Pre-processing and quality control of sequence data

Barbera van Schaik

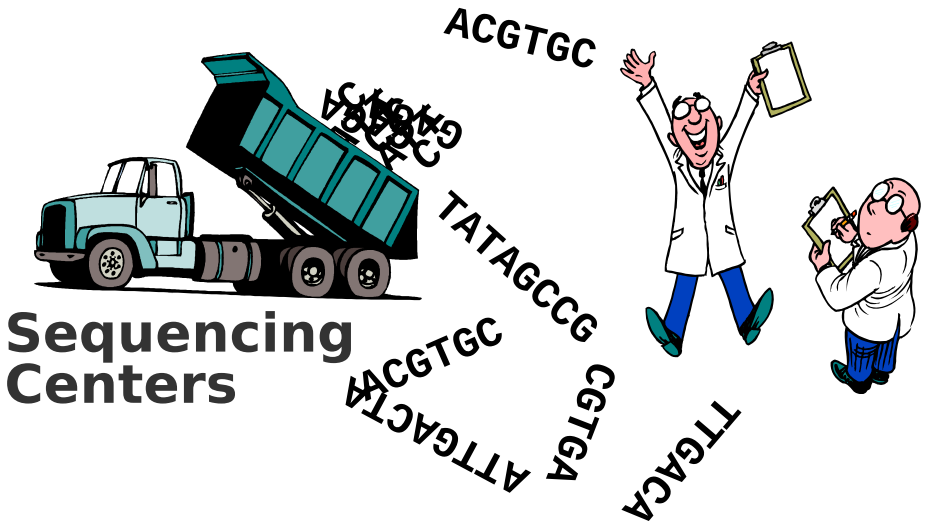
EDS - Bioinformatics and Biomedical Computing

b.d.vanschaik@amsterdamumc.nl

Topic: quality control and prepare data for the interesting stuff

Keep

Throw away



Pre-processing **Quality control**

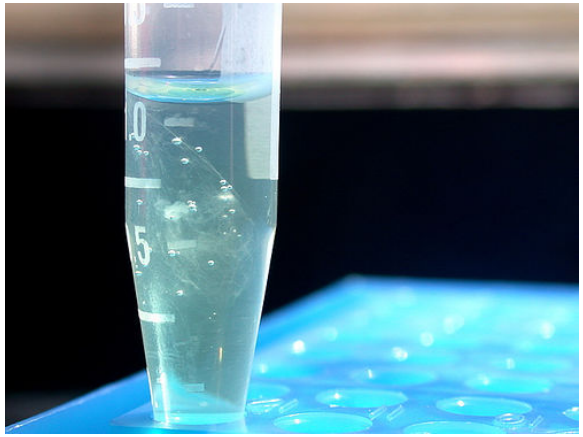
Search for mutations

Expression profiling

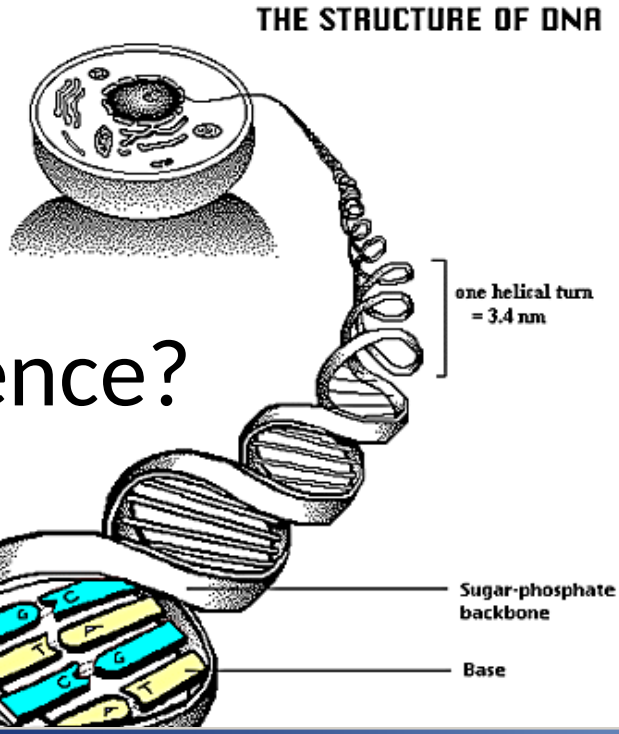
Analysis Z

Pre-processing and quality control of sequence data

- In what format is sequence data provided?
- Is the data of good quality?
- How do you pre-process raw data for further analysis?
- How to align sequences to a reference genome?
- How do you judge if the alignment went well?

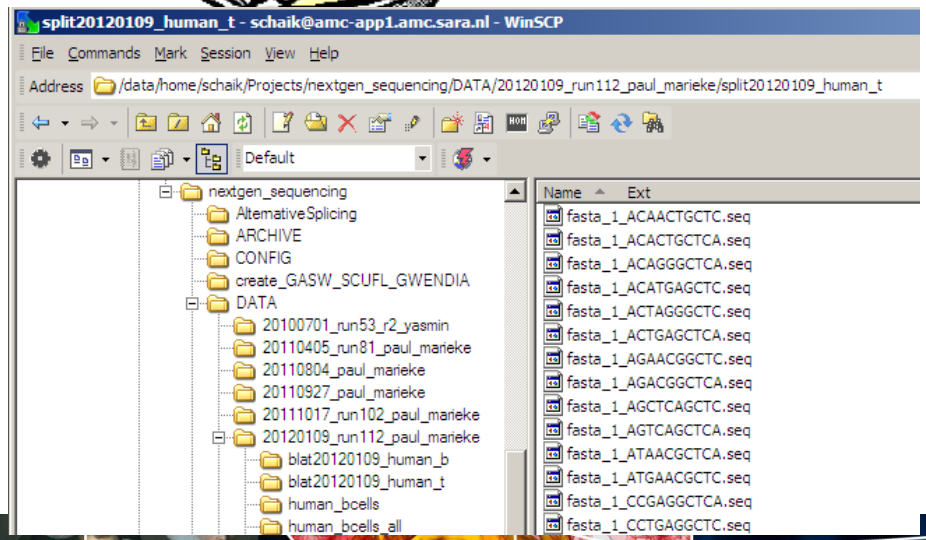


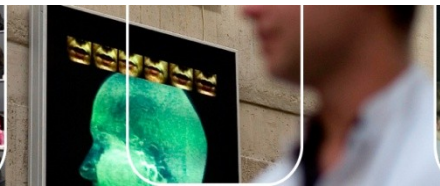
DNA sequence?



>hg19_dna range=chr16:226703-227520 5'pad=0 3'pad=0 strand=+ repeatMasking=none

GAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGG
CCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCC
CTGGAGAGGTGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCCTCGCCCGC
CCGGACCCACAGGCCACCCTCAACCGTCCTGGCCCCGGACCCAAACCCCA
CCCCTCACTCTGCTTCTCCCCGCAGGATGTTCCTGTCCTTCCCCACCACC
AAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAA
GGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACG
TGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCAC
AAGCTTCGGGTGGACCCGGTCAACTTCAAGGTGAGCGGCGGGCCGGGAGC
GATCTGGGTCGAGGGGCGAGATGGCGCCTTCCTCGCAGGGCAGAGGATCA
CGCGGGTTGCGGGAGGTGTAGCGCAGGCGGCGGCTGCGGGCCTGGGCCCT
CGGCCCCACTGACCCTCTTCTCTGCACAGCTCCTAAGCCACTGCCTGCTG
GTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGC
CTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAAT
ACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCC
CCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAAT

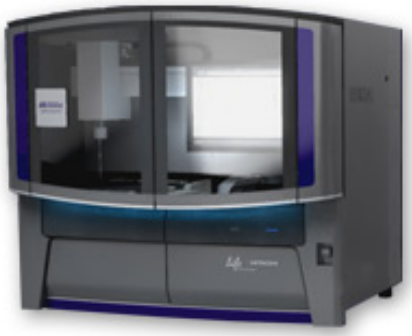


Selection of brands

Roche - 454

Illumina - HiSeq



Applied biosystems - Solid

Information from the sequencer

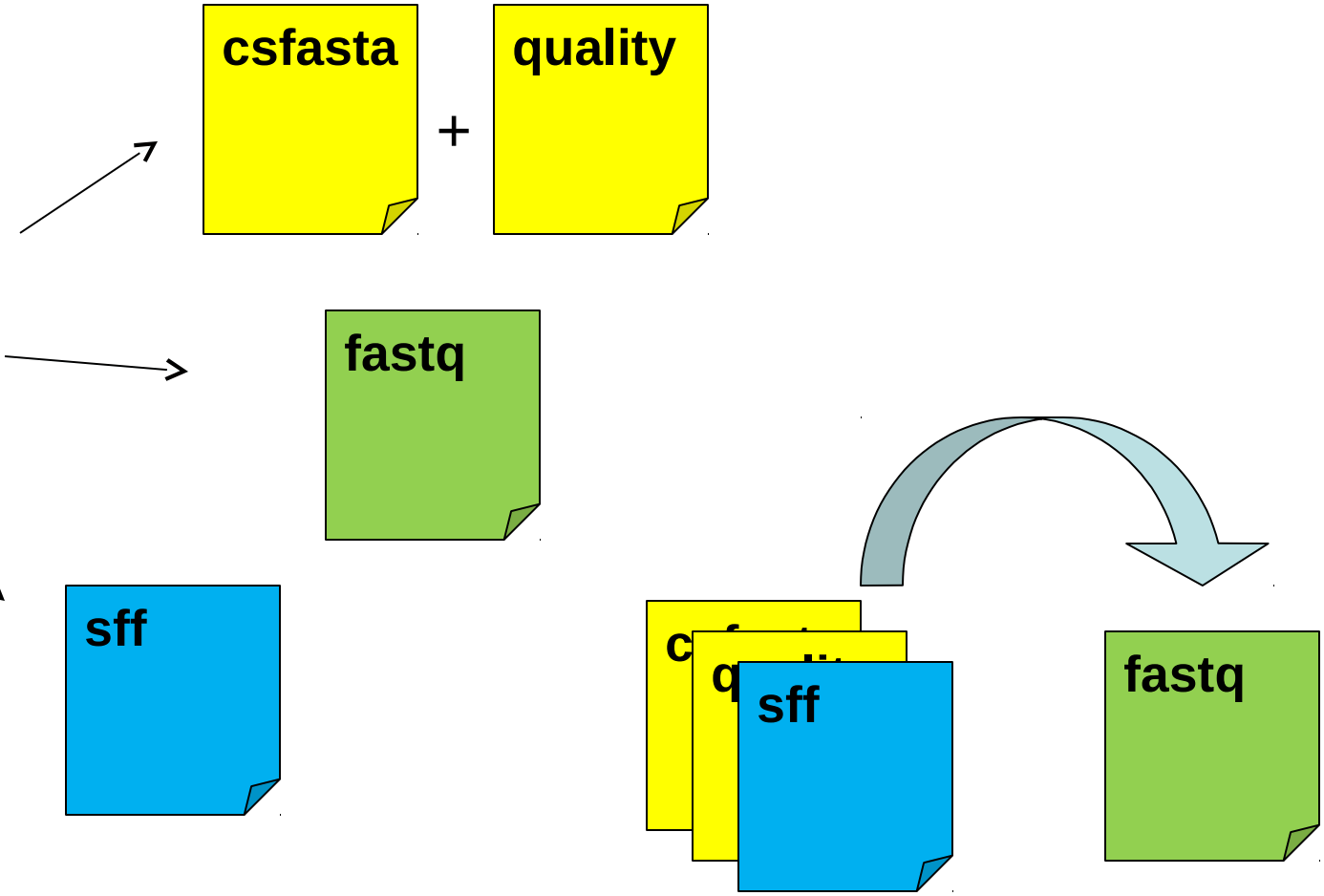
F49EYDB01CB123 length=67 xy=0840_1277 region=1 run=R_2009_11_0

ID Sequence length Coordinate on slide

- Name of the sequences
- Sequences AGAGTAGCGTCGTCGTCGACGACGTCGTCGTCC
- Quality scores 34 34 34 34 34 34 34 34 31 30 31 31 34 32 30 31 31
- Optional: intensities A,0.87 C,0.00 G,1.20 T,0.30 A,0.54 C,0.50 G,

Every brand provides the data in a different format

- ABI solid
- Illumina
- Roche



Standard format: fastq

Fastq entry for one sequence

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

One fastq file contains
millions or billions of sequences

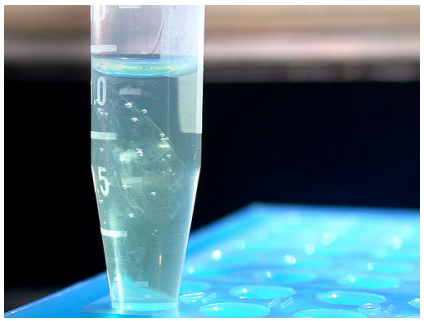
fastq

Image from: http://humanwarnings.hubpages.com/hub/Art-Has-No-Standards

Different types of sequences

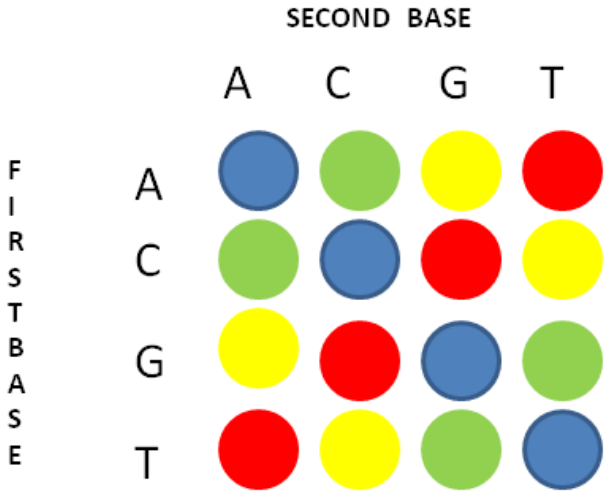
- base space
- color space (Solid)
- double encoded

```
A C G T T A G
T 3 1 3 1 0 3 2
T T C T C A T G
```



Most tools were/are designed for base space

Convert base space into color space



	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

AA	0
CA	1
GA	2
TA	3
AC	1
CC	0
GC	3
TC	2
AG	2

Convert this sequence into color space:

A C T C T T C T G G

A 1 2 2 2 0 2 2 1 0

A C G G G A G G C A

To double encoded

0 = A

1 = C

2 = G

3 = T

Now convert color space to double encoded sequence

Different reporting of quality values

- Phred scores (Sanger)
- Solexa scores (older version of Illumina)
- Illumina 1.3+
- Illumina 1.8+
- Solid (comparable with Sanger)

http://en.wikipedia.org/wiki/Fastq

Phred score

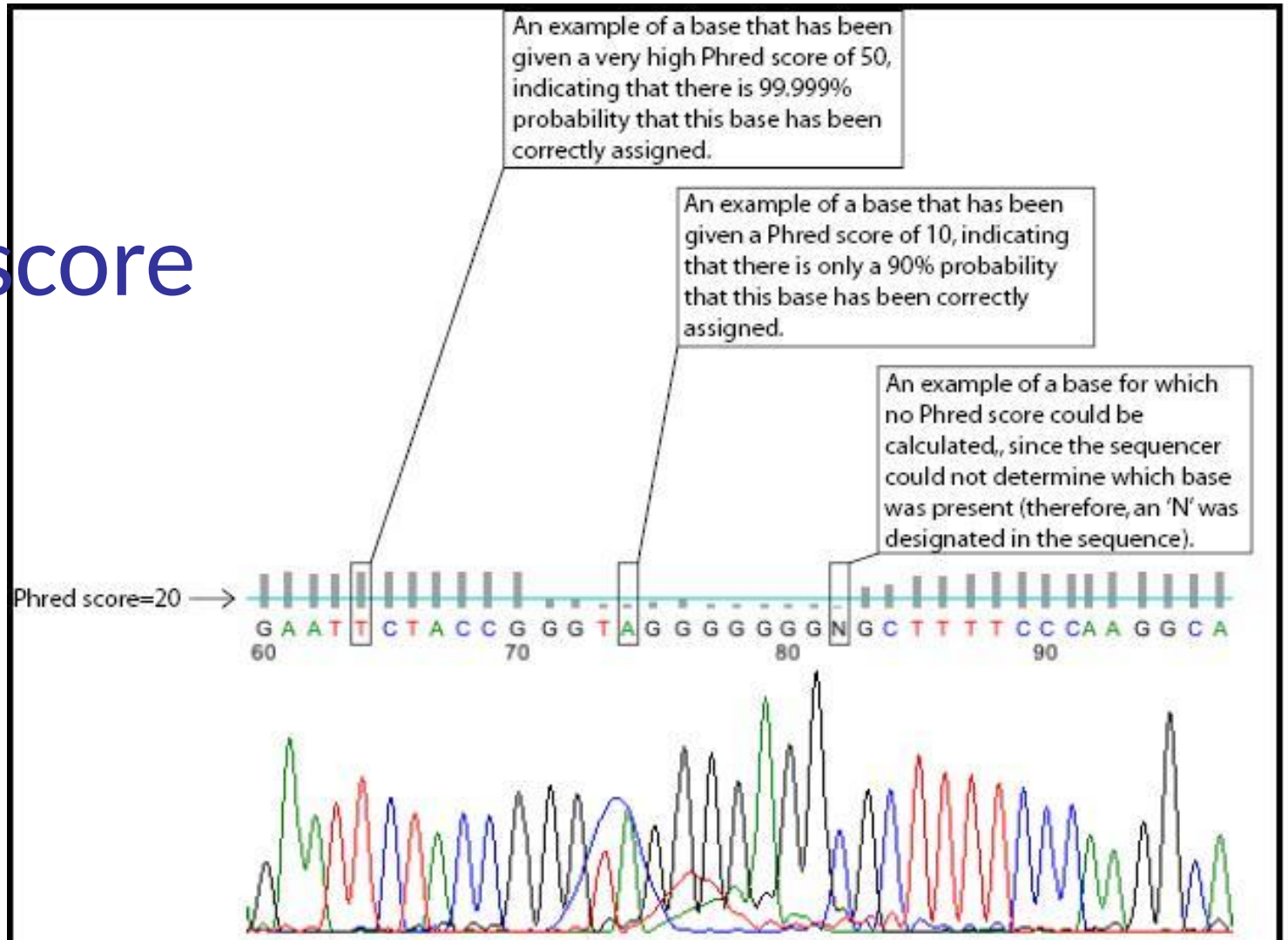


Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

Phred scores

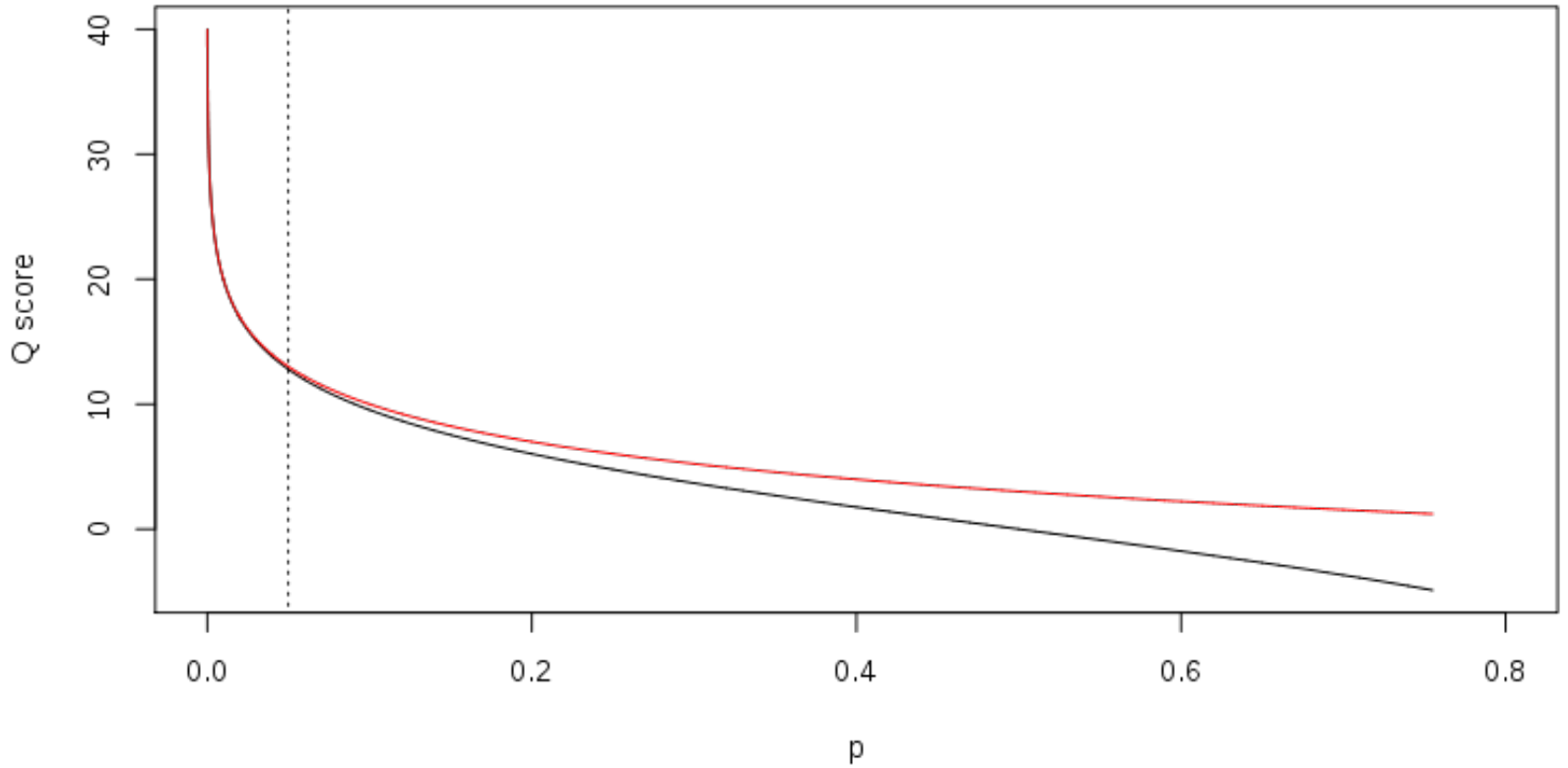
$$Q_{\text{sanger}} = -10 \log_{10} p$$

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 10000	99.99 %
50	1 in 100000	99.999 %

Solexa scores

$$Q_{\text{solexa-prior to v.1.3}} = -10 \log_{10} \frac{p}{1-p}$$



Probability of base call errors. Relationship between Q and p using the **Sanger (red)** and **Solexa (black)** equations. The vertical dotted line indicates p = 0.05, or equivalently, Q ≈ 13.

Ascii encoding

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Fastq entry

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
   |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Exercise: quality scores

A base has an ascii score: **h**

(Illumina 1.3+)

1. To which ascii number does this correspond?
2. What is the Phred score?
3. What does this mean in terms of error rate?

Exercise: quality scores

A base has an ascii score: **h**

(Illumina 1.3+)

1. To which ascii number does this correspond?
2. What is the Phred score?
3. What does this mean in terms of error rate?

1. h = character 104
2. 104 (ascii) – 64 (illumina 1.3 scheme) = 40 (phred score)
3. Phred score 40 = 1 error in 10,000 (accuracy 99.99%)

Summary:
In what format is sequence data provided?

- **Information**
  - sequences and quality scores
- **File formats**
  - sff, csfasta+quality, fastq
- **Sequence formats**
  - base space, color space, double encoded
- **Quality score format**
  - ascii, but with different schemes

Pre-processing and quality control of sequence data

- In what format is sequence data provided?
- **Is the data of good quality?**
- How do you pre-process raw data for further analysis?
- How to align sequences to a reference genome?
- How do you judge if the alignment went well?

Is the raw data of good quality?

- What is a good quality experiment?
- Example of a program for quality control
- What do the metrics mean?

Number of sequences

Basic Statistics

Measure	Value
Filename	run045-20110814.fastq
File type	Conventional base calls
Total Sequences	65905535
Sequence length	34
%GC	53

Back to summary

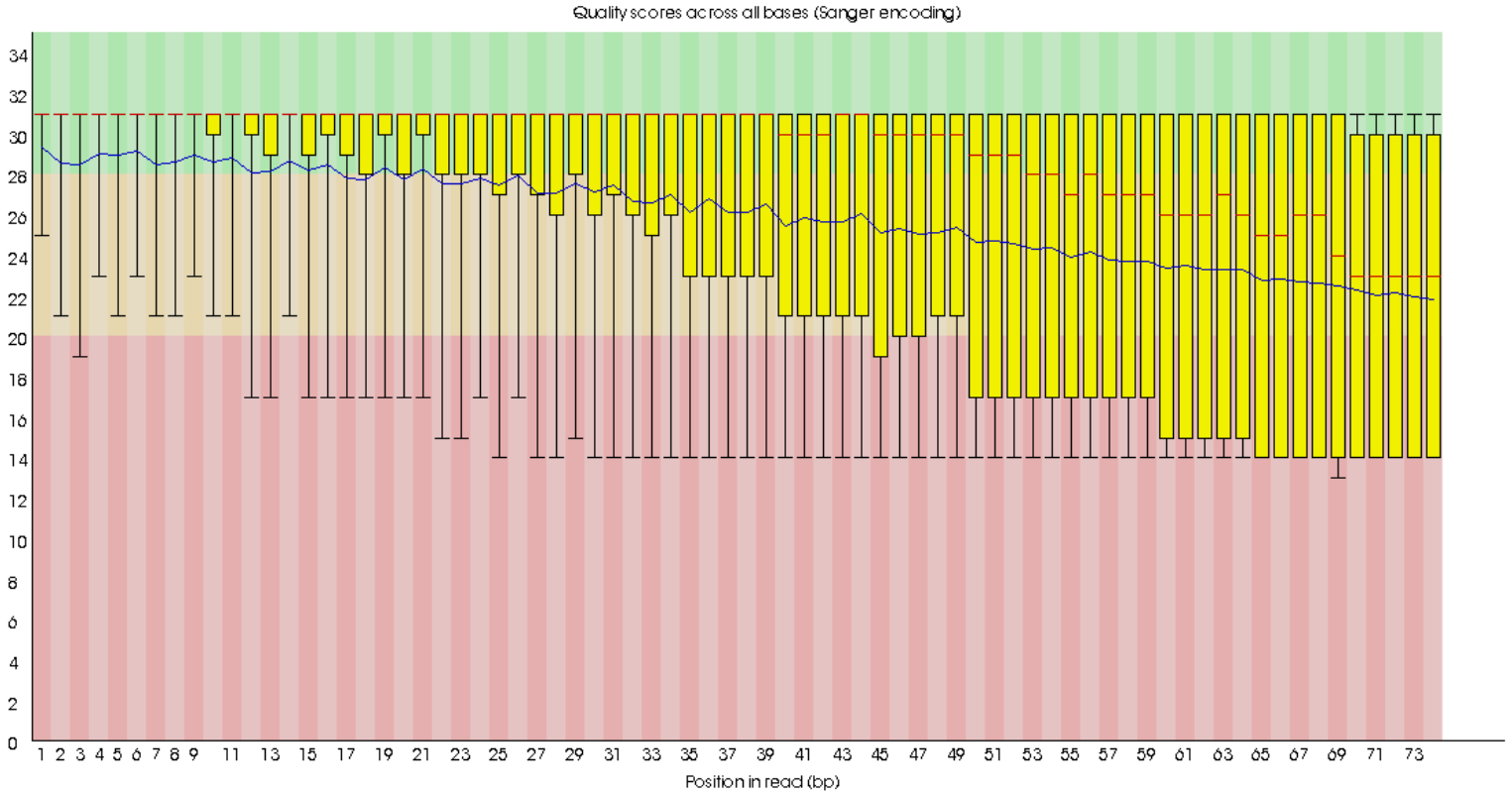
FastQC to check the quality

http://www.youtube.com/watch?v=bz93ReOv87Y

0:00-4:00 Quality of the sequence reads

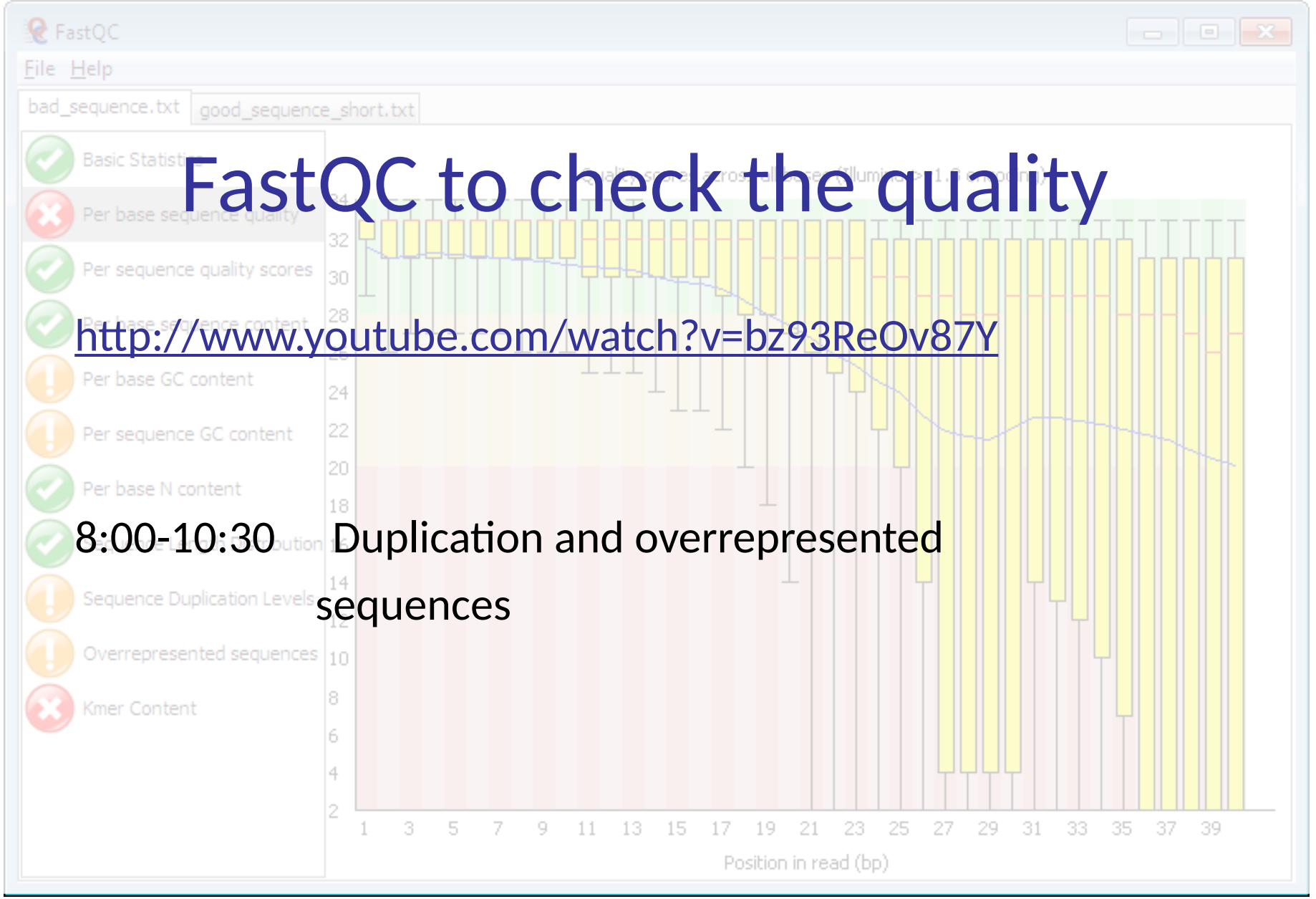
Quality of the sequences

Per base sequence quality



Back to summary

FastQC to check the quality

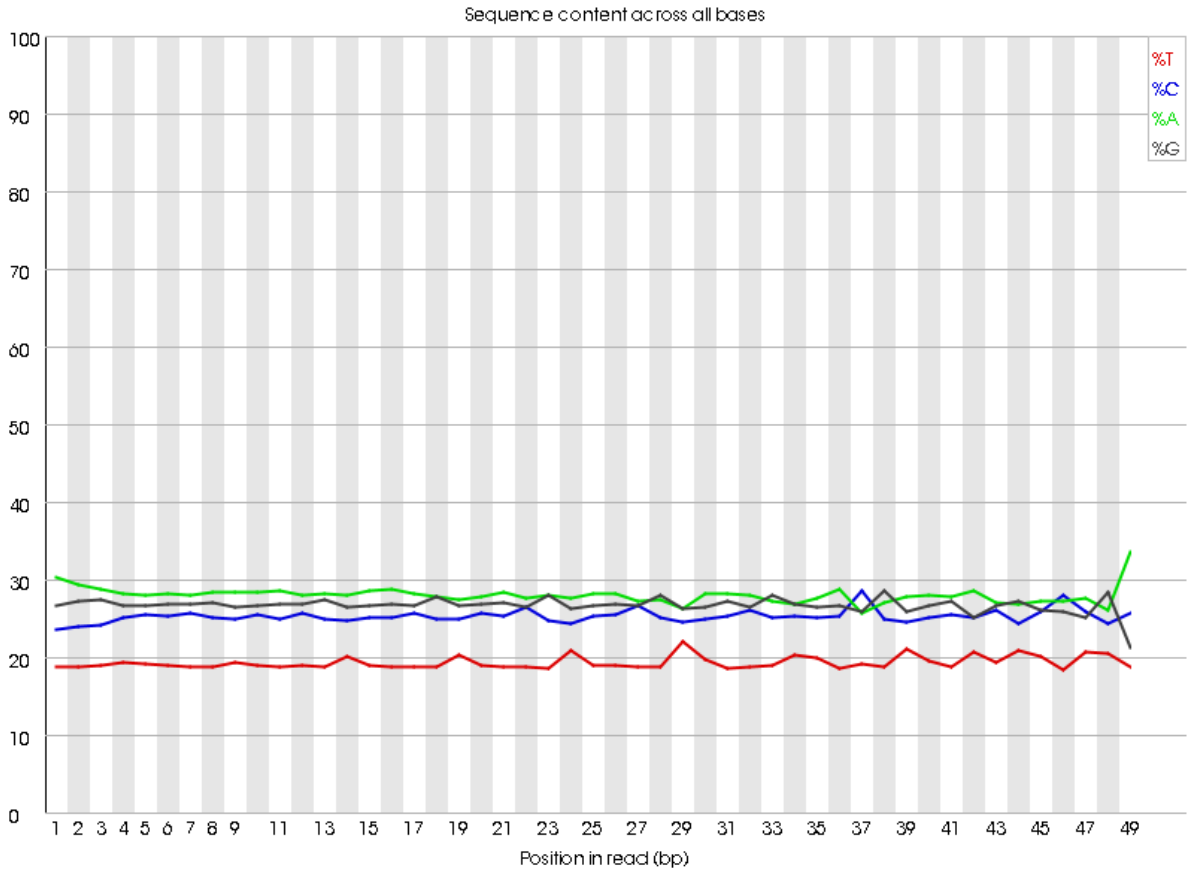


http://www.youtube.com/watch?v=bz93ReOv87Y

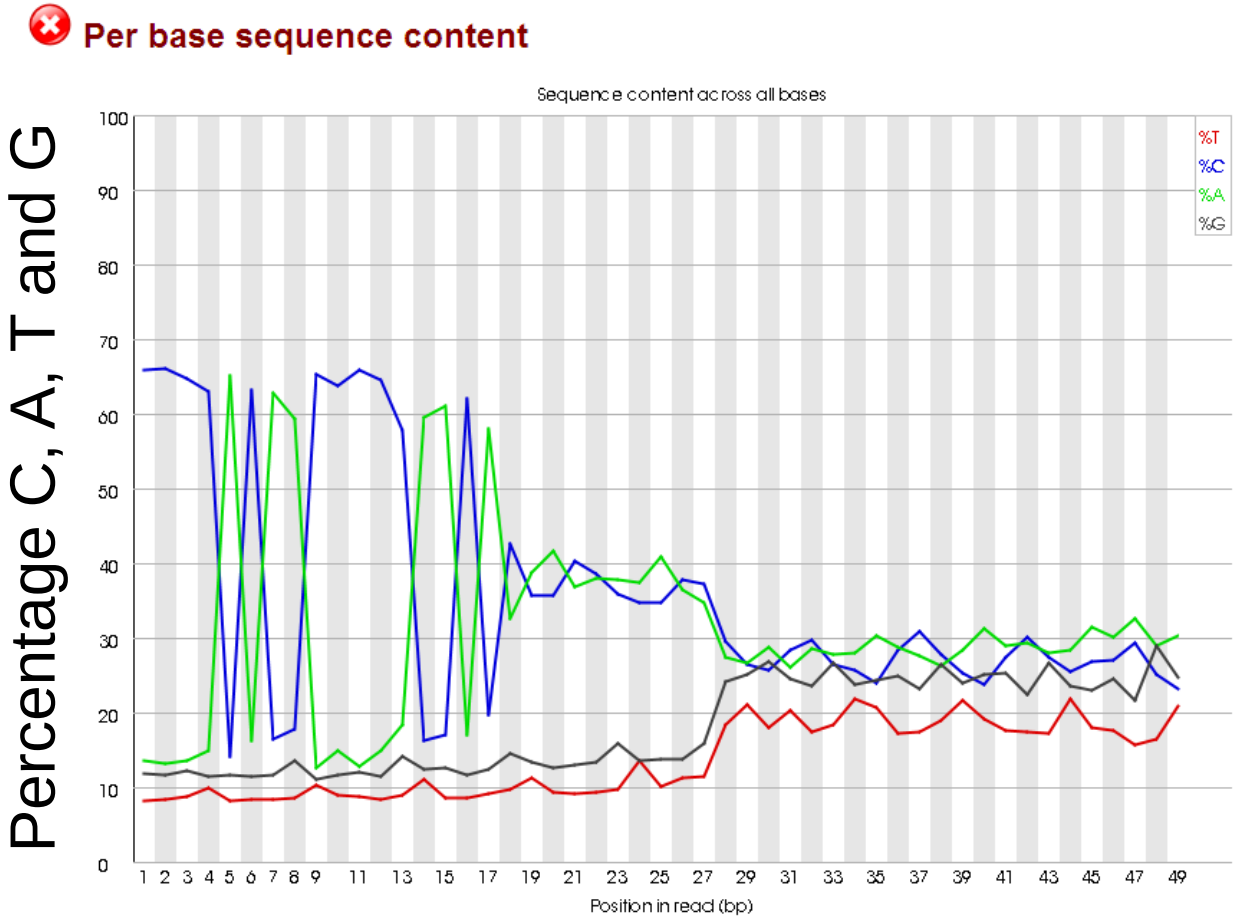
8:00-10:30 Duplication and overrepresented sequences

Normal nucleotide content

Per base sequence content



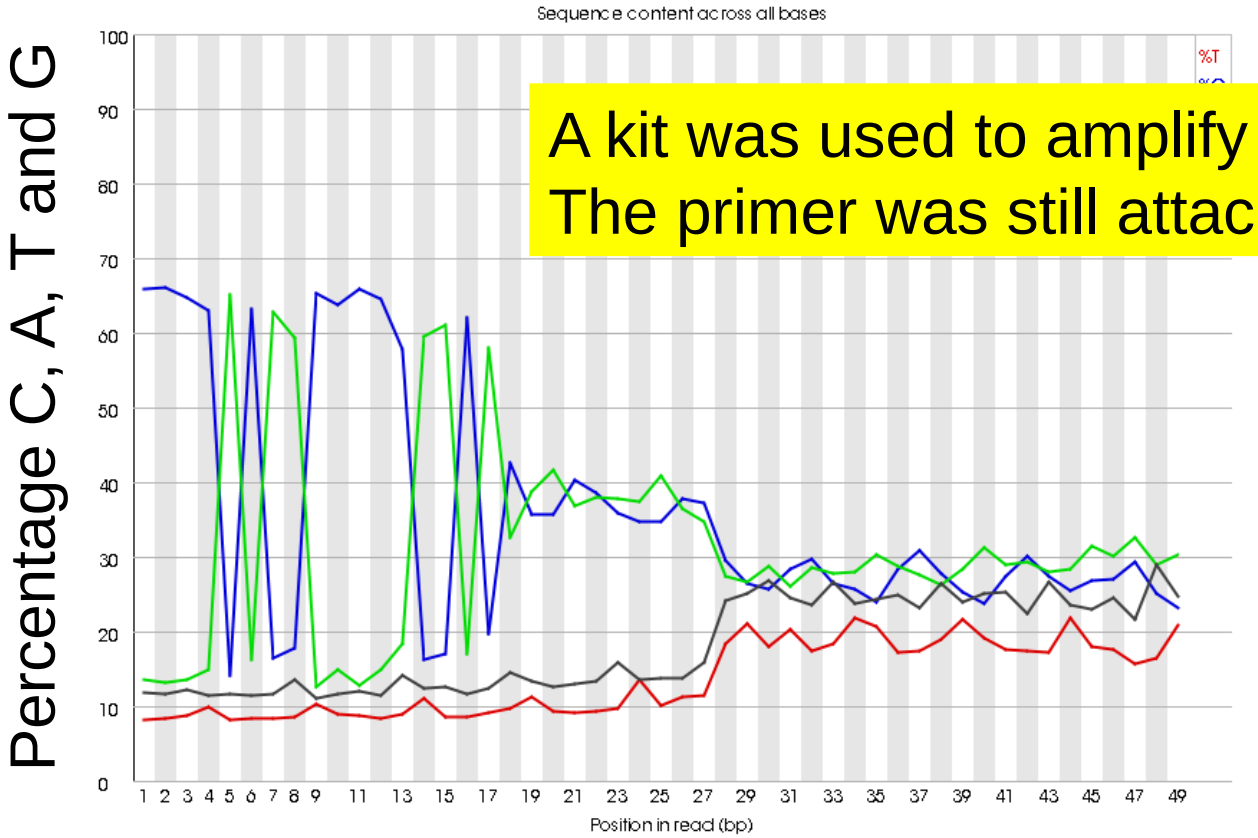
Exercise: What went wrong here?



Position in the reads

Exercise: What went wrong here?

Per base sequence content



A kit was used to amplify the sequences
The primer was still attached to the fragment

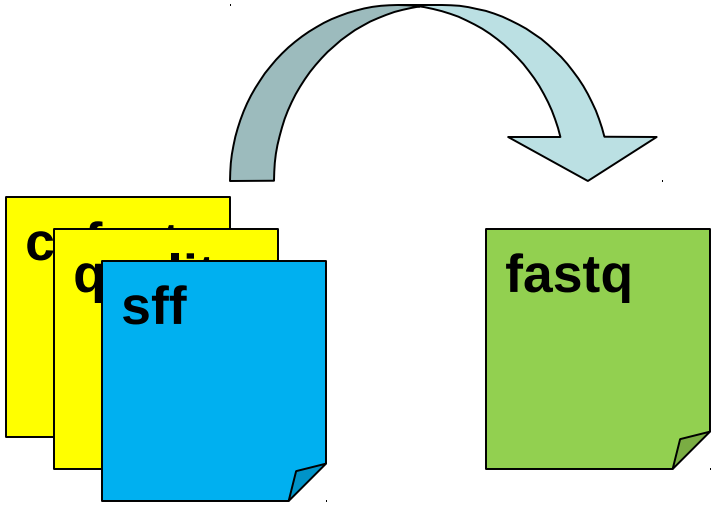
Back to summary

Position in the reads

Pre-processing and quality control of sequence data

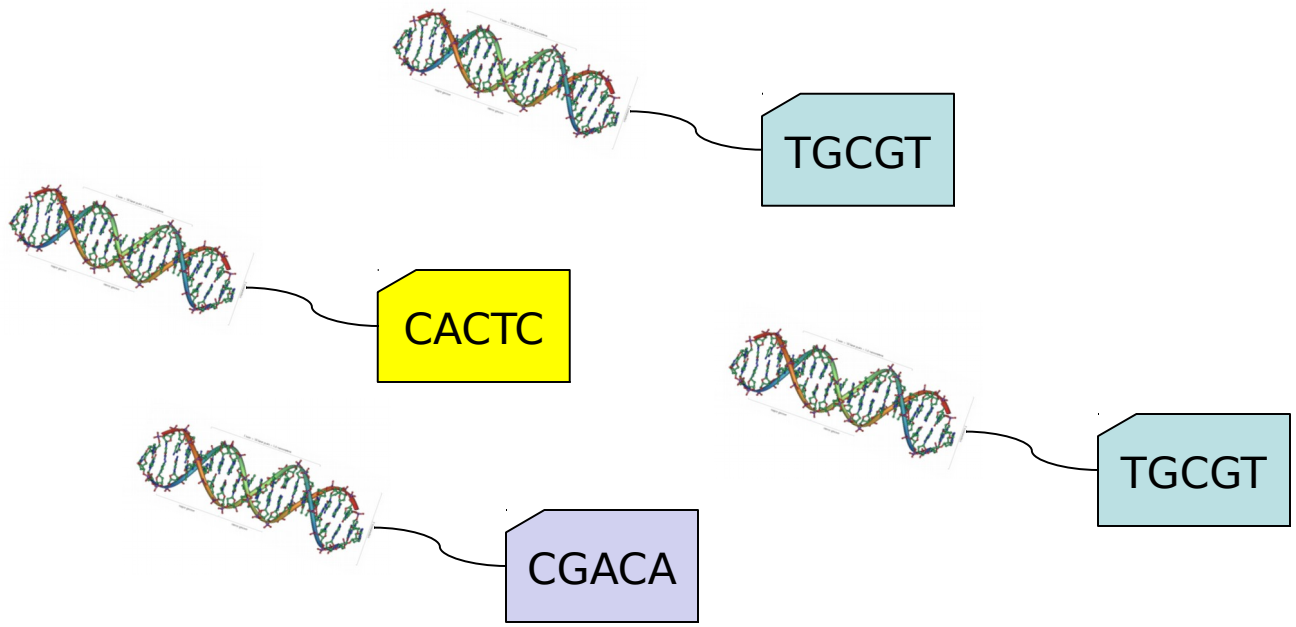
- In what format is sequence data provided?
- Is the data of good quality?
- **How do you pre-process raw data for further analysis?**
- How to align sequences to a reference genome?
- How do you judge if the alignment went well?

Data conversion



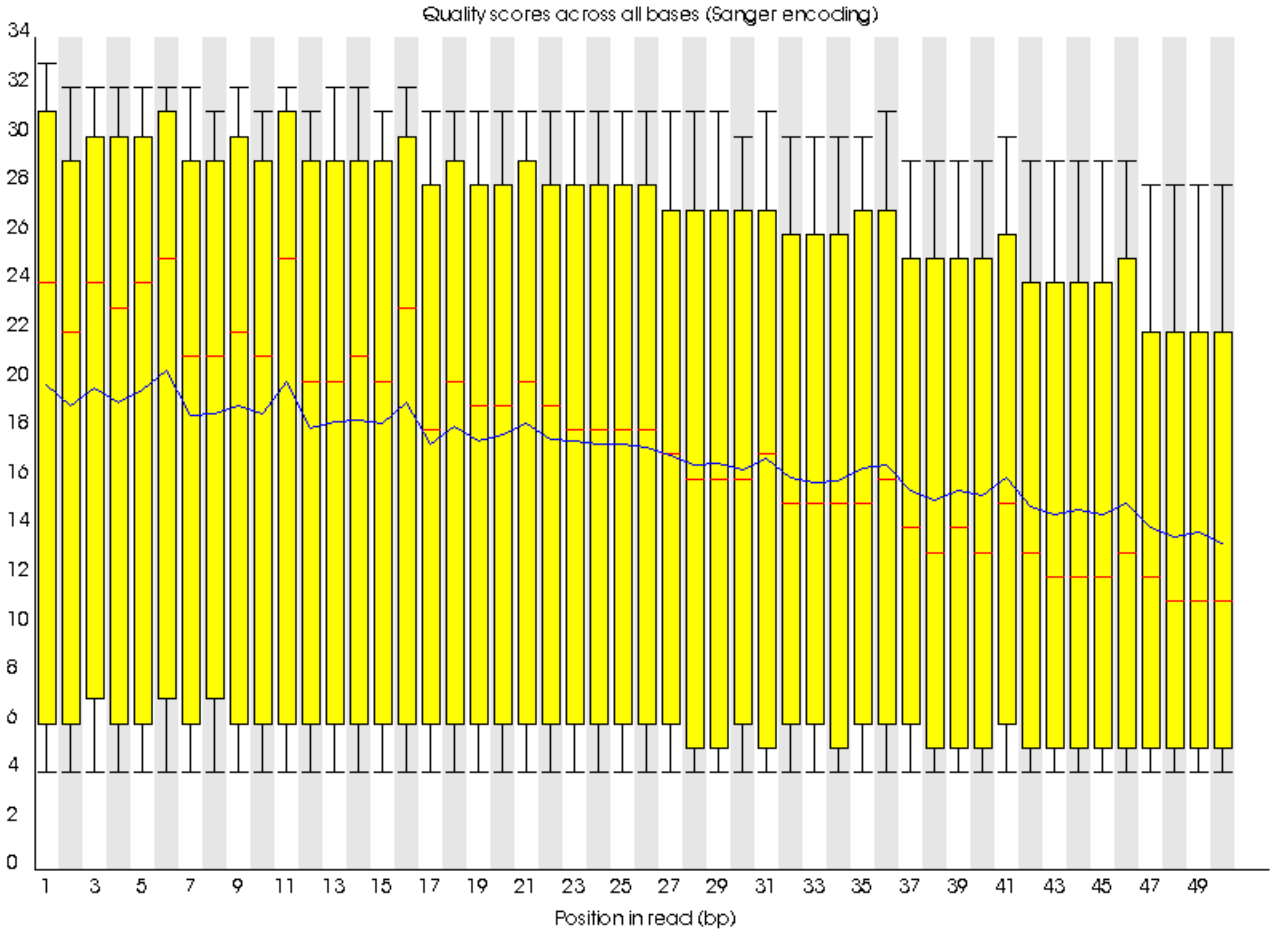
Most programs only work with fastq format

Data pre-processing – split per sample

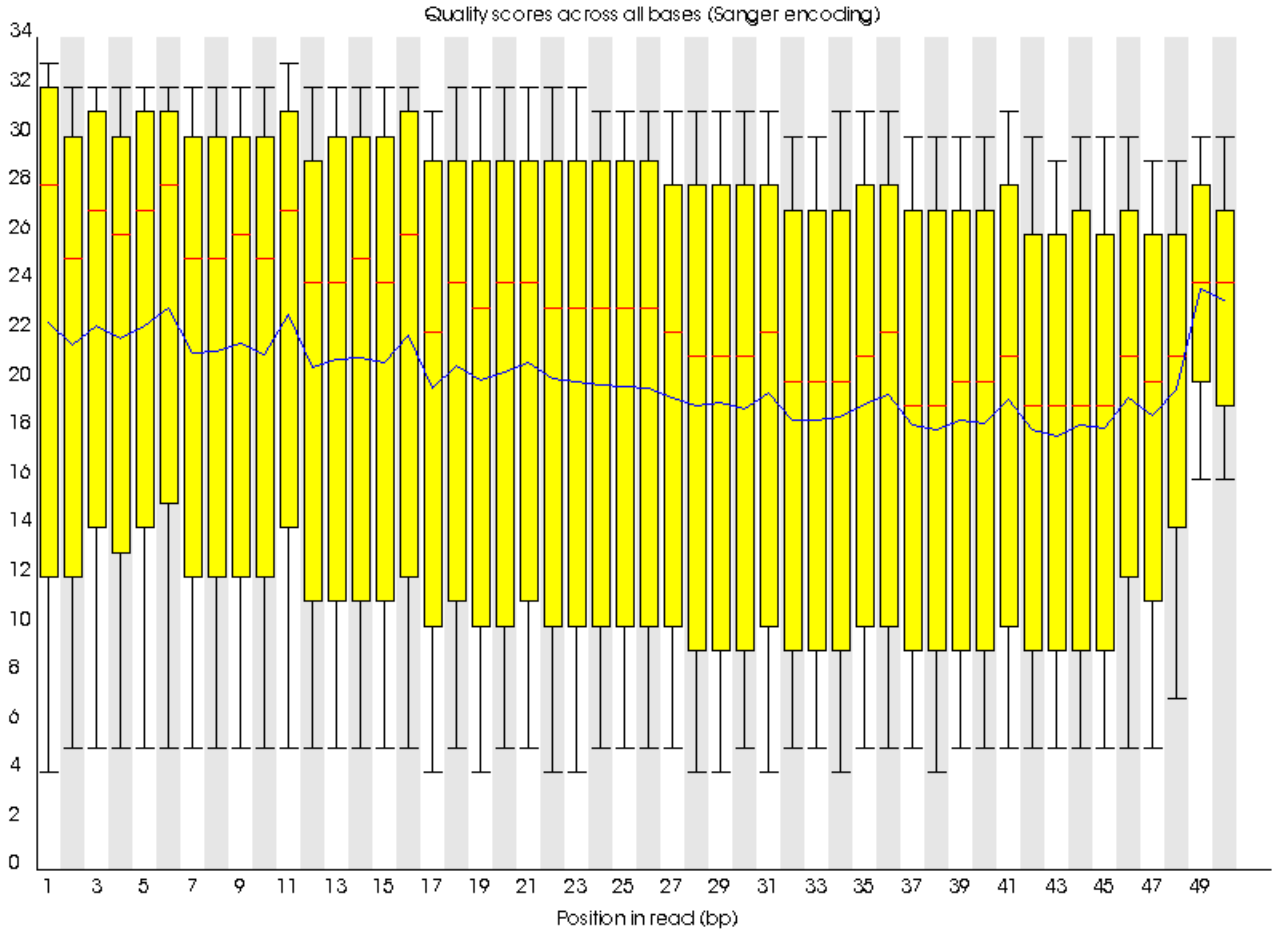



```
>E9108QN01BVB2T length=238 xy=0649_3411 region=1 run=R_2008_05_06_17_51_52_
CACTCCAGGAAACAGCTATGACCTCTGCCTGGAAAGCCAGGTGCCTGTGGGCAGAGCCCA
GGACCACAGGGCCAGGGGTATCTCGTGTTCCTGTCCTGGCCGCGGATCTTCTTCTCCATC
TCAGCGTCTGTCAGAGTCTCCAGCAGTGGGCACCACTGGTCCGCATCGCCCGTGTTCCGG
ATGGCAATCTCCACTGTGGGCAGAGGGTTCTCGCTACGAGGAGGGAGGCAGTGAGAGG
```

Quality trimming - before



Quality trimming - after



Summary:
How do you pre-process raw data for further analysis?

- Convert to a suitable data format
- Split experiment per sample
- Check the quality of the raw data
- Trim raw data on quality

Pre-processing and quality control of sequence data

- In what format is sequence data provided?
- Is the data of good quality?
- How do you pre-process raw data for further analysis?
- **How to align sequences to a reference genome?**
- How do you judge if the alignment went well?

How to align sequences to a reference genome?

- Dynamic programming doesn’t work for NGS
- The alternative: BWA (and others)
- Alignment in base- and colorspace

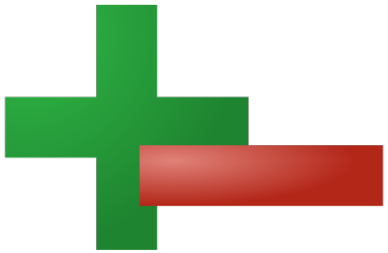
Pairwise sequence alignment with dynamic programming

+ True optimization
- Time to compute: O(n x m)

Not suitable for next generation sequencing

(unless you do this highly parallel on many computers at the same time)

BLAST



+ Works very good for large databases and not too many experiment sequences
- Index for the database is stored and accessed from disk

Accessing the disk every time is much slower than accessing the memory

BWA

- Based on Burrows-Wheeler transform
- Origin: data compression (bzip2)
- Does smart sorting of the database

David Wheeler
Michael Burrows

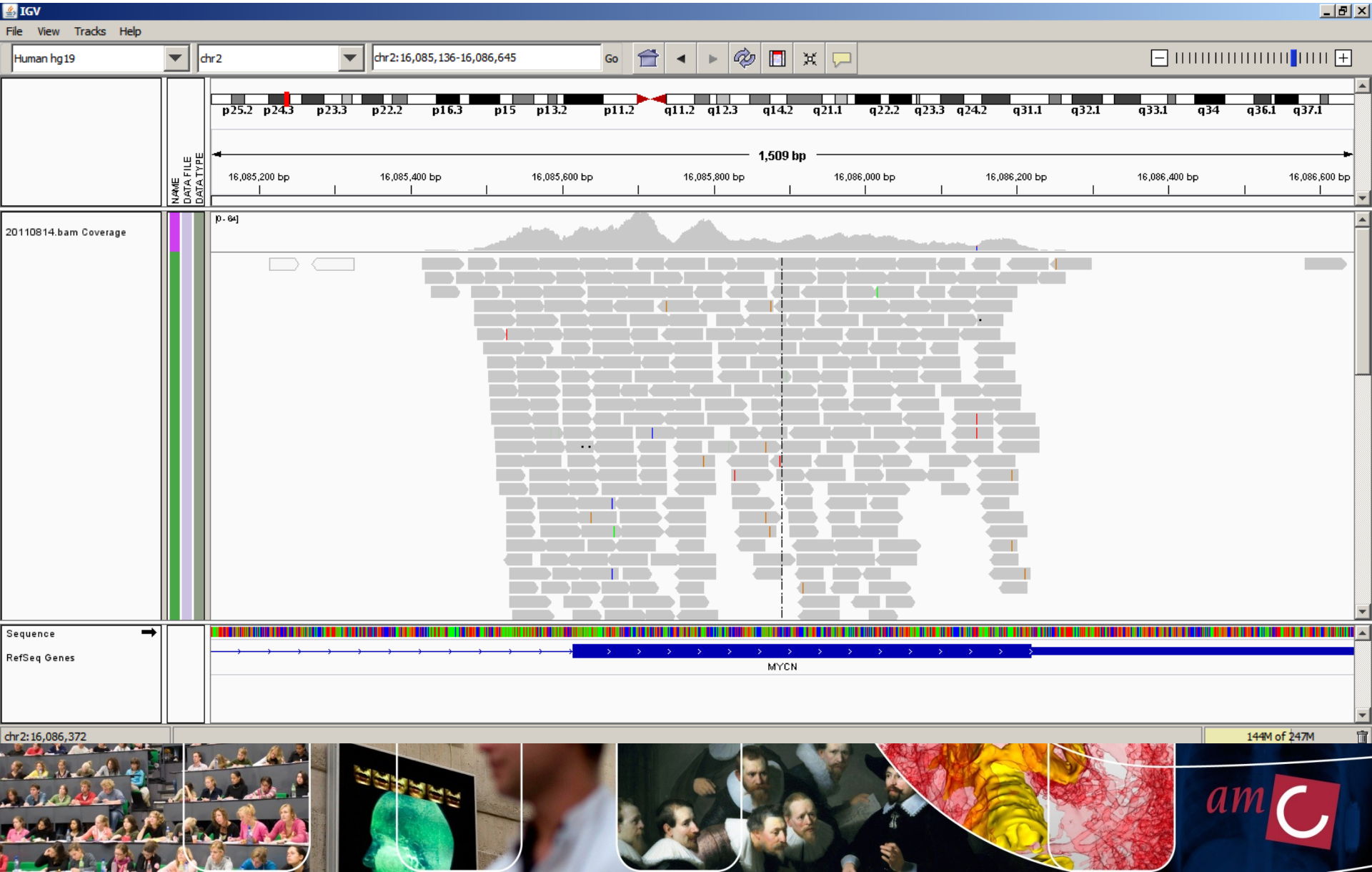
http://bio-bwa.sourceforge.net/

BWA

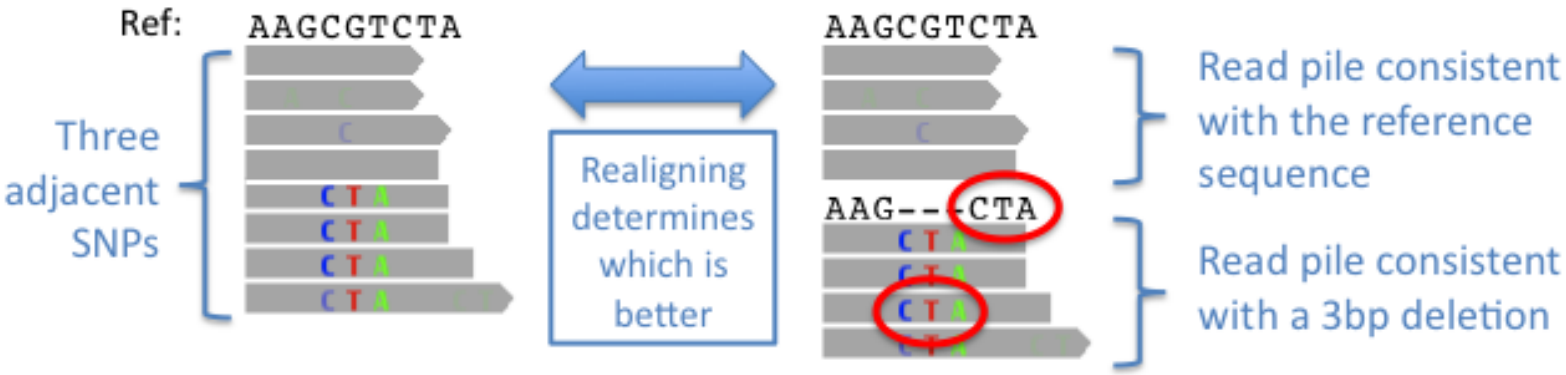
+ Stores the entire database in memory
+ Time to compute: O(m)
- Database can not be larger than 4GB

Suitable for human genome and next generation sequence data

Sequence alignment in genome viewer

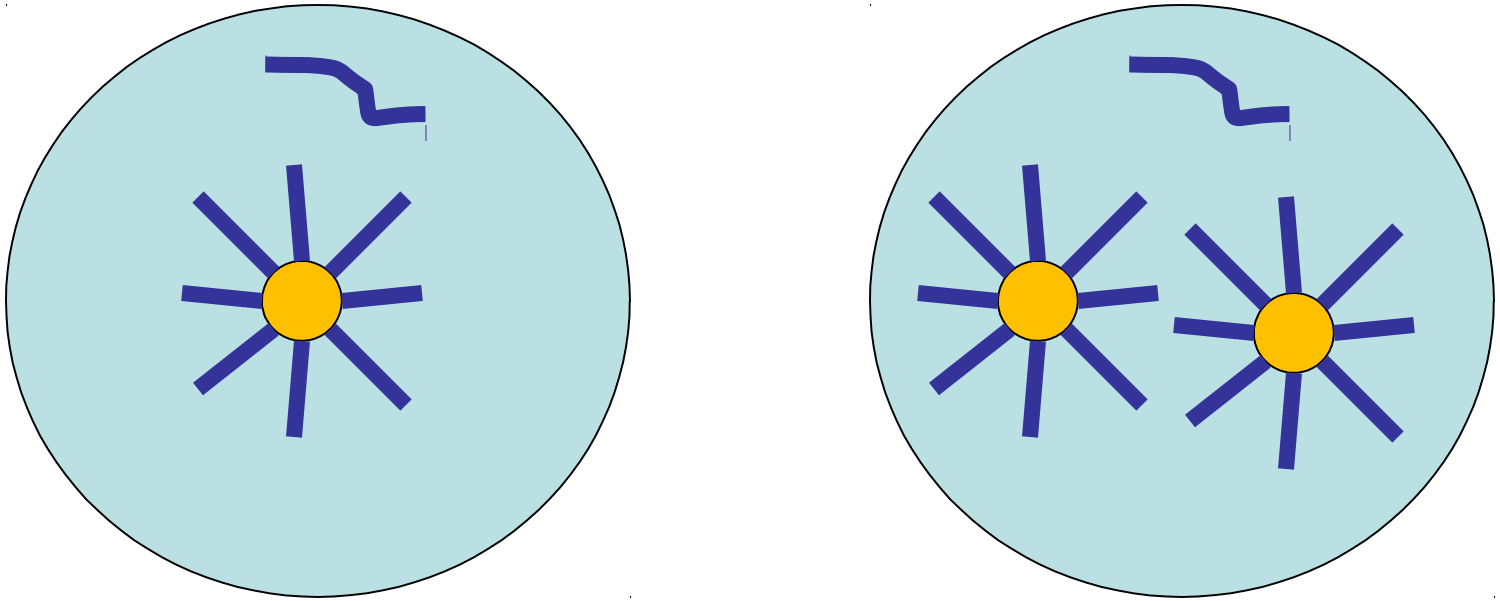


A closer look at the alignments: fix alignment errors



- BWA aligns sequences one by one
- Has problems with indels (insertions/deletions)
- Local realignment step to correct for these errors

Remove duplication artifacts from alignment files



Result: two sequences with one origin
Can be identified, because they will align to the same position

Summary

How to align sequences to a reference genome?

- You need a fast sequence alignment program like BWA (puts database in memory, time to align is linear to amount of sequences)
- Artifacts like alignment errors and duplication errors need to be fixed or removed

Pre-processing and quality control of sequence data

- In what format is sequence data provided?
- Is the data of good quality?
- How do you pre-process raw data for further analysis?
- How to align sequences to a reference genome?
- **How do you judge if the alignment went well?**

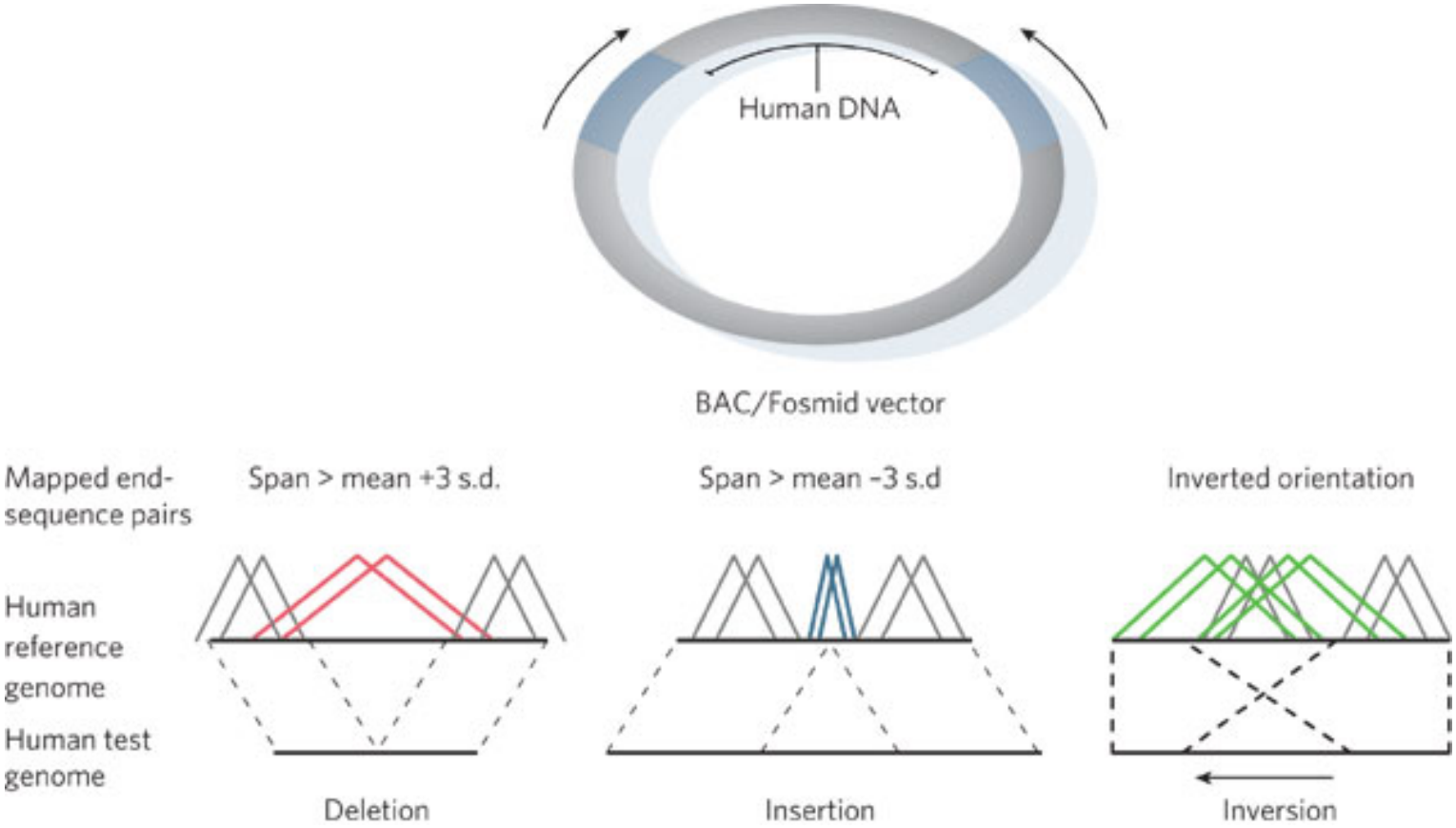
How can sequences align?

On a unique position
On many positions
With low confidence
Not

Check nr of aligned sequences (Picard tools)

CATEGORY	FIRST_OF_PAIR	SECOND_OF_PAIR	PAIR
TOTAL_READS	62003448	62003448	124006896
PF_READS	62003448	62003448	124006896
PCT_PF_READS	1	1	1
PF_NOISE_READS	598	522	1120
PF_READS_ALIGNED	58856753	58618988	117475741
PCT_PF_READS_ALIGNED	0.94925	0.945415	0.947332
PF_HQ_ALIGNED_READS	55113393	54966470	110079863
PF_HQ_ALIGNED_BASES	4942400720	4911919258	9854319978
PF_HQ_ALIGNED_Q20_BASES	4789247244	4660056419	9449303663
PF_HQ_MEDIAN_MISMATCHES	0	0	0
PF_HQ_ERROR_RATE	0.003681	0.005545	0.00461
MEAN_READ_LENGTH	90	90	90
READS_ALIGNED_IN_PAIRS	58225227	58225318	116450545
PCT_READS_ALIGNED_IN_PAIRS	0.98927	0.993284	0.991273
BAD_CYCLES	0	0	0
STRAND_BALANCE	0.49934	0.499927	0.499633
PCT_CHIMERAS	0.004945	0.004958	0.004952
PCT_ADAPTER	0.000023	0.000016	0.000019

Paired-end sequencing



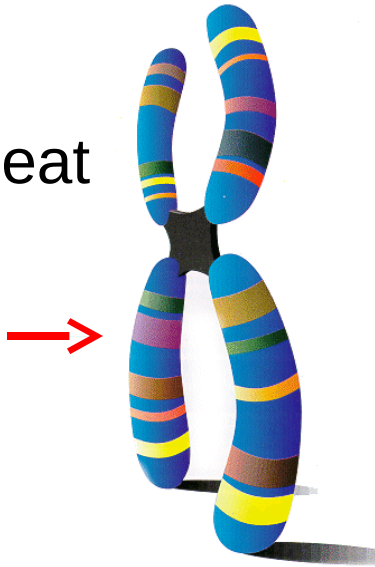
The Human Genome Structural Variation Working Group, Nature 2007

After alignment: check the insert size

MEDIAN_INSERT_SIZE		478
MIN_INSERT_SIZE		1
MAX_INSERT_SIZE		240953475
MEAN_INSERT_SIZE		474.443017
STANDARD_DEVIATION		33.616421
READ_PAIRS		50613089
PAIR_ORIENTATION	FR	

How do pairs align?

Not
Unique and expected insert size
Unique, but larger/smaller insert size
Unique, but reversed
On different chromosomes
One unique, the other in a repeat
Etc….



Check if most sequences are properly paired (samtools flagstat)

124006896 + 0 in total (QC-passed reads + QC-failed reads)
14993642 + 0 duplicates
117476853 + 0 mapped (94.73%:nan%)
124006896 + 0 paired in sequencing
62003448 + 0 read1
62003448 + 0 read2
115408830 + 0 properly paired (93.07%:nan%)
116451572 + 0 with itself and mate mapped
1025281 + 0 singletons (0.83%:nan%)
713462 + 0 with mate mapped to a different chr
594419 + 0 with mate mapped to a different chr (mapQ>=5)

Summary:
How do you judge if the alignment went well?

- Percentage of aligned sequence reads
- Percentage of properly paired sequences
- Percentage of duplicate reads
- … and several other metrics

- Have a closer look at the alignments

Pre-processing and quality control of sequence data

- In what format is sequence data provided?
- How do you check if the data is of good quality?
- How do you pre-process raw data for further analysis?
- How to align sequences to a reference genome?
- How do you judge if the alignment went well?