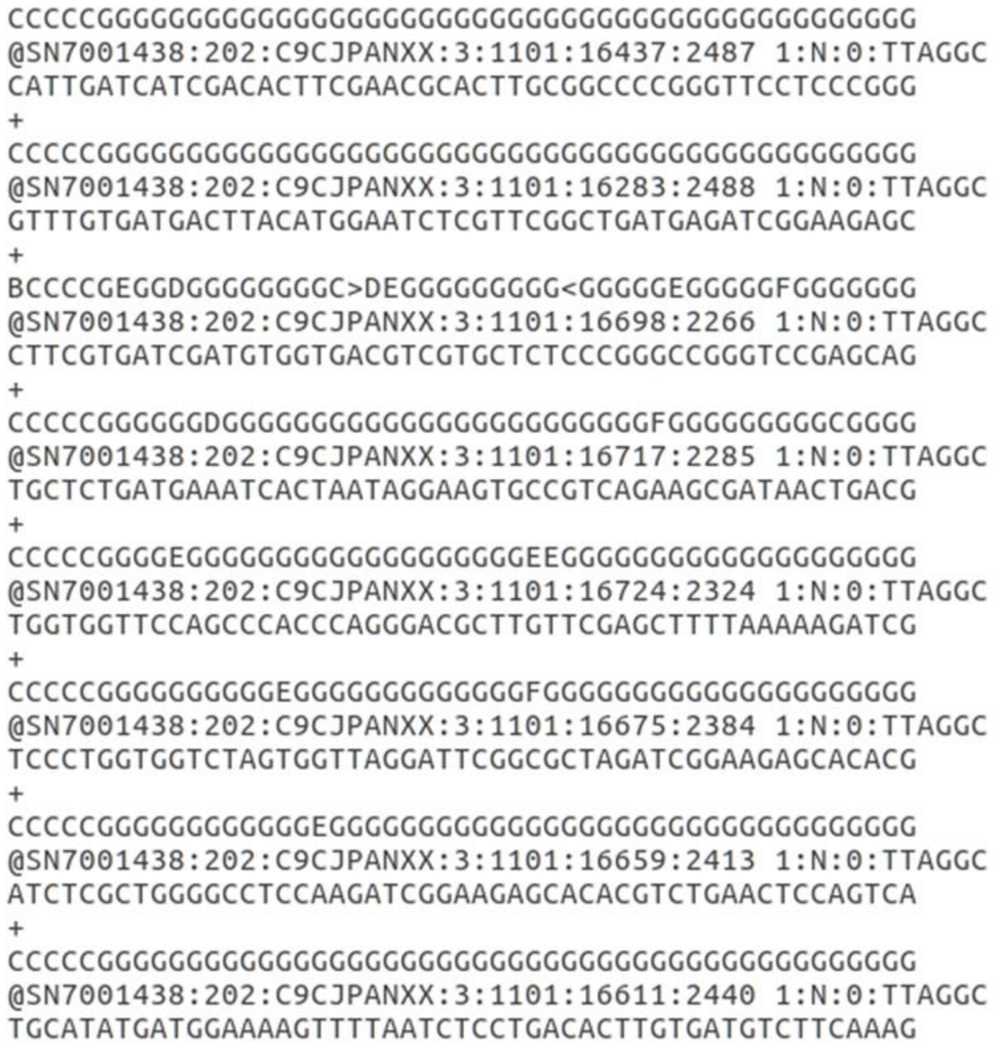
Slides based on Luigi Grassi’s material

RNA-SEQ DATA ANALYSIS:

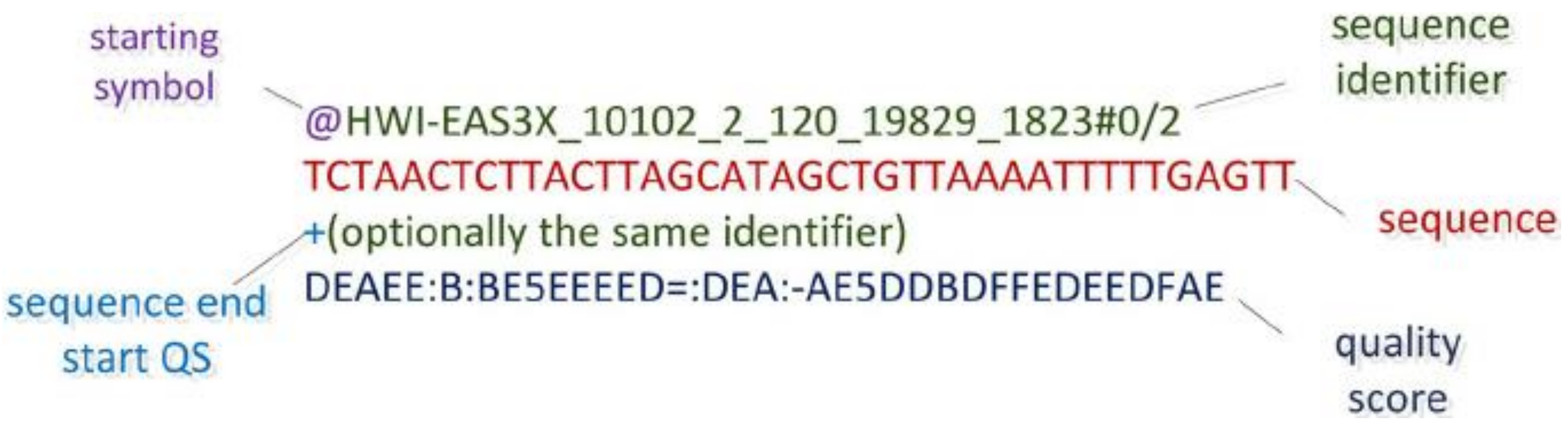
Guillermo Parada Gonzalez
(guillermo.parada@sanger.ac.uk)

Input data: FASTQ



```
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@SN7001438:202:C9CJPANXX:3:1101:16437:2487 1:N:0:TTAGGC
CATTGATCATCGACACTTCGAACGCACTTGCGGCCCCGGGTTCCTCCCGGG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@SN7001438:202:C9CJPANXX:3:1101:16283:2488 1:N:0:TTAGGC
GTTTGTGATGACTTACATGGAATCTCGTTCGGCTGATGAGATCGGAAGAGC
+
BCCCCGEGGDGGGGGGGGGC>DEGGGGGGGG<GGGGGEGGGGGFGGGGGGG
@SN7001438:202:C9CJPANXX:3:1101:16698:2266 1:N:0:TTAGGC
CTTCGTGATCGATGTGGTGACGTCGTGCTCTCCCGGGCCGGGTCCGAGCAG
+
CCCCCGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGCGGGG
@SN7001438:202:C9CJPANXX:3:1101:16717:2285 1:N:0:TTAGGC
TGCTCTGATGAAATCACTAATAGGAAGTGCCGTCAGAAGCGATAACTGACG
+
CCCCCGGGGEGGGGGGGGGGGGGGGGGGGEEGGGGGGGGGGGGGGGGGGGG
@SN7001438:202:C9CJPANXX:3:1101:16724:2324 1:N:0:TTAGGC
TGGTGGTTCCAGCCCACCCAGGGACGCTTGTTCGAGCTTTTAAAAAGATCG
+
CCCCCGGGGGGGGGGEGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGG
@SN7001438:202:C9CJPANXX:3:1101:16675:2384 1:N:0:TTAGGC
TCCCTGGTGGTCTAGTGGTTAGGATTCGGCGCTAGATCGGAAGAGCACACG
+
CCCCCGGGGGGGGGGGGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@SN7001438:202:C9CJPANXX:3:1101:16659:2413 1:N:0:TTAGGC
ATCTCGCTGGGGCCTCCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@SN7001438:202:C9CJPANXX:3:1101:16611:2440 1:N:0:TTAGGC
TGCATATGATGGAAAAGTTTTAATCTCCTGACACTTGTGATGTCTTCAAAG
```

Input data: Fastq



starting symbol

sequence identifier

@HWI-EAS3X_10102_2_120_19829_1823#0/2

TCTAACTCTTACTTAGCATAGCTGTTAAAATTTTTGAGTT

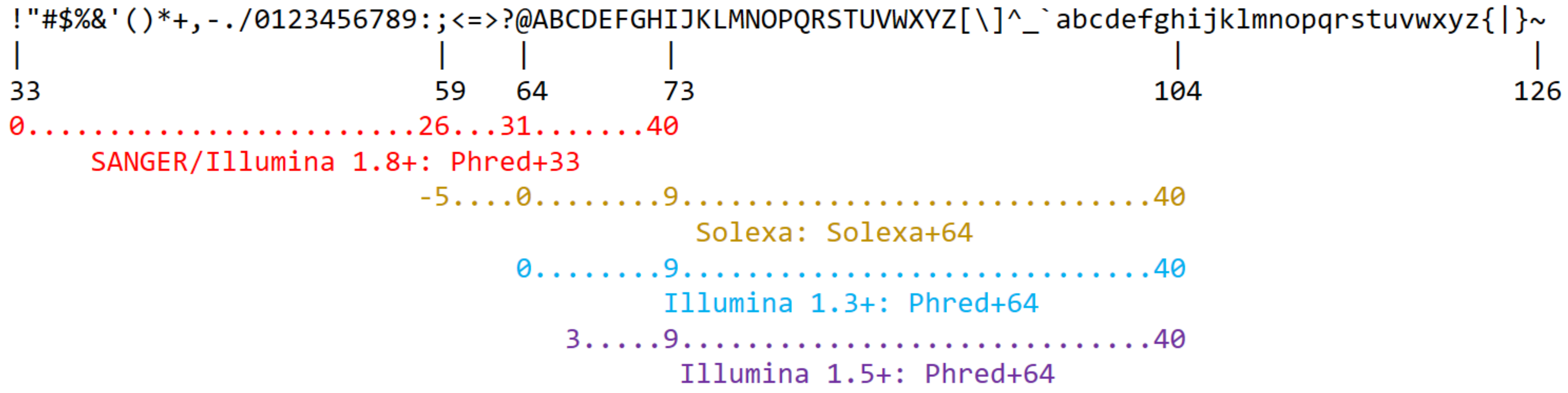
sequence

+(optionally the same identifier)

sequence end start QS

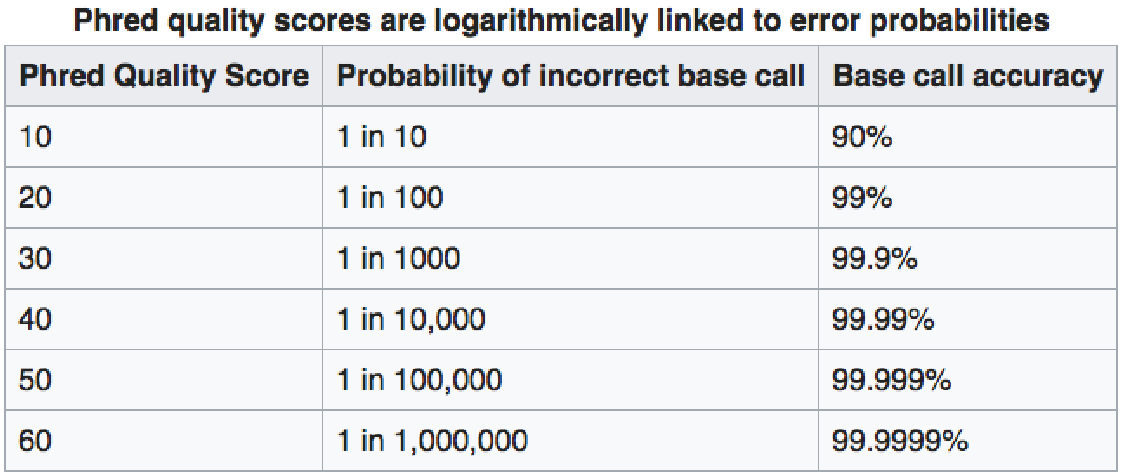
DEAEE:B:BE5EEEED=:DEA:-AE5DDBDFFEDEEDFAE

quality score



```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                  126
0........................26...31.......40
     SANGER/Illumina 1.8+: Phred+33
                         -5....0........9.............................40
                                          Solexa: Solexa+64
                               0........9.............................40
                                        Illumina 1.3+: Phred+64
                                  3.....9.............................40
                                         Illumina 1.5+: Phred+64
```

Input data: Fastq

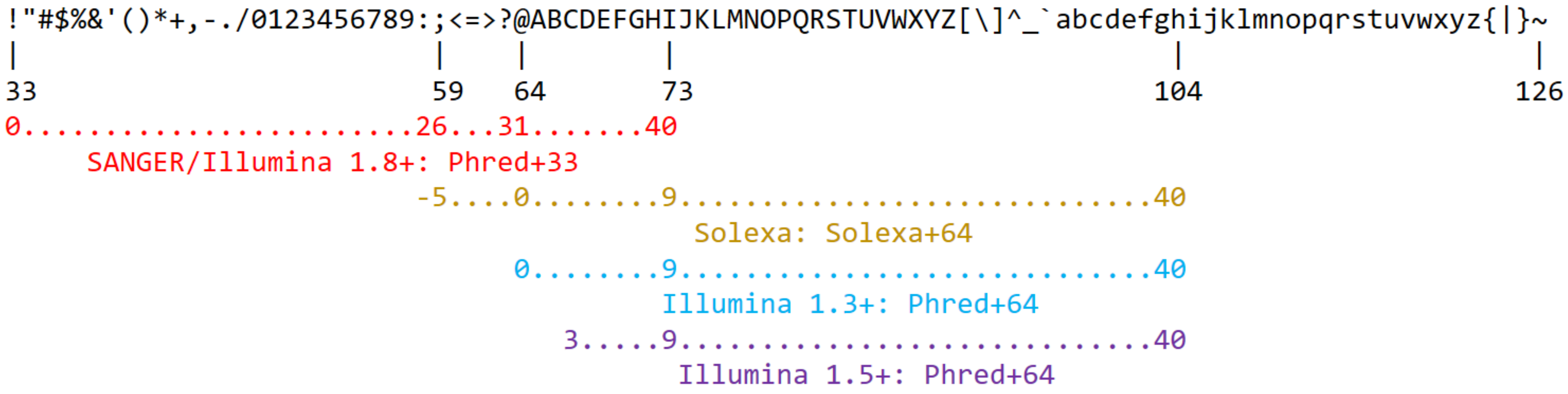


Phred quality scores are logarithmically linked to error probabilities

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%

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

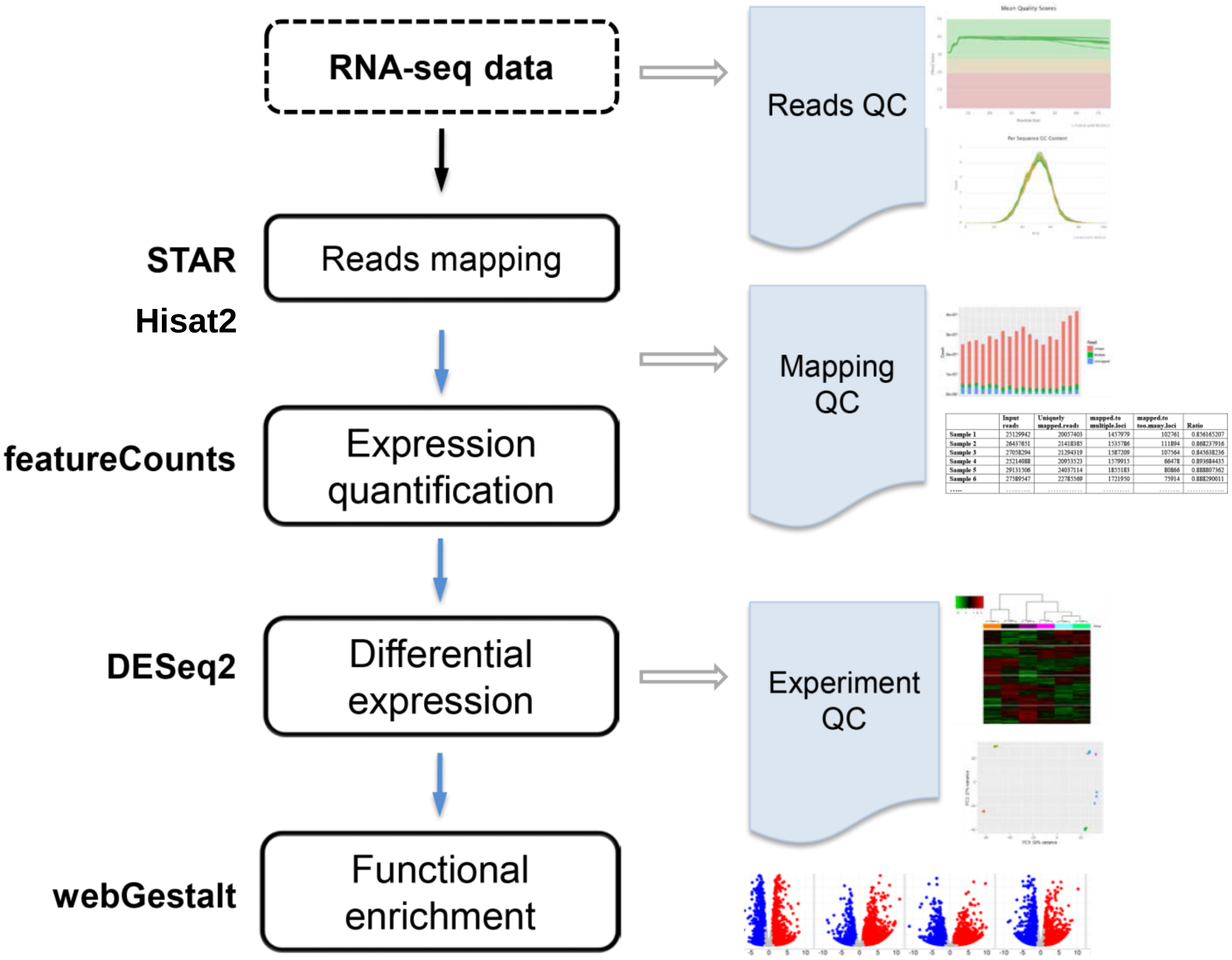
$$P = 10^{\frac{-Q}{10}}$$



```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
     SANGER/Illumina 1.8+: Phred+33
                         -5....0........9.............................40
                                          Solexa: Solexa+64
                               0........9.............................40
                                        Illumina 1.3+: Phred+64
                                  3.....9.............................40
                                         Illumina 1.5+: Phred+64
```

Step 3 – A typical RNA-seq analysis workflow*

* if the reference genome is available



RNA-seq data

STAR

Hisat2

Reads mapping

featureCounts

Expression quantification

DESeq2

Differential expression

webGestalt

Functional enrichment

Reads QC

Mapping QC

Experiment QC

	Input reads	Uniquely mapped.reads	mapped.to multiple.loci	mapped.to too.many.loci	Ratio
Sample 1	25129942	20057403	1457979	102761	0.856165207
Sample 2	26437651	21418385	1535786	111894	0.868237916
Sample 3	27058294	21294319	1587209	107564	0.845638236
Sample 4	25214088	20953523	1579915	66478	0.893684435
Sample 5	29131506	24037114	1855183	80866	0.888807362
Sample 6	27589547	22785569	1721950	75914	0.888290011
.....	....	....	....	....	....

(VANGARD)

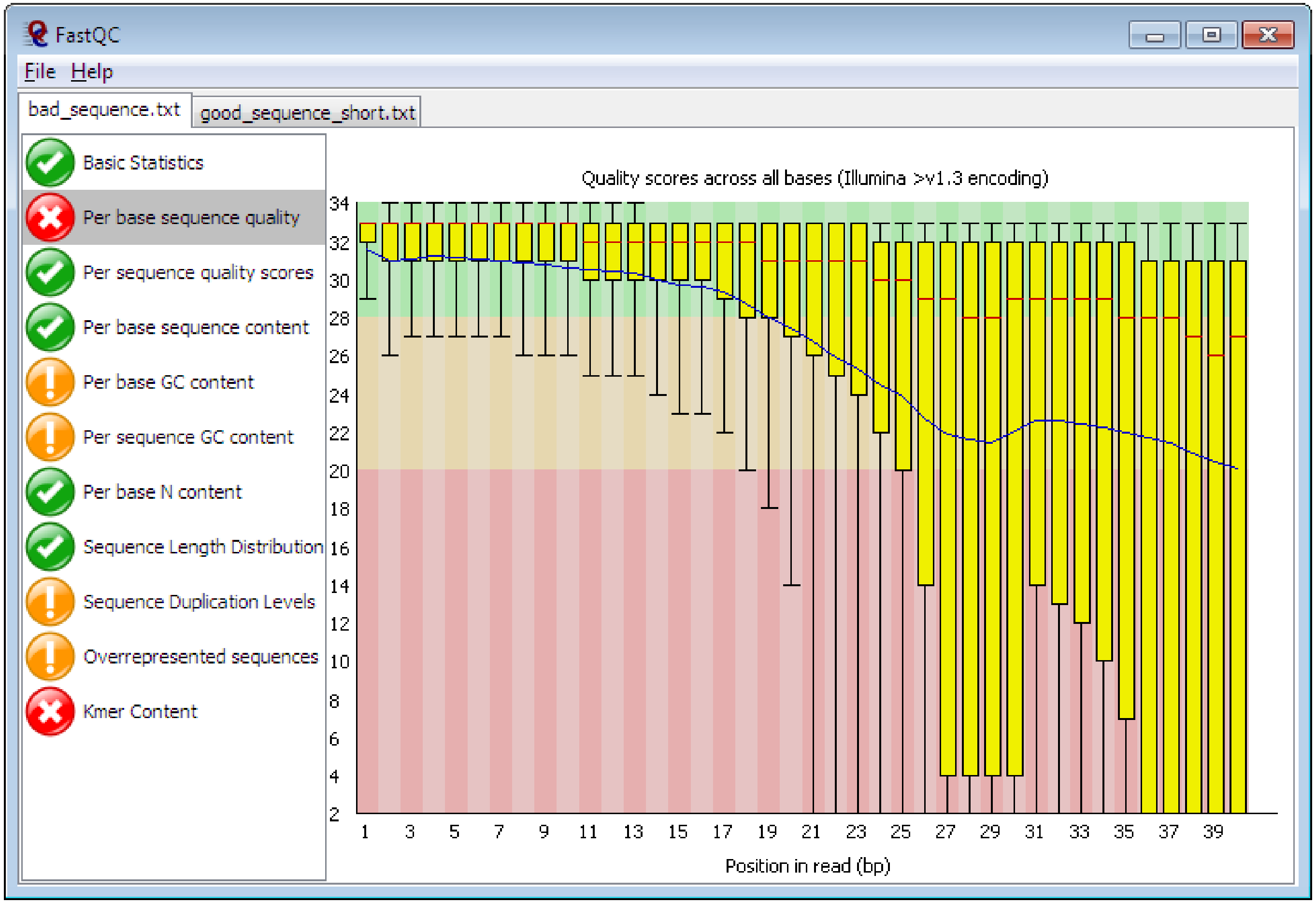
Quality Control

- Essential for downstream analysis.

- Decide sensibly on which data can be filtered out from the downstream analysis.

- You might find yourself going back to that step several times during downstream analysis.

FastQC



FastQC

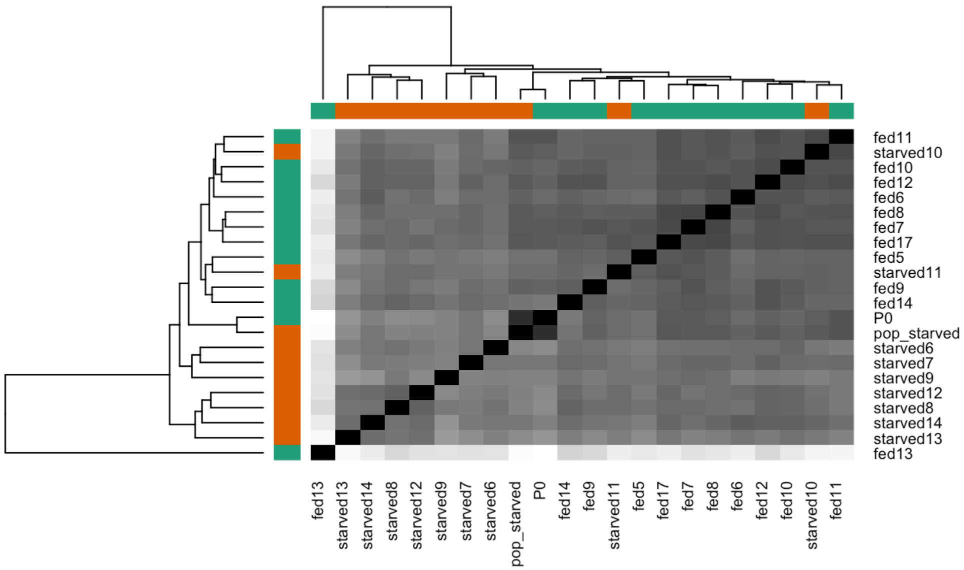
File Help

bad_sequence.txt | good_sequence_short.txt

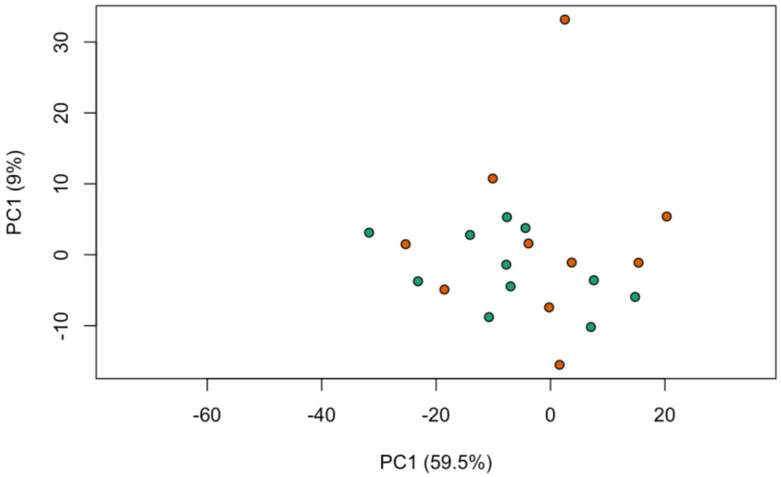
- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Kmer Content

Filtering outliers

Sample Distance Matrix

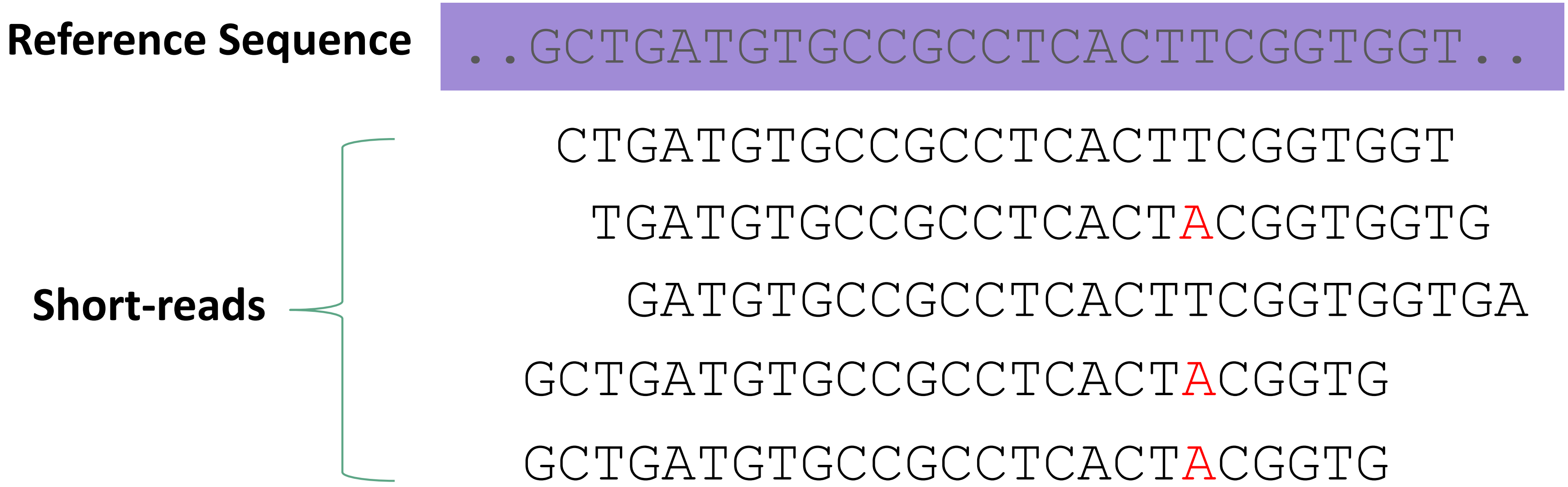


PCA Biplot



Alignment

AIM: Given a reference sequence and a set of short reads, align each read to the reference sequence

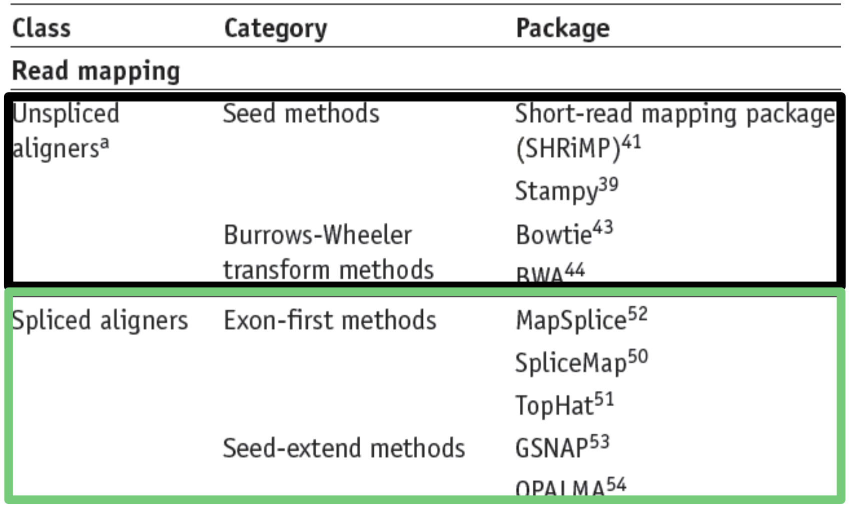


Reference Sequence

Short-reads

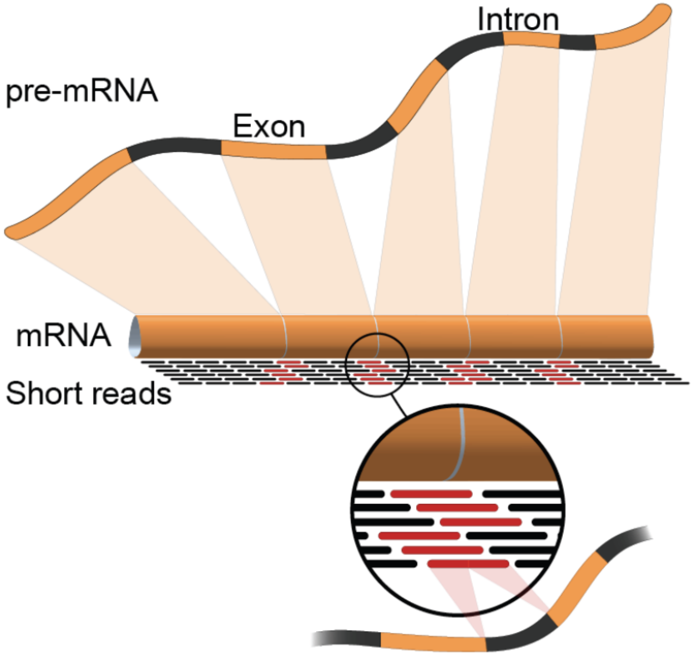
```
..GCTGATGTGCCGCCTCACTTCGGTGGT..
   CTGATGTGCCGCCTCACTTCGGTGGT
    TGATGTGCCGCCTCACTACGGTGGTG
     GATGTGCCGCCTCACTTCGGTGGTGA
  GCTGATGTGCCGCCTCACTACGGTG
  GCTGATGTGCCGCCTCACTACGGTG
```

Alignment



Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		OPALMA[54]

STAR
Hisat2

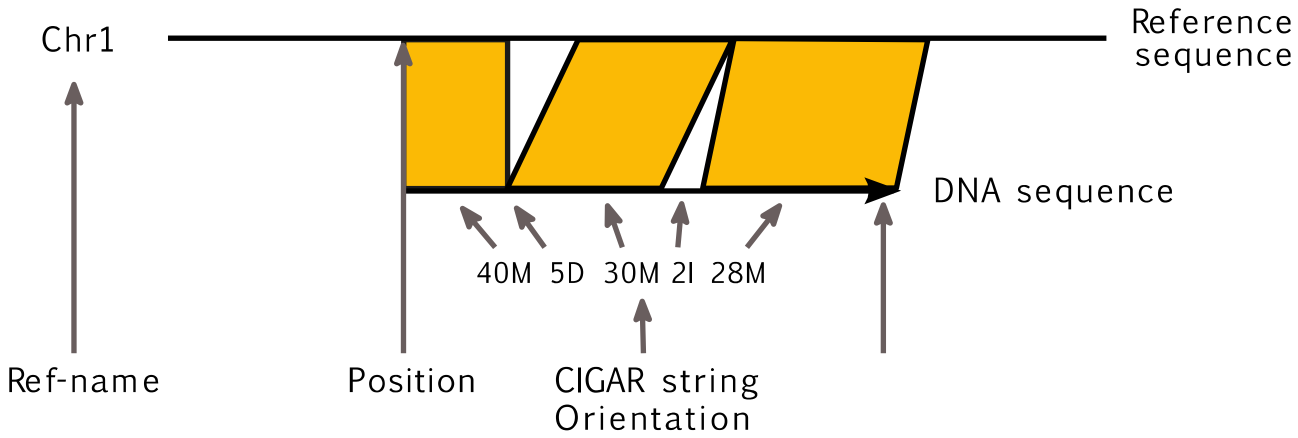


Garber, M., et al. (2011), Nature Methods, 8(6), 469–477.

Aliments are reported as SAM

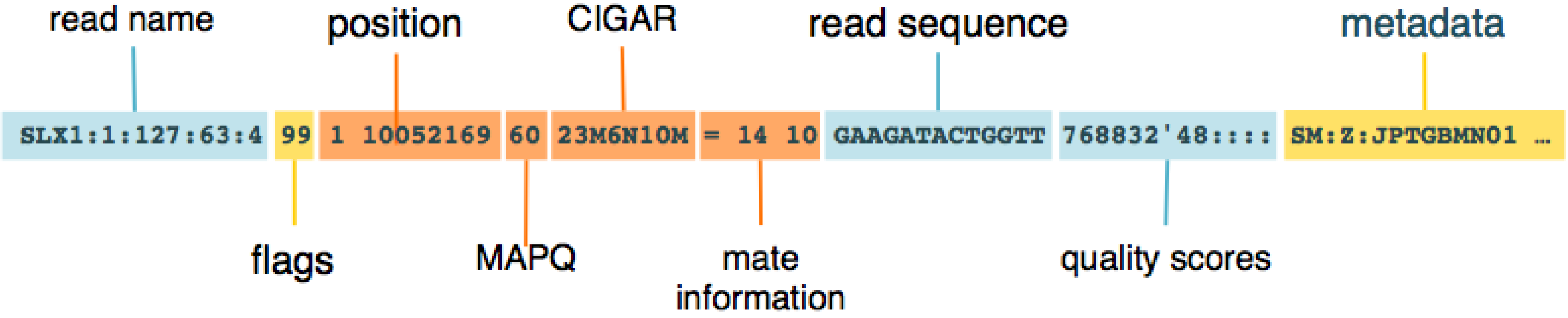
SAM (Sequence Alignment/Map) format

- Unified format for storing read alignments to a reference genome
- Developed by the 1000 Genomes Project group (2009)
- One record (a single DNA fragment alignment) per line describing alignment between fragment and reference
- 11 fixed columns + optional key:type:value tuples



SAM FORMAT

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



read name

position

CIGAR

read sequence

metadata

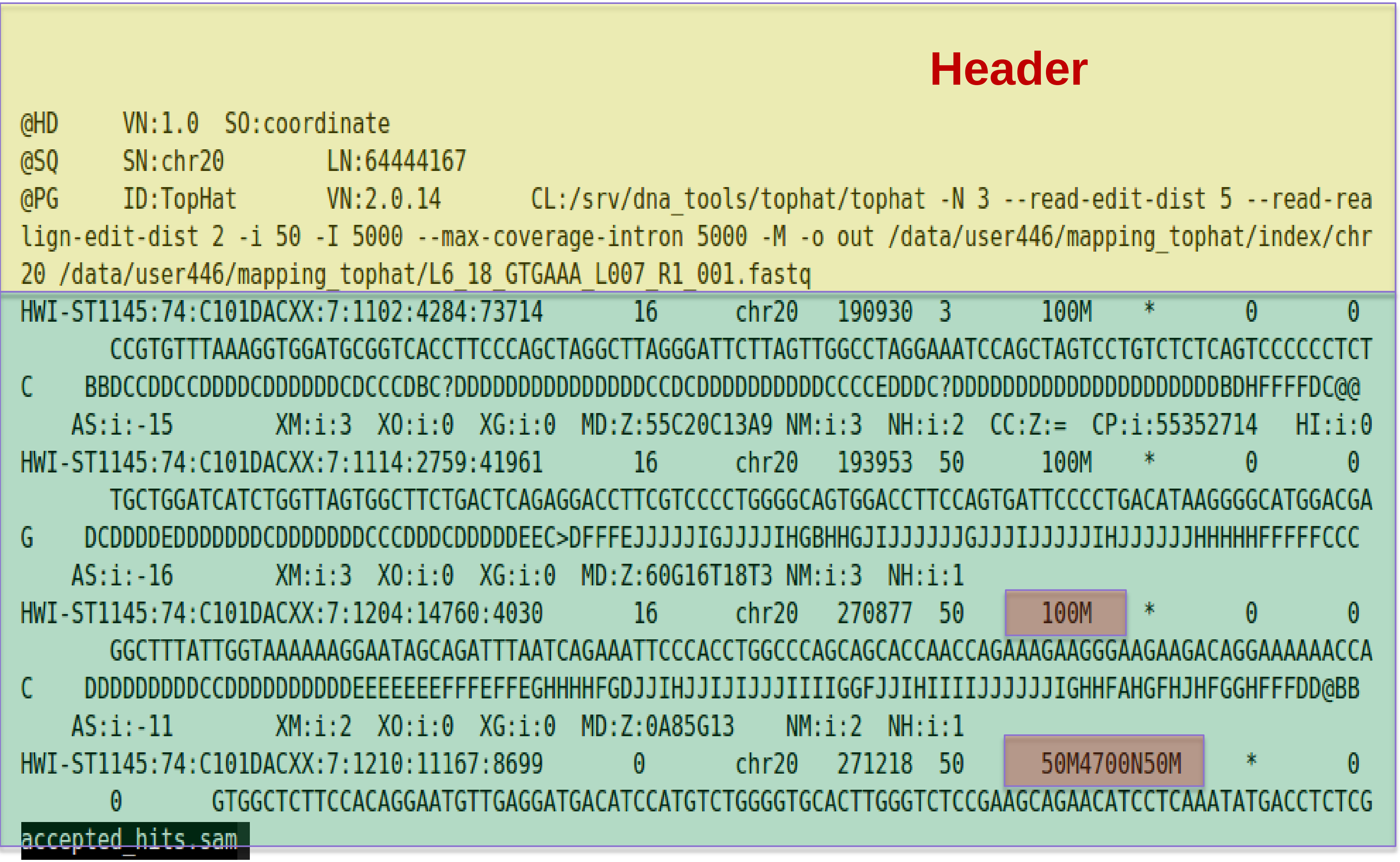
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …

flags

MAPQ

mate information

quality scores



Header

```
@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?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?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
accepted_hits.sam
```

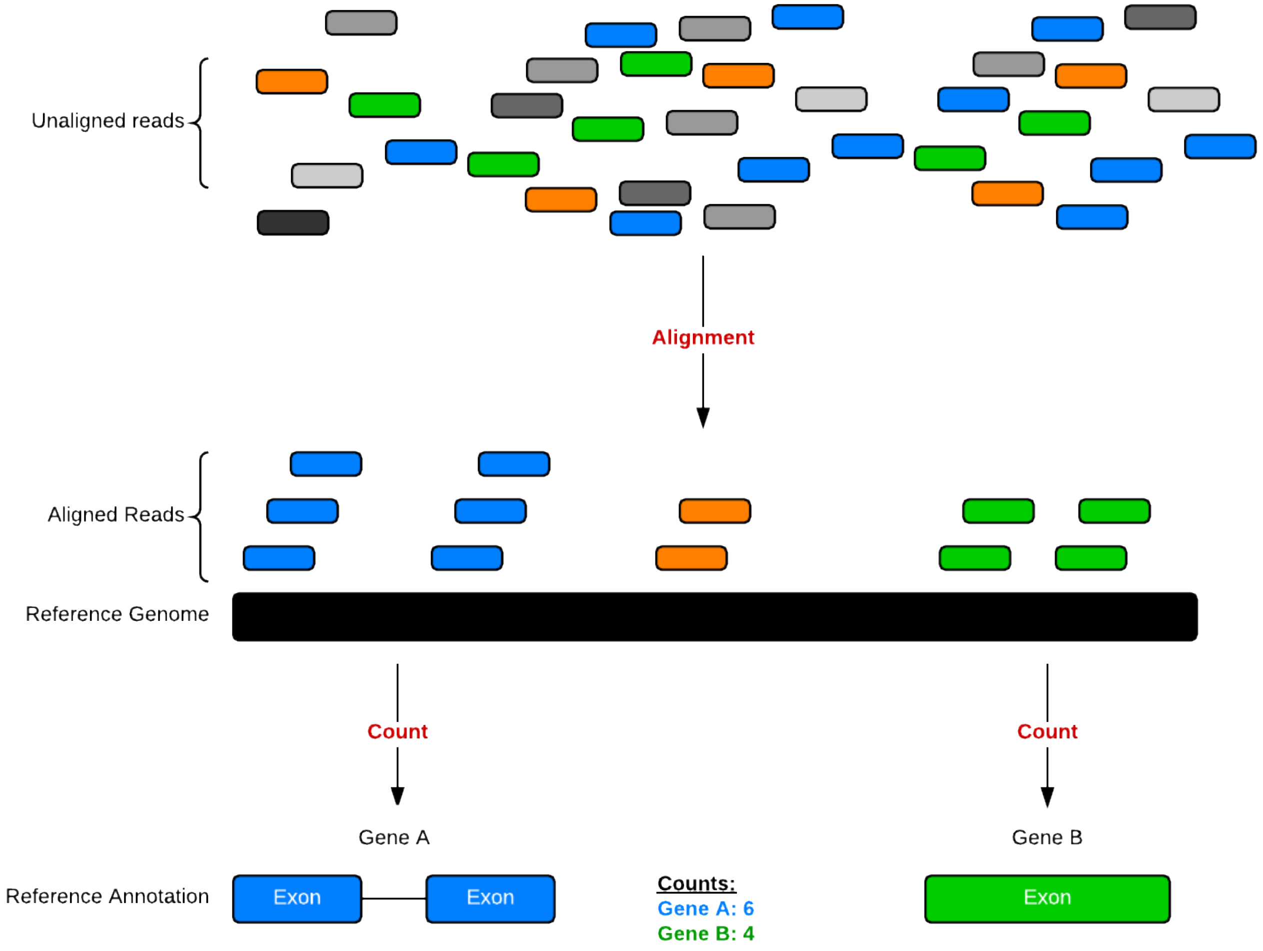
Aligments

SAM tools

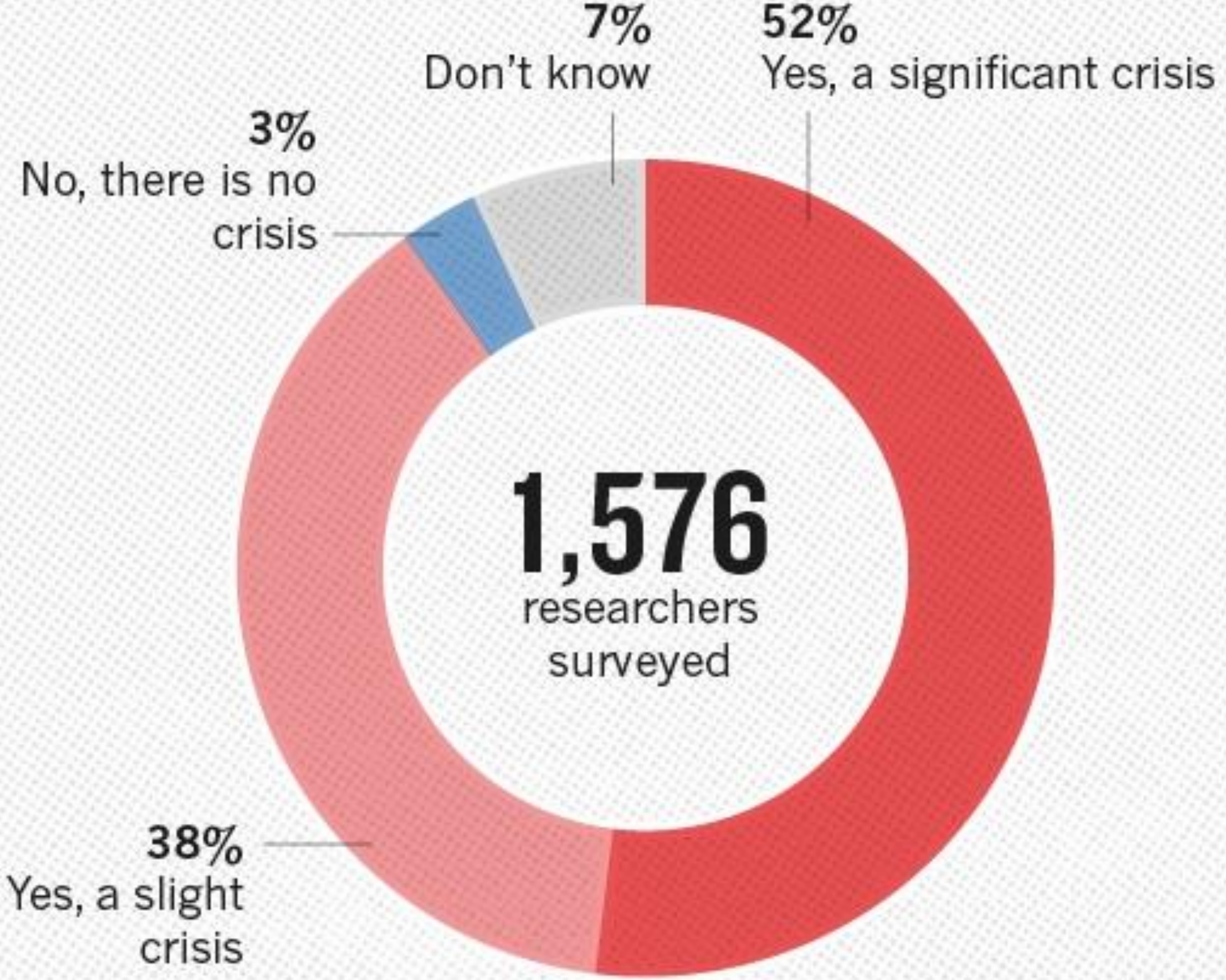
MANUAL: http://www.htslib.org/doc/samtools-1.1.html

Utility	Description
view	Convert between sam/bam format, and filter alignment file
sort	Sort alignments by genomic position
index	Creates a new index file that allows fast look up, generating *.sam.sai or *.bam.bai files. These files are required by some genome browsers
mpileup	Creates pileup format, i.e. BCF files, which gives overlapping read bases or indels for each genomic position. Can be used for variant calling
flagstat	Summary alignment statistics
merge	Merge multiple bam files into one bam aligment file. For example, if you have one bam file for each tile, combine all into one bam file for the sample
rmdup	remove potential PCR duplicates
bam2fq	convert bam to FASTQ format

Read count

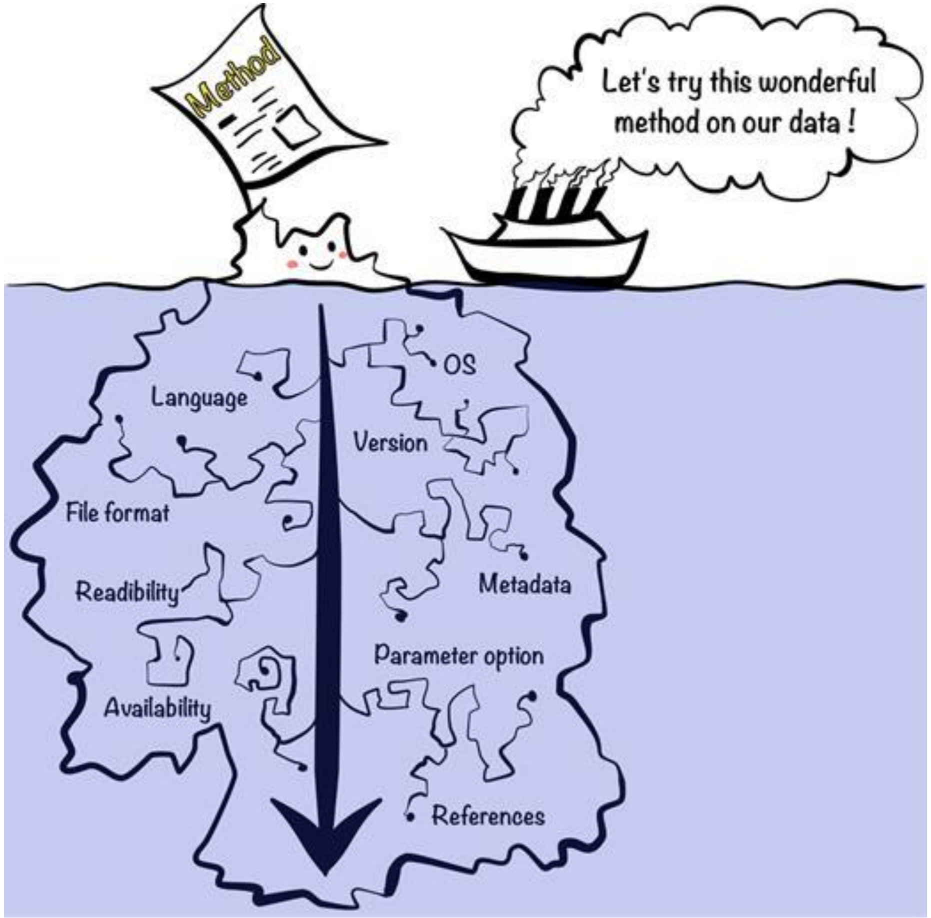

Unaligned reads
Alignment
Aligned Reads
Reference Genome
Count
Count
Gene A
Gene B
Reference Annotation
Exon
Exon
Counts:
Gene A: 6
Gene B: 4
Exon

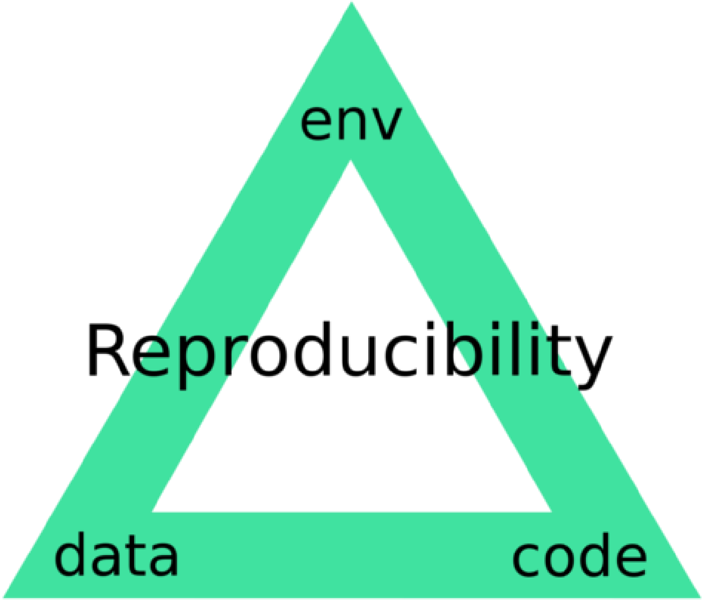
IS THERE A REPRODUCIBILITY CRISIS?



©nature

Reproducibility





Kim et al (2018) Experimenting with reproducibility: a case study of robustness in bioinformatic