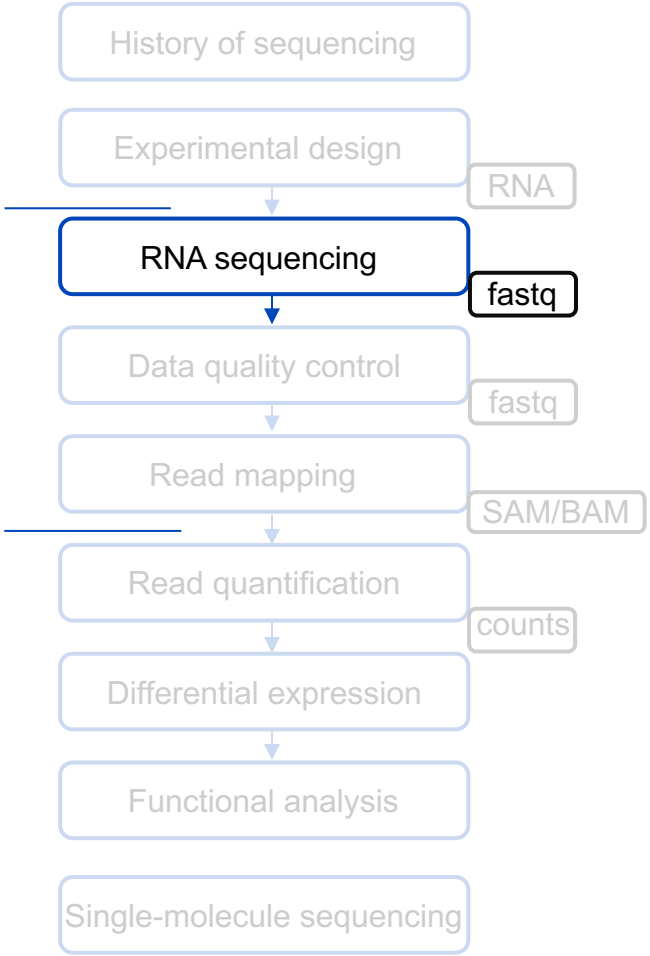
RNA sequencing

- part 1.2 -

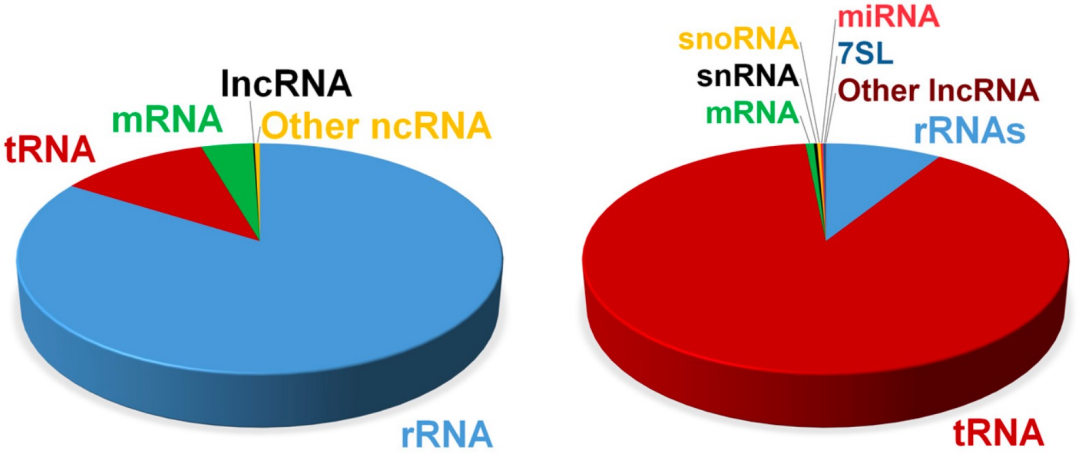
Deutsches Krebsforschungszentrum
Angewandte Bioinformatik (Prof. Dr. Benedikt Brors)
Dr. Óscar González-Velasco

Outline


History of sequencing
Experimental design
RNA
RNA sequencing
fastq
Data quality control
fastq
Read mapping
SAM/BAM
Read quantification
counts
Differential expression
Functional analysis
Single-molecule sequencing

RNA sequencing – in brief

Composition of RNA in mammalian cells



RNA by mass

RNA by number of molecules

Palazzo et al. (2015) *Front. Genetics 6(2)*

- poly-A enrichment
- rRNA depletion (e.g. riboZero)
- size selection for small RNAs

Sequence data format

FASTQ files

How to store sequencing data?

- FASTQ files – text-based ***raw sequencing data***
- SAM - Sequence Alignment Map (SAM) text-based format aligned data
- BAM – Binary compressed version of SAM
- CRAM - Compressed Reference-oriented Alignment Map for aligned data

CRAM COMPRESSION RATE

File format	File size (GB)
SAM	7.4
BAM	1.9
CRAM lossless	1.4
CRAM 8 bins	0.8
CRAM no quality scores	0.26

Illumina: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

Sequence data format

FASTQ files

How to store sequencing data?

- FASTQ files – text-based ***raw sequencing data***
- SAM - Sequence Alignment Map (SAM) text-based format aligned data
- BAM – Binary compressed version of SAM
- CRAM - Compressed Reference-oriented Alignment Map for aligned data

CRAM COMPRESSION RATE

File format	File size (GB)
SAM	7.4
BAM	1.9
CRAM lossless	1.4
CRAM 8 bins	0.8
CRAM no quality scores	0.26

Illumina: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

Sequence data format

FASTQ files

```
@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

- Text files with **.fastq** extension
- Standarised format across platforms
- Each read is constituted by 4 lines of text
- Size: often MB to GB

Illumina: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

Sequence data format

Line 1 ='@' character, followed by information about the sequencing run (also a sequence identifier)

```
@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

```
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>:<UMI> <read>:<is filtered>:<control number>:<index>
```

Table 1 FASTQ File Elements

Element	Requirements	Description
@	@	Each sequence identifier line starts with @.
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID.
<run number>	Numerical	Run number on instrument.
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number.
<tile>	Numerical	Tile number.
<x_pos>	Numerical	X coordinate of cluster.
<y_pos>	Numerical	Y coordinate of cluster.
<UMI>	Restricted characters: A/T/G/C/N	Optional, appears when UMI is specified in sample sheet. UMI sequences for Read 1 and Read 2, seperated by a plus [+].
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end.
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise.
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X and NextSeq systems, control specification is not performed and this number is always 0.
<index>	Restricted characters: A/T/G/C/N	Index of the read.

Illumina: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

Sequence data format

```
@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Line 2 = raw sequence letters

Illumina: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

Sequence data format

```
@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Line 3 = A separator, which is simply a plus (+) sign and is optionally followed by the info in Line 1

Illumina: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

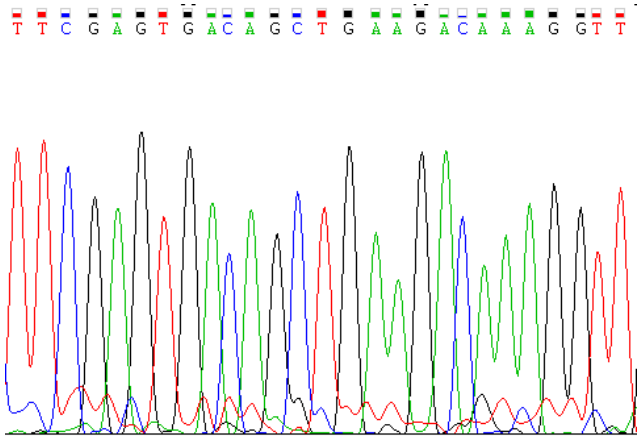
Sequence data format

```
@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Line 4 = encodes (ASCII) the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

Illumina: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

Phred scores



Phred quality scores measure base call accuracy

- signal intensity
- signal to noise ratio
- base position and composition of the read

accuracy limited owing to

- low intensity
- low diversity (many of the same color) => especially for first few bases

Further reading: Phred algorithm (Ewing and Green 1998, Genome Res. 8(3):186)

Phred scores

Phred quality scores measure base call accuracy

P: error probability of a given base call $\mathbf{P_{err} = 10^{(-Q/10)}}$

Phred score $\mathbf{Q = -10 \cdot \log_{10} P}$

Phred scores

Phred quality scores measure base call accuracy

P: error probability

Phred score $\mathbf{Q = -10 \cdot \log_{10} P}$

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%

Harvard Chen Bioinformatics Core Training Material

Phred scores

Phred quality scores measure base call accuracy

P: error probability

Phred score $\mathbf{Q = -10 \cdot \log_{10} P}$

RNA-seq

DNA-seq

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%

Harvard Chen Bioinformatics Core Training Material

Phred scores

Phred quality scores measure base call accuracy

Phred scores are stored as ASCII characters in the FASTQ file

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
!	33	0	1.00E+00	0.000%	1	2,858,034,764
"	34	1	7.94E-01	20.567%	1	2,270,217,709
#	35	2	6.31E-01	36.904%	2	1,803,298,025
$	36	3	5.01E-01	49.881%	2	1,432,410,537
%	37	4	3.98E-01	60.189%	3	1,137,804,133
&	38	5	3.16E-01	68.377%	3	903,789,949
`	39	6	2.51E-01	74.881%	4	717,905,874
(	40	7	2.00E-01	80.047%	5	570,252,906
)	41	8	1.58E-01	84.151%	6	452,967,984
*	42	9	1.26E-01	87.411%	8	359,805,259
+	43	10	1.00E-01	90.000%	10	285,803,476
,	44	11	7.94E-02	92.057%	13	227,021,771
-	45	12	6.31E-02	93.690%	16	180,329,803
.	46	13	5.01E-02	94.988%	20	143,241,054
/	47	14	3.98E-02	96.019%	25	113,780,413
0	48	15	3.16E-02	96.838%	32	90,378,995
1	49	16	2.51E-02	97.488%	40	71,790,587
2	50	17	2.00E-02	98.005%	50	57,025,291
3	51	18	1.58E-02	98.415%	63	45,296,798
4	52	19	1.26E-02	98.741%	79	35,980,526
5	53	20	1.00E-02	99.000%	100	28,580,348
6	54	21	7.94E-03	99.206%	126	22,702,177
7	55	22	6.31E-03	99.369%	158	18,032,980

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
8	56	23	5.01E-03	99.499%	200	14,324,105
9	57	24	3.98E-03	99.602%	251	11,378,041
:	58	25	3.16E-03	99.684%	316	9,037,899
;	59	26	2.51E-03	99.749%	398	7,179,059
<	60	27	2.00E-03	99.800%	501	5,702,529
=	61	28	1.58E-03	99.842%	631	4,529,680
>	62	29	1.26E-03	99.874%	794	3,598,053
?	63	30	1.00E-03	99.900%	1,000	2,858,035
@	64	31	7.94E-04	99.921%	1,259	2,270,218
A	65	32	6.31E-04	99.937%	1,585	1,803,298
B	66	33	5.01E-04	99.950%	1,995	1,432,411
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253
G	71	38	1.58E-04	99.984%	6,310	452,968
H	72	39	1.26E-04	99.987%	7,943	359,805
I	73	40	1.00E-04	99.990%	10,000	285,803
J	74	41	7.94E-05	99.992%	12,589	227,022
K	75	42	6.31E-05	99.994%	15,849	180,330
L	76	43	5.01E-05	99.995%	19,953	143,241
M	77	44	3.98E-05	99.996%	25,119	113,780
N	78	45	3.16E-05	99.997%	31,623	90,379
O	79	46	2.51E-05	99.997%	39,811	71,791

https://drive5.com/usearch/manual/quality_score.html

Phred scores

Phred quality scores measure base call accuracy

Phred scores are stored as ASCII characters in the FASTQ file

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
!	33	0	1.00E+00	0.000%	1	2,858,034,764
"	34	1	7.94E-01	20.567%	1	2,270,217,709
#	35	2	6.31E-01	36.904%	2	1,803,298,025
$	36	3	5.01E-01	49.881%	2	1,432,410,537
%	37	4	3.98E-01	60.189%	3	1,137,804,133
&	38	5	3.16E-01	68.377%	3	903,789,949
`	39	6	2.51E-01	74.881%	4	717,905,874
(	40	7	2.00E-01	80.047%	5	570,252,906
)	41	8	1.58E-01	84.151%	6	452,967,984
*	42	9	1.26E-01	87.411%	8	359,805,259
+	43	10	1.00E-01	90.000%	10	285,803,476
,	44	11	7.94E-02	92.057%	13	227,021,771
-	45	12	6.31E-02	93.690%	16	180,329,803
.	46	13	5.01E-02	94.988%	20	143,241,054
/	47	14	3.98E-02	96.019%	25	113,780,413
0	48	15	3.16E-02	96.838%	32	90,378,995
1	49	16	2.51E-02	97.488%	40	71,790,587
2	50	17	2.00E-02	98.005%	50	57,025,291
3	51	18	1.58E-02	98.415%	63	45,296,798
4	52	19	1.26E-02	98.741%	79	35,980,526
5	53	20	1.00E-02	99.000%	100	28,580,348
6	54	21	7.94E-03	99.206%	126	22,702,177
7	55	22	6.31E-03	99.369%	158	18,032,980

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
8	56	23	5.01E-03	99.499%	200	14,324,105
9	57	24	3.98E-03	99.602%	251	11,378,041
:	58	25	3.16E-03	99.684%	316	9,037,899
;	59	26	2.51E-03	99.749%	398	7,179,059
<	60	27	2.00E-03	99.800%	501	5,702,529
=	61	28	1.58E-03	99.842%	631	4,529,680
>	62	29	1.26E-03	99.874%	794	3,598,053
?	63	30	1.00E-03	99.900%	1,000	2,858,035
@	64	31	7.94E-04	99.921%	1,259	2,270,218
A	65	32	6.31E-04	99.937%	1,585	1,803,298
B	66	33	5.01E-04	99.950%	1,995	1,432,411
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253
G	71	38	1.58E-04	99.984%	6,310	452,968
H	72	39	1.26E-04	99.987%	7,943	359,805
I	73	40	1.00E-04	99.990%	10,000	285,803
J	74	41	7.94E-05	99.992%	12,589	227,022
K	75	42	6.31E-05	99.994%	15,849	180,330
L	76	43	5.01E-05	99.995%	19,953	143,241
M	77	44	3.98E-05	99.996%	25,119	113,780
N	78	45	3.16E-05	99.997%	31,623	90,379
O	79	46	2.51E-05	99.997%	39,811	71,791

https://drive5.com/usearch/manual/quality_score.html

Sequence data format

FASTQ files

Read ID
Sequence
+
Quality score
(Phred + 33)

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

Phred score is provided for each base

Illumina: https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Sequence data format

FASTQ files

Read ID
Sequence
+
Quality score
(Phred + 33)

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

Phred score is provided for each base

C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253

69 – 33(ASCII) = Q score 36

Illumina: https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Sequence data format

FASTQ files

Read ID
Sequence
+
Quality score
(Phred + 33)

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

Phred score is provided for each base

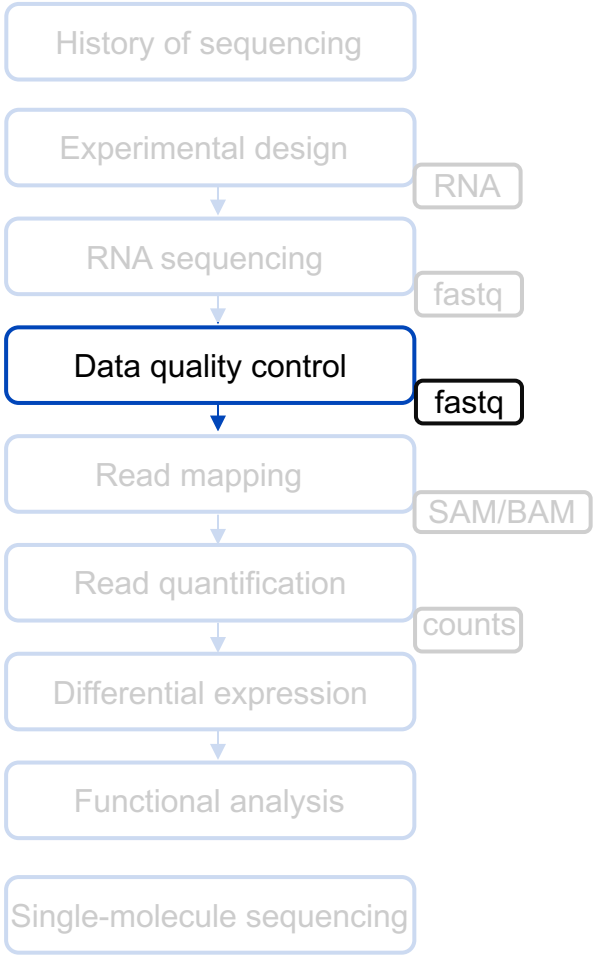
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253

69 – 33(ASCII) = Q score 36

$$P_{err} = 10^{(-Q/10)} = 10^{(-36/10)} = \boxed{0.0002511886}$$

Illumina: https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Outline


History of sequencing
Experimental design
RNA
RNA sequencing
fastq
Data quality control
fastq
Read mapping
SAM/BAM
Read quantification
counts
Differential expression
Functional analysis
Single-molecule sequencing

Data quality assessment

```
@HWI-M03127:41:ACE13:1:2109:11596:14331 1:N:0:GGAGACAAGGGA
TACGGAGGGTGCGAGCGTTGTTCGGAATTATTGGGCGTAAAGCGCGTGTAGGCGGTTTGTTAAGTCTGGTGTGAAAGCCCTGGGC
TCAACCTGGGAAGTGCATTGGATACTGGCAAACTTGAGTACGGGAGAGGATAGTGGAATTTCGAGTGTAGGGGTGAAATCCGTAG
ATATTCGAAGGAACACCGGTGGCGAAGGCGGCTATCTGGACCGATACTGACGCTGGGACGCGAAAGCGTGGGGAGCAAACAGG
+
BBBBBBBDDBBBGAEGGEEGGGHGFEFFHHHHHHHGGGGGHHHGGGFGGHHGHGGGGGGHHFFHHGHHGHHHGFHHHHHHHGHGH
HGGGGHHHHGHHGGGGHHHHHHHHHHHHHHHHHGHHHGHHHHHGCGFHGGHHHHHGHHHHHHHGH1FEGEHHHHGD?E/EFDEGE
HGHGAHHG2HGAE>>/FHGGFGGG/EEEGFFHGF1G1GGGEEDBHGAEEB/CGB?EEAGEEB00CECGGFFFFBB1FFAAAA>
@HWI-M03127:41:ACE13:1:1113:6675:5716 1:N:0:GGAGACAAGGGA
GACGTAGGGGGCCAGCGTTGTTCGGAACTACTGGGTGTAAAGGGTTCGTAGGCGGTGCGGCAAGTTGGGAGTGAAATCTCTGGGC
TTAACCCAGAGGCTGCTTCCAAAACTGCTGTGCTCGAGTGTGAGAGAGGCGCGTGGAATTGCAGGTGTAGCGGTGAAATGCGTAG
ATATCTGCAGGAACACCCGTGGCGAAAGCGGCGCGCTGGATCACTACTGACGCTGAGGAACGAAAGCTAGGGGAGCAAACAGG
+
11AAA1F?1ADAEEGGGAEEFED0/AFEHHFHHHHGAEHHHHF?GGGGGHGEGGGGEE/>>E/FHG1FCGGGFHHHHHHHGHHHF
GGHHHHEFECG/CGHGHHFCG><GHHFHHHHHHHHGCHGGGHHFHHGGGGGG1<ACBHHHGGHGH1@EEE/HGBB2CEA>/GDGF
FEB1BFB1DEF?//>///>B/EEGFA/EAAA?EEGB1FHGCFHHFB0AEA?EFGFHGE0B0F3BEBGGGGFFDDBC3F>>>A>
@HWI-M03127:41:ACE13:1:2108:7969:19134 1:N:0:GGAGACAAGGGA
TACGGAGGATGCAAGCGTTATCCGGATTTACTGGGTTTAAAGGGTGCGTAGGTGGGTCTGTAAGTCAGTGGTGAAATCTCCGAGC
TTAACTCGGAAACTGCCATTGATACTATAGGTCTTGAATTATCTGGAGGTAAGCGGAATATGTCATGTAGCGGTGAAATGCTTAG
ATATGACATAGAACACCAATTGCGAAGGCAGCTGGCTACACAAATATTGACACTGAGGCACGAAAGCGTGGGGATCAAACAGG
```

https://2017-lapaz-assembly.readthedocs.io/en/latest/qiime1.html

Data quality assessment

```
@HWI-M03127:41:ACE13:1:2109:11596:14331 1:N:0:GGAGACAAGGGA
TACGGAGGGTGCGAGCGTTGTTCGGAATTATTGGGCGTAAAGCGCGTGTAGGCGGTTTGTTAAGTCTGGTGTGAAAGCCCTGGGC
TCAACCTGGGAAGTGCATTGGATACTGGCAAACTTGAGTACGGGAGAGGATAGTGGAATTTCGAGTGTAGGGGTGAAATCCGTAG
ATATTCGAAGGAACACCGGTGGCGAAGGCGGCTATCTGGACCGATACTGACGCTGGGACGCGAAAGCGTGGGGAGCAAACAGG
+
BBBBBBBDDBBBGAEGGEEGGGHGFEFFHHHHHHHGGGGGHHHGGGFGGHHGHGGGGGGHHFFHHGHHGHHHGFHHHHHHHGHGH
HGGGGHHHHGHHGGGGHHHHHHHHHHHHHHHHHGHHHGHHHHHGCGFHGGHHHHHGHHHHHHHGH1FEGEHHHHGD?E/EFDEGE
HGHGAHHG2HGAE>>/FHGGFGGG/EEEGFFHGF1G1GGGEEDBHGAEEB/CGB?EEAGEEB00CECGGFFFFBB1FFAAAA>
@HWI-M03127:41:ACE13:1:1113:6675:5716 1:N:0:GGAGACAAGGGA
GACGTAGGGGGCCAGCGTTGTTCGGAACTACTGGGTGTAAAGGGTTCGTAGGCGGTGCGGCAAGTTGGGAGTGAAATCTCTGGGC
TTAACCCAGAGGCTGCTTCCAAAACTGCTGTGCTCGAGTGTGAGAGAGGCGCGTGGAATTGCAGGTGTAGCGGTGAAATGCGTAG
ATATCTGCAGGAACACCCGTGGCGAAAGCGGCGCGCTGGATCACTACTGACGCTGAGGAACGAAAGCTAGGGGAGCAAACAGG
+
11AAA1F?1ADAEEGGGAEEFED0/AFEHHFHHHHGAEHHHHF?GGGGGHGEGGGGEE/>>E/FHG1FCGGGFHHHHHHHGHHHF
GGHHHHEFECG/CGHGHHFCG><GHHFHHHHHHHHGCHGGGHHFHHGGGGGG1<ACBHHHGGHGH1@EEE/HGBB2CEA>/GDGF
FEB1BFB1DEF?//>///>B/EEGFA/EAAA?EEGB1FHGCFHHFB0AEA?EFGFHGE0B0F3BEBGGGGFFDDBC3F>>>A>
@HWI-M03127:41:ACE13:1:2108:7969:19134 1:N:0:GGAGACAAGGGA
TACGGAGGATGCAAGCGTTATCCGGATTTACTGGGTTTAAAGGGTGCGTAGGTGGGTCTGTAAGTCAGTGGTGAAATCTCCGAGC
TTAACTCGGAAACTGCCATTGATACTATAGGTCTTGAATTATCTGGAGGTAAGCGGAATATGTCATGTAGCGGTGAAATGCTTAG
ATATGACATAGAACACCAATTGCGAAGGCAGCTGGCTACACAAATATTGACACTGAGGCACGAAAGCGTGGGGATCAAACAGG
```

https://2017-lapaz-assembly.readthedocs.io/en/latest/qiime1.html

Errors

Sample
- contamination
- fixation artifacts

Library preparation (bias)
- fragmentation
- ligation
- amplification
- GC bias
- Contamination

Sequencing
- chemical
- optical read-out
- computational

Instrument	Primary Errors	Single-pass Error Rate (%)	Consensus Error Rate (%)
ABI 3730xl (capillary)	substitutions	0.1-1	0.1-1
Roche 454 – All models	indels	1	1
Illumina – All models	substitutions	~0.1	~0.1
Ion Torrent – all chips	indels	~1	≤1
Oxford Nanopore	deletions	2-3	~0.1
PacBio RS	indels	13-15	~0.1

https://www.sciencedirect.com/science/article/pii/S0198885921000628
https://www.nature.com/articles/s41598-018-29325-6
https://nanoporetech.com/

Data quality control software

FastQC

- read/sequence/mapping quality, nucleotide distribution, bias
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTX

- quality statistics, nucleotide distribution
- http://hannonlab.cshl.edu/fastx_toolkit/

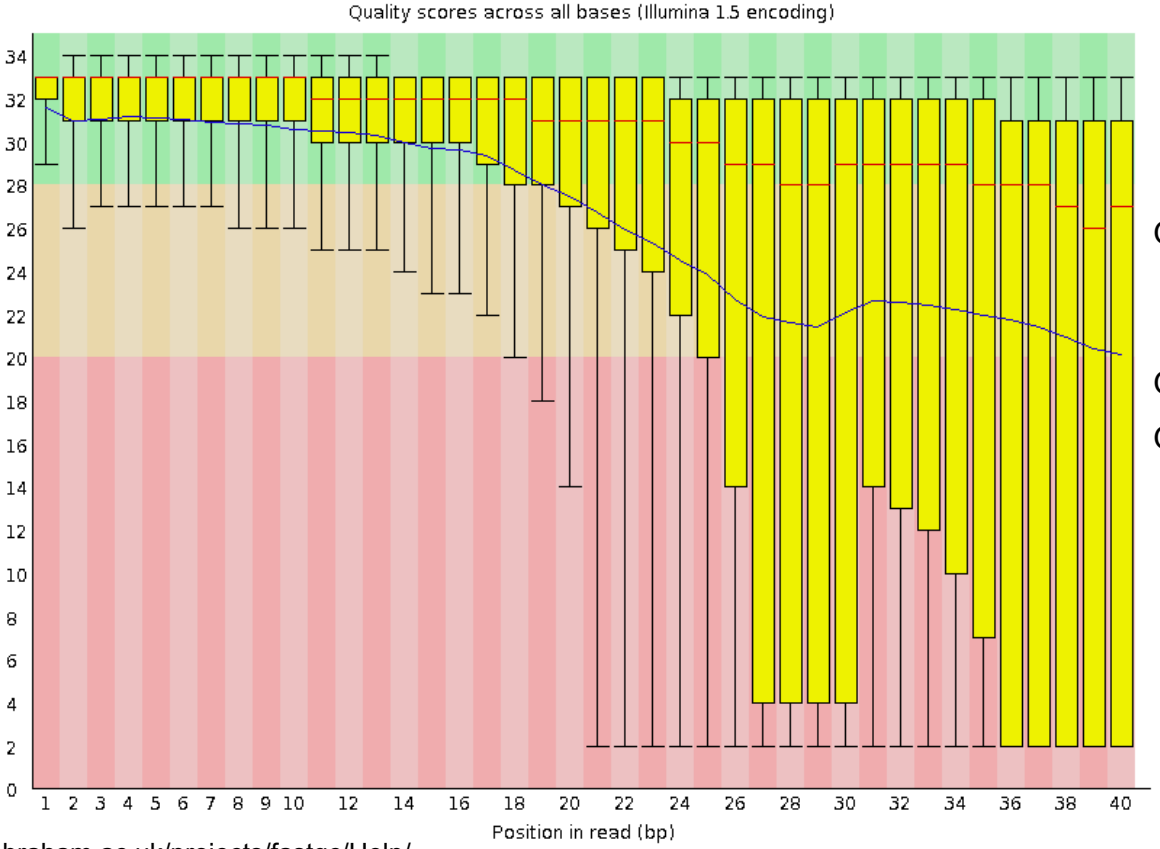
MultiQC

- combine QC results of different samples
- https://multiqc.info/

FastqScreen

- contamination, search against set of libraries
- https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

FastQC: Sequence quality per base



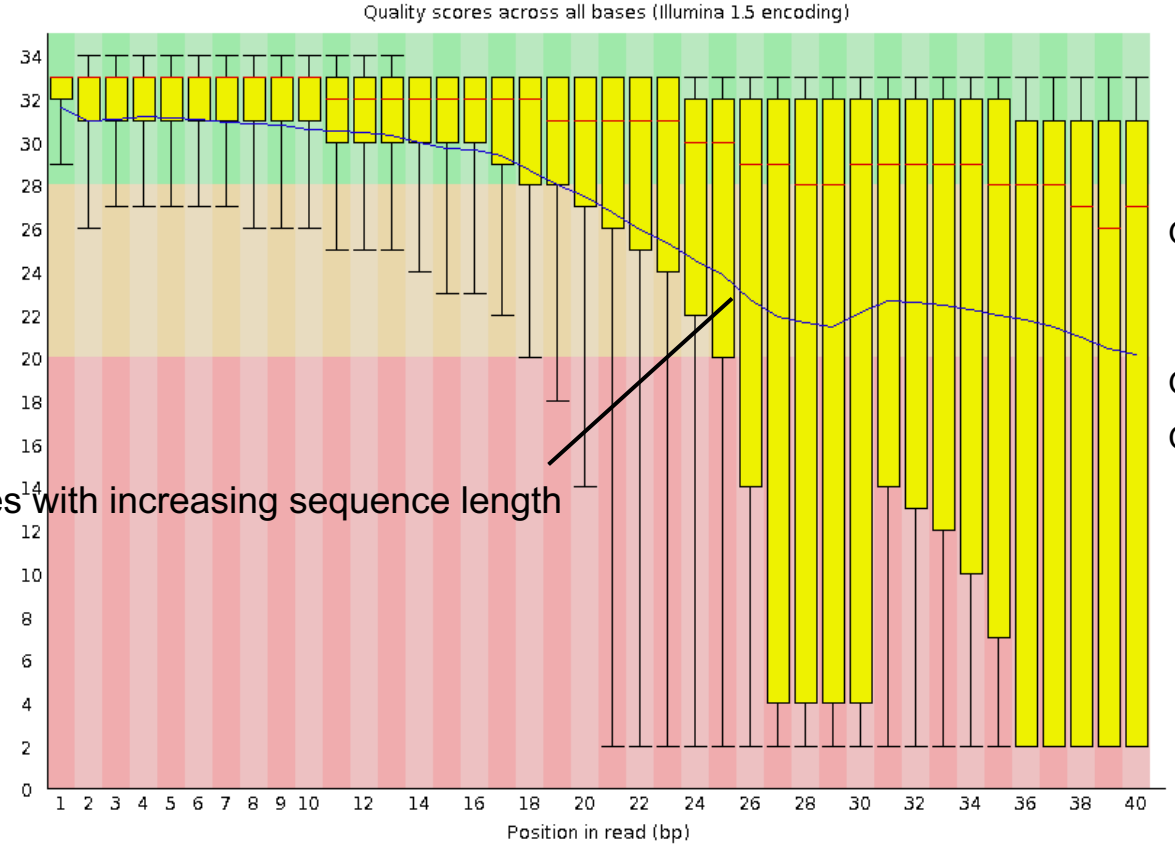
Q>28 => good

Q>20 => reasonable

Q<20 => poor

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

FastQC: Sequence quality per base



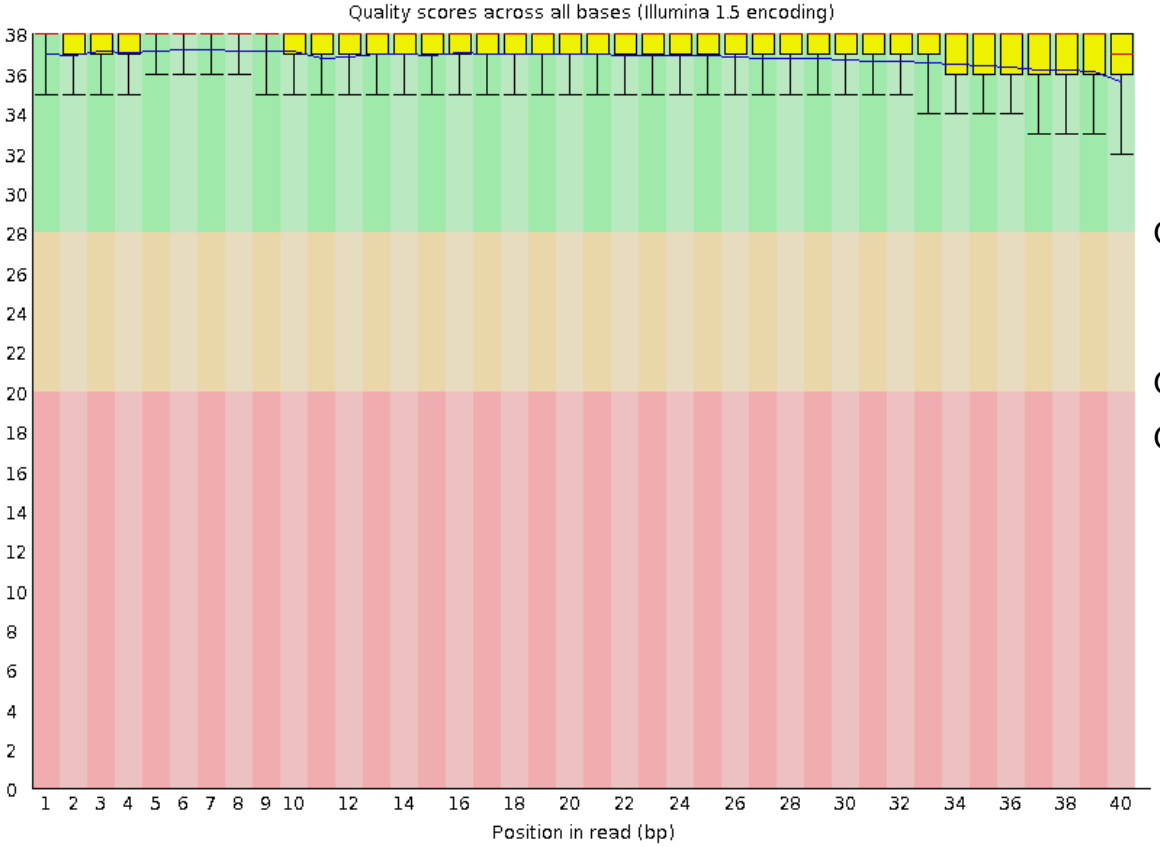
Quality decreases with increasing sequence length

Q>28 => good

Q>20 => reasonable

Q<20 => poor

FastQC: Sequence quality per base



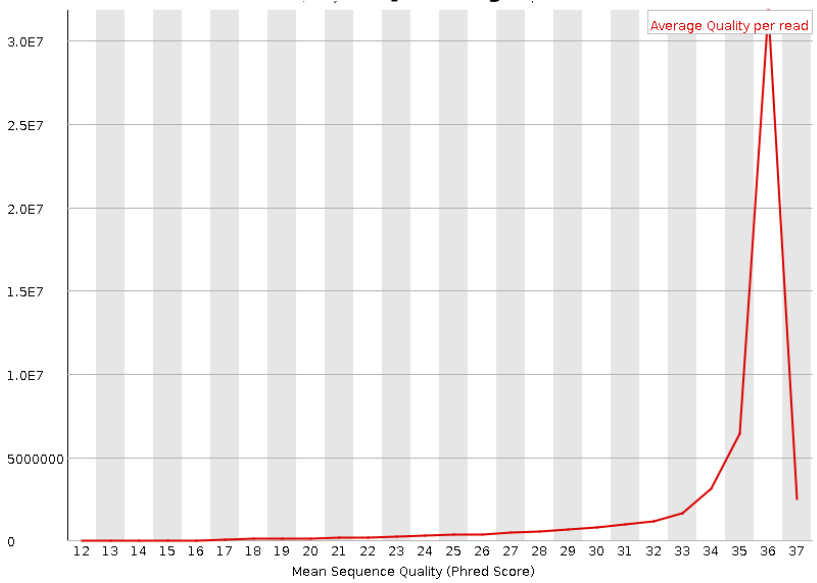
Q>28 => good

Q>20 => reasonable

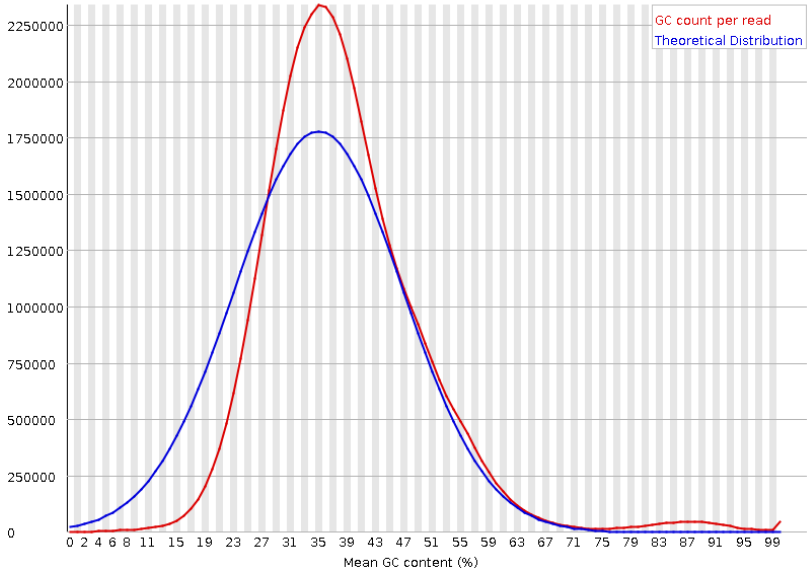
Q<20 => poor

FastQC: read quality

Distribution of quality scores

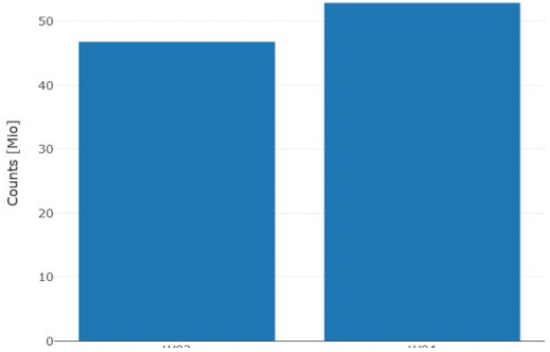


GC content

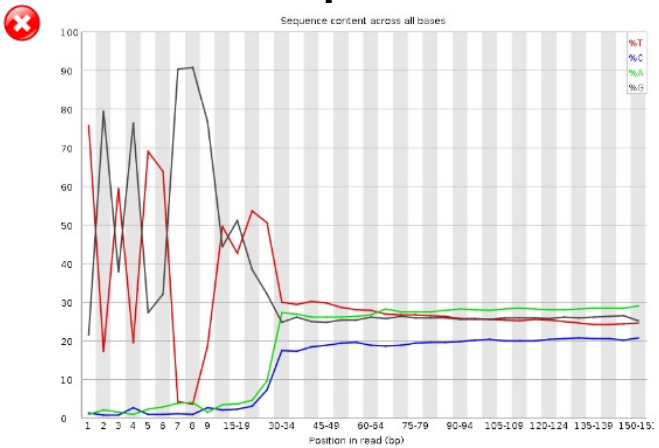


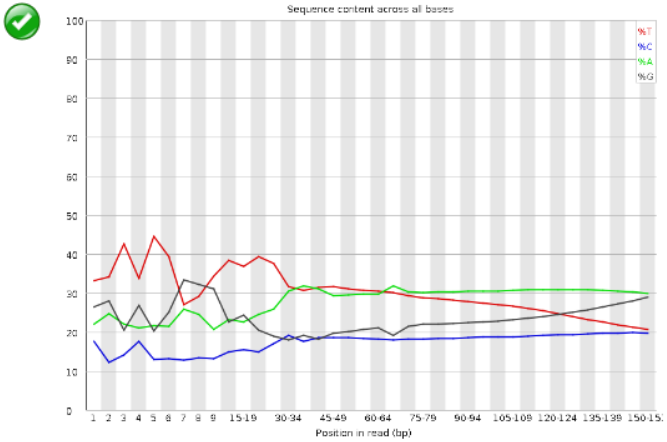
FastQC: read quality

Total number of reads



Per base sequence content





Overrepresented sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGAAGAGAATCTCGGTT	628359	1.3439386795454222	TruSeq Adapter, Index 2 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGAAGAGAATCTCGGGT	509641	1.0900237803265467	TruSeq Adapter, Index 2 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGAAGAGAATCGCGGTT	222525	0.4759380460307644	TruSeq Adapter, Index 2 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGAAGAGAATCGCGGGT	143867	0.3077037585363801	TruSeq Adapter, Index 2 (97% over 36bp)

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

Data preprocessing

Trimming (removal of bases from beginning/end)

- low quality reads
- adapter sequences

Filtering (removal of bad reads)

- low quality reads
- contamination sequences
- repeats
- short reads (<20 bp slow down mapping)

Masking

- substitute low quality base calls by “N“

Preprocessing software

PRINSEQ
- quality control, data preprocessing (trimming, filtering, reformating)
- http://prinseq.sourceforge.net/

Trimmomatic
- adapter trimming, quality filtering
- Bolger et al. (2014) Bioinformatics
- http://www.usadellab.org/cms/?page=trimmomatic

FlexBar (FAR)
- adapter removal, barcode detection
- Dodt et al. (2012) Biology 1(3):895
- https://sourceforge.net/projects/flexbar/

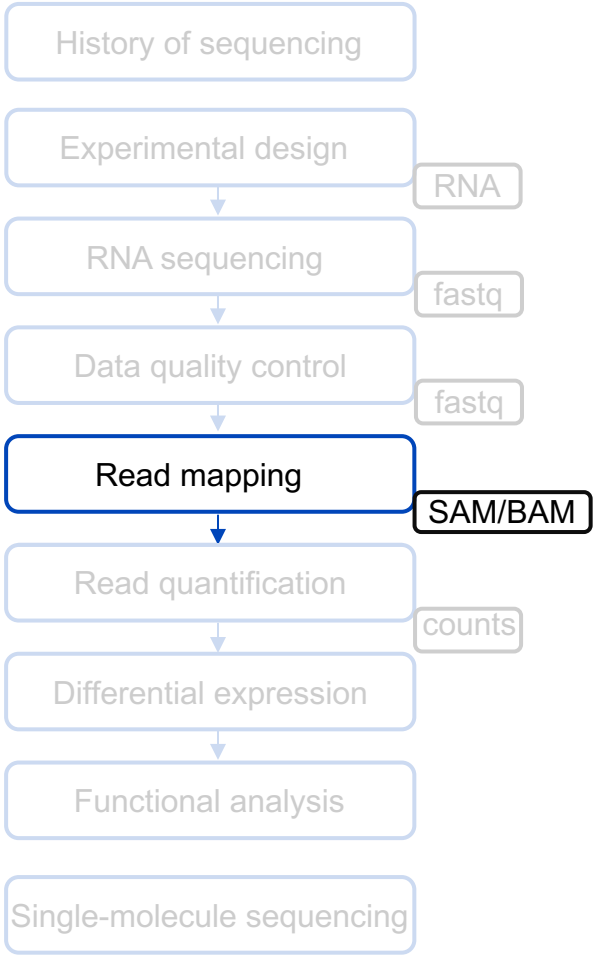
FASTX
- quality statistics, trimming, adapter removal, quality filtering, demultiplexing
- http://hannonlab.cshl.edu/fastx_toolkit/

TagCleaner
- tag prediction, adapter trimming, demultiplexing
- http://tagcleaner.sourceforge.net/

DeconSeq
- removal of contaminating sequences
- http://deconseq.sourceforge.net/

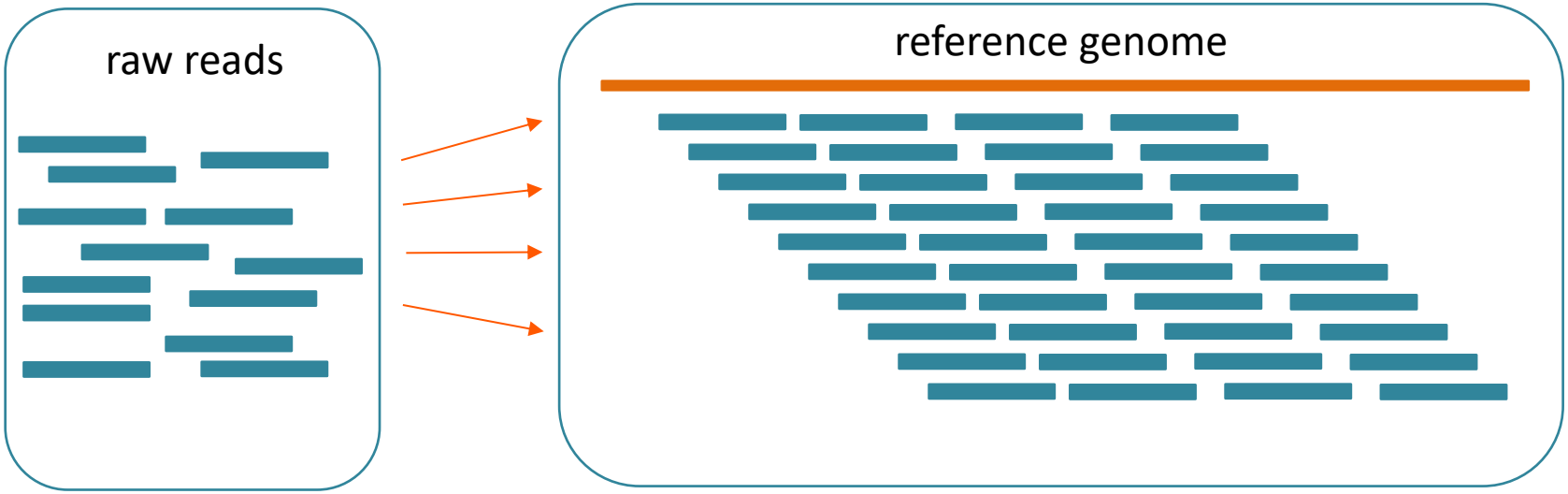
Adapted from Functional Genomics Center Zurich

Outline


History of sequencing
Experimental design
RNA
RNA sequencing
fastq
Data quality control
fastq
Read mapping
SAM/BAM
Read quantification
counts
Differential expression
Functional analysis
Single-molecule sequencing

Read mapping

Locate reads with respect to a reference sequence



Spliced versus unspliced aligners

unspliced aligners

- align continuous reads
- without splice gaps
  => when aligning exon to reference genome: at the site of an intron mismatches appear
  => aligner stops and trims rest of the read
- e.g. BLAST, BWA, Bowtie, Soap

Spliced versus unspliced aligners

unspliced aligners

- align continuous reads
- without splice gaps
  => when aligning exon to reference genome: at the site of an intron mismatches appear
  => aligner stops and trims rest of the read
- e.g. BLAST, BWA, Bowtie, Soap

spliced aligners

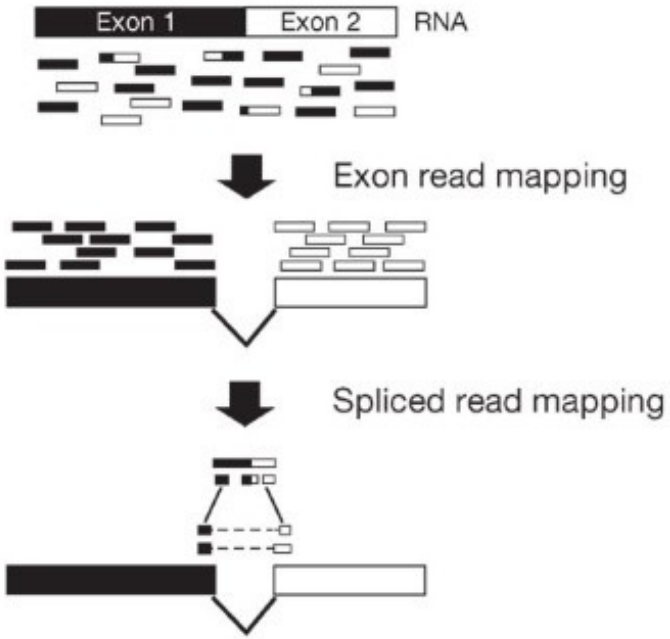
- allows for introns
- exon-first approach: some aligners employ unspliced aligners first, then map the rest (split reads into smaller segments and align independently)
  $\Rightarrow$ reads are not trimmed at the beginning of an intron
  $\Rightarrow$ possible to detect splice junctions
- e.g. BLAT, GMAP, STAR, TopHat

Narrandes et al. (2018) J. of Cancer 9(13):2249

Spliced aligners: exon first

- step 1: alignment of exonic reads
- step 2: split remaining reads into smaller pieces and map to genome

- fast
- less computational resources
- e.g. TopHat, SpliceMap, MapSplice, STAR



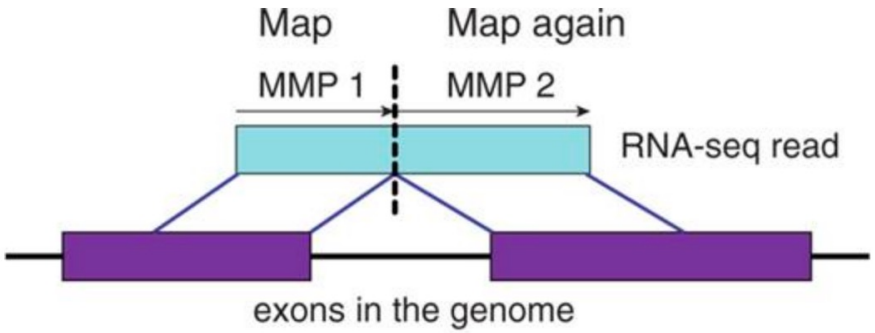
Garber et al. (2011) Nat. Methods 8(6):469

Spliced aligners: STAR

Spliced Transcripts Alignment to a Reference (STAR)

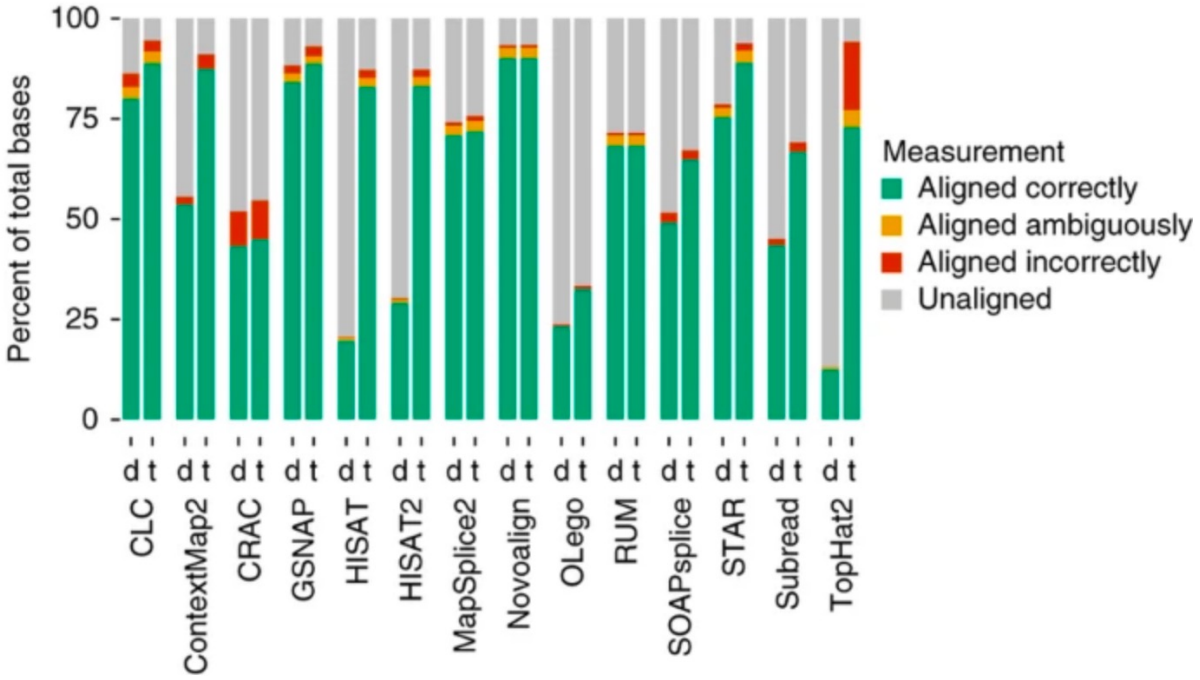
- based on BWT
- STAR looks for longest exact match for each read (Maximal Mappable Prefixes: MMP)
- seeds: different parts of read that are mapped separately
- sequential search of unmapped parts of read
- extension of MMPs if mismatches occur

- works for short and long reads
- very high mapping speed
- error tolerant
- memory intensive



Dobin et al. (2012) Bioinformatics

Comparison aligners



Barruzo et al. (2016) Nature Methods 14:135

Mapping output

Sequence Alignment/Map (SAM)

- text format
- storage of read alignments to reference genome

Binary Alignment/Map (BAM)

- “binary SAM“
- compressed, containing index (bai) (~25% of sam file)

  => index for fast retrieval of alignments

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

https://medium.com/@shilparaopradeep/samtools-guide-learning-how-to-filter-and-manipulate-with-sam-bam-files-2c28b25d29e8

Mapping output

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

Header lines starting with @ symbol describing various metadata for *all* reads

```
@HD VN:1.6 SO:coordinate        – BAM header line
@SQ SN:chr1 LN:248956422        – Reference sequence dictionary entries
@SQ SN:chr2 LN:242193529
@RG ID:RG1 SM:SAMPLE_A          – Read group(s)
```

Records containing structured read information (1 line per read/record)

Read name	Flags	Position	MAPQ	CIGAR	Mate information	Read sequence	Phred quality scores	Metadata
SlX1:1:127:63:4	99	chr1 10052169	60	23M3D10M	= 14 10	GAAGATACTGGTT	768832'48::::	RG:Z:A ...

– **Mapping** information summarizes **position, quality, and structure** for each **read**

– **Mate** information points to the **other read in a pair**

https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

Mapping output

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

Header lines starting with @ symbol describing various metadata for *all* reads

```
@HD VN:1.6 SO:coordinate        – BAM header line
@SQ SN:chr1 LN:248956422        – Reference sequence dictionary entries
@SQ SN:chr2 LN:242193529
@RG ID:RG1 SM:SAMPLE_A          – Read group(s)
```

Records containing structured read information (1 line per read/record)

Read name → SlX1:1:127:63:4
Flags → 99
Position → chr1 10052169
MAPQ → 60
CIGAR → 23M3D10M
Mate information → = 14 10
Read sequence → GAAGATACTGGTT
Phred quality scores → 768832'48::::
Metadata → RG:Z:A ...

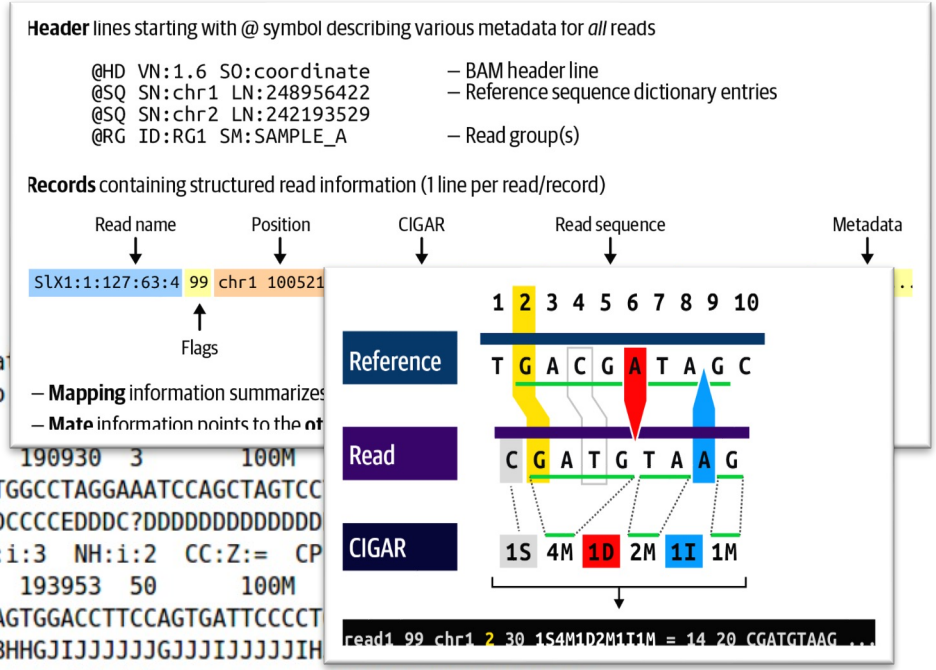
– **Mapping** information summarizes **position, quality, and structure** for each **read**

– **Mate** information points to the **other read in a pair**

https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

Mapping output

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCC
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDD
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCT
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIH
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```



https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

Mapping quality control

Poor mapping

Sample/Reference information
- library of low quality/contaminated
- reference of low quality
- large difference between sample and reference

- paired end reads more likely to be mapped
- repeats in repetitive regions

Technical
- corrupted files
- choice of mapping software
- alignment parameters

Mapping quality control

Poor mapping

Sample/Reference information
- library of low quality/contaminated
- reference of low quality
- large difference between sample and reference

- paired end reads more likely to be mapped
- repeats in repetitive regions

Technical
- corrupted files
- choice of mapping software
- alignment parameters

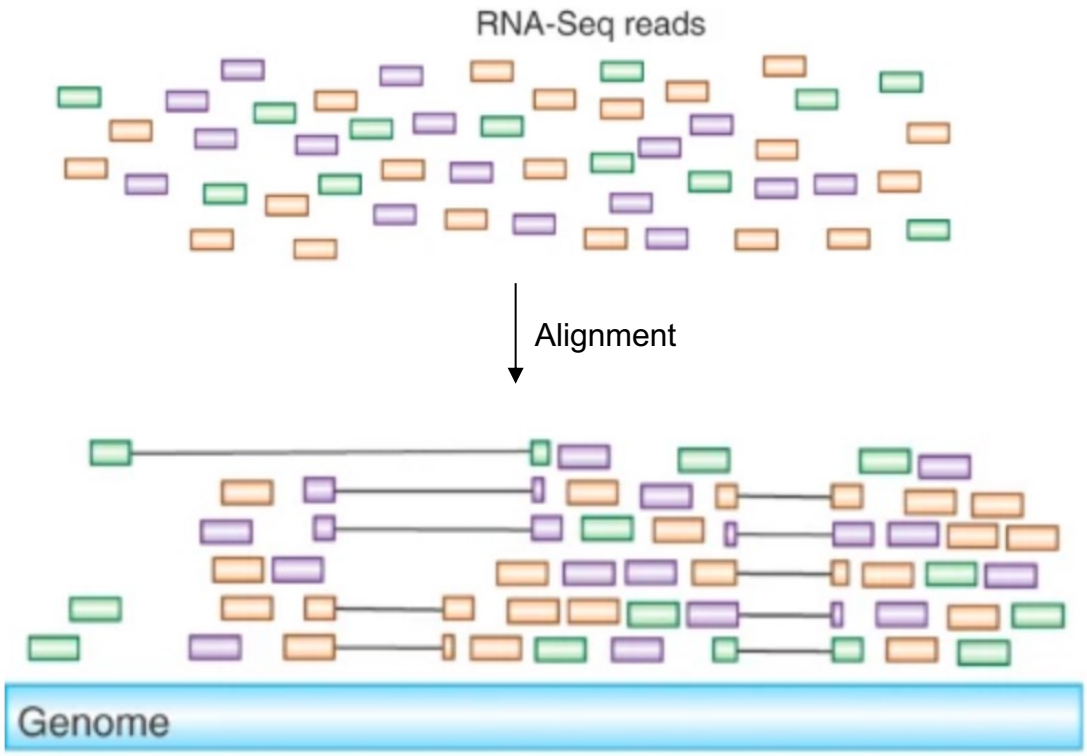
Quality score

- mapP: probability of wrong read alignment
- mapQ = $-10\log_{10}$mapP
  (range 1-255)

⇒ SNV callers usually consider mapping quality scores

⇒ SV callers only use chimeric reads with high mapping quality (e.g. >=30)

Aligned sequences



Adapted from Haas and Zody (2010) Nat. Biotechnology 28:421

Further reading

Elaine Mardis (2007) Nature Milestones. A brief history of (DNA sequencing) time

ENCODE RNA-seq good practice:
http://genome.ucsc.edu/ENCODE/protocols/dataStandards/RNA_standards_v1_2011_May.pdf

FastQC:
https://www.youtube.com/watch?v=bz93ReOv87Y
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/
https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-reads-to-counts/tutorial.html

Alignment:
Engström et al. (2013) Nat. Methods. Systematic evaluation of spliced alignment programs for RNA-seq data
Fonseca et al. (2012) Bioinformatics. Tools for mapping high-throughput sequencing data
Shang et al. (2014) BioMed Res. Int. Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis
Dobin et al. (2013) Bioinformatics. STAR: ultrafast universal RNA-seq aligner
Barruzo et al. (2016) Nature Methods. Simulation-based comprehensive benchmarking of RNA-seq aligners

Narrandes and Xu (2018) J. Cancer. Gene Expression Detection Assay for Cancer Clinical Use

RNA sequencing

- part 1 -

Deutsches Krebsforschungszentrum
Angewandte Bioinformatik (Prof. Dr. Benedikt Brors)
Dr. Óscar González-Velasco (oscar.gonzalezvelasco@dkfz-heidelberg.de)