MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Next-generation sequencing and quality control:

An introduction
2016

Sebastian Schmeier
s.schmeier@massey.ac.nz
http://sschmeier.com/bioinf-workshop/

THE ENGINE
OF THE NEW
NEW ZEALAND

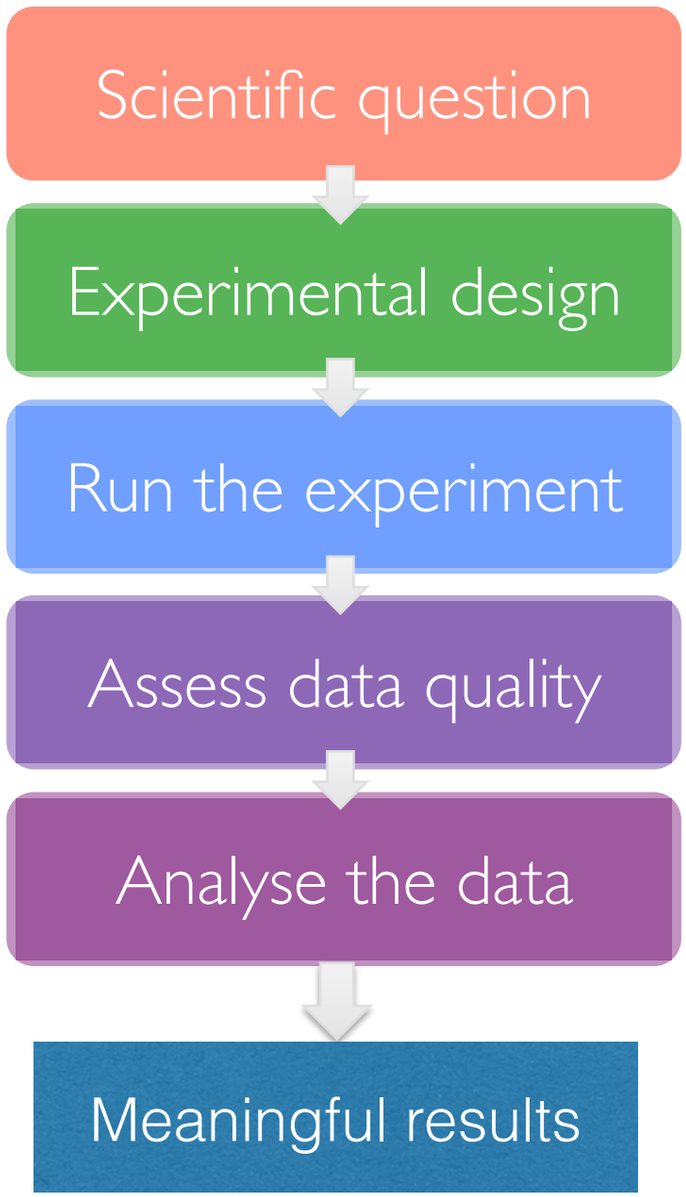
Overview

- Typical workflow of a genomics experiment
- Genome versus transcriptome
- DNA sequencing
- The FASTQ-file format
- Sources of sequencing errors
- Quality assessment of a sequencing run
- Pre-processing sequencing data

Learning outcomes

- Being able to describe what next generation sequencing is.
- Being able to describe the FastQ-sequence format and understand what information it contains and how it can be used.
- Describe what the Phred-based quality score is.
- Be able to describe the sources of sequencing errors.
- Being able to compute, investigate and evaluate the quality of sequence data from a sequencing experiment.
- Be able to distinguish between a good and a bad sequencing run.
- Being able to describe the steps involved in cleaning sequencing data.

Typical workflow of a genomics experiment



The essential part to make all downstream analysis work

This is where the biological experiment happens

This is the bottleneck of the whole experiment

Genome versus transcriptome

Genome

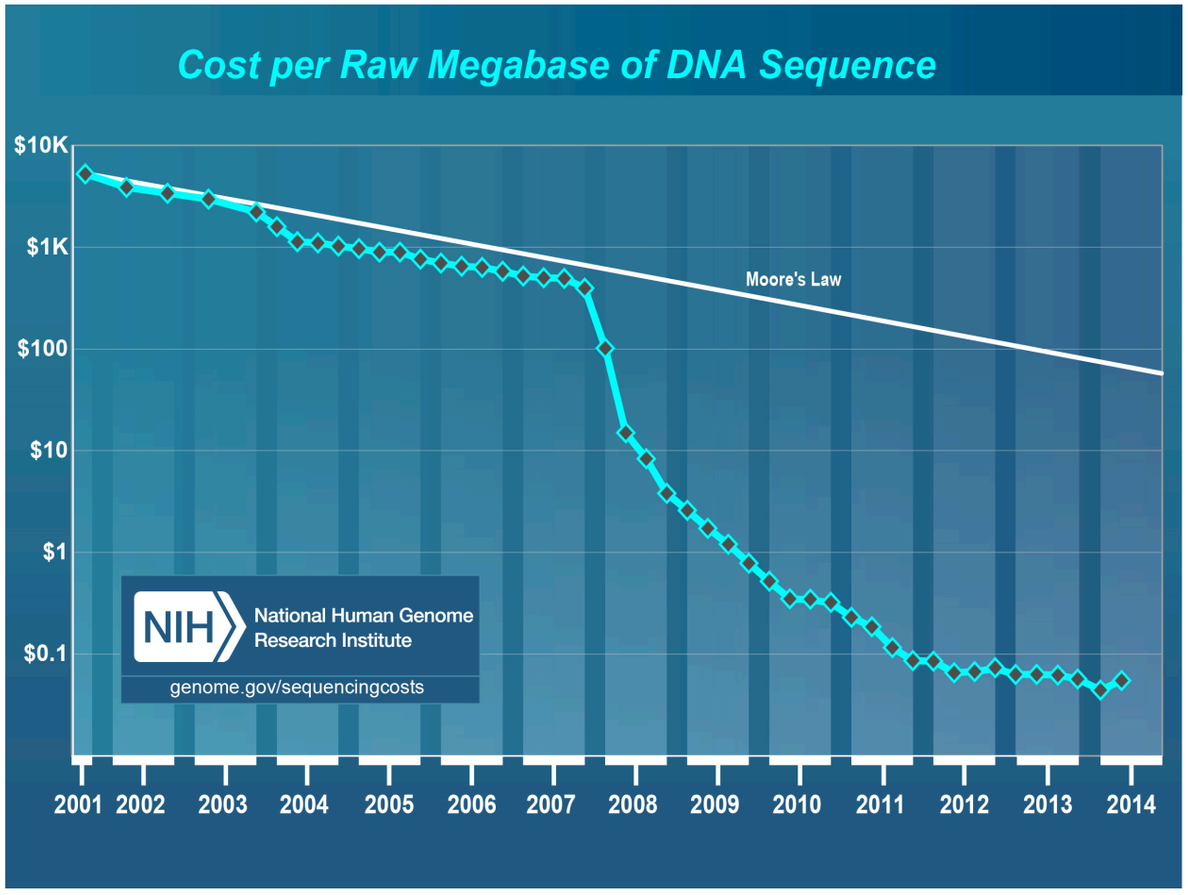
The entirety of an organism's ancestral information. It is encoded either in DNA or, for many types of viruses, in RNA.

Transcriptome

The set of all RNA molecules, including messenger RNA, ribosomal RNA, transfer RNA, and other non-coding RNA produced in one or a population of cells

Name	Base Pairs	
HIV	9,749	9.7kb
E.Coli	4,600,000	4.6MB
Yeast	12,100,000	12.1Mb
Drosophila	130,000,000	130MB
Homo sapiens	**3,200,000,000**	**3.2GB**
marbled lungfish	130,000,000,000	130Gb
"Amoeba" dubia	670,000,000,000	670Gb

DNA sequencing



DNA sequencing is the process of determining the nucleotide order of a given DNA fragment.

First-generation sequencing:

1977 Sanger sequencing method development (chain-termination method)

2001, Sanger method produced a draft sequence of the human genome

Next-generation sequencing (NGS)

Demand for low-cost sequencing has driven the development of high-throughput sequencing (or NGS) technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently

2004 454 Life Sciences marketed a parallelized version of pyrosequencing

Result of a sequencing run

Short read sequences

- The result of NGS technology are a collection of short nucleotide sequences (reads) of varying length (~40-400nt) depending on the technology used to generate the reads
- Usually a reads quality is good at the beginning of the read and errors accumulate the longer the read gets
  => **IMPORTANT**

Illumina sequencing

MiSeq:

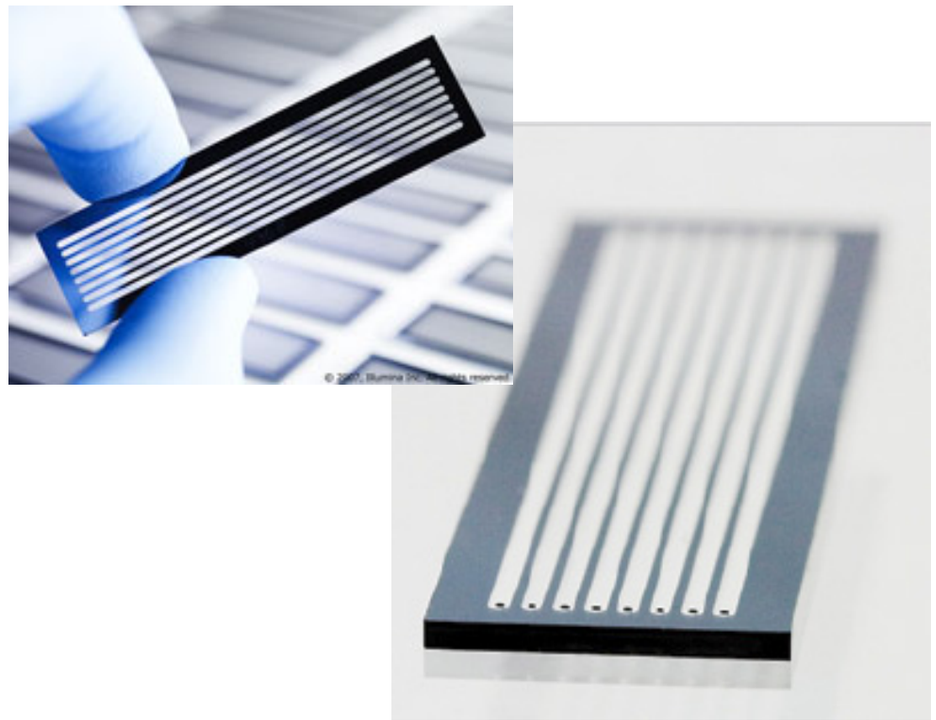
- Bench-top sequencer
- Produces around 30 million reads/run
- Reads are up to 250nt

HiSeq:

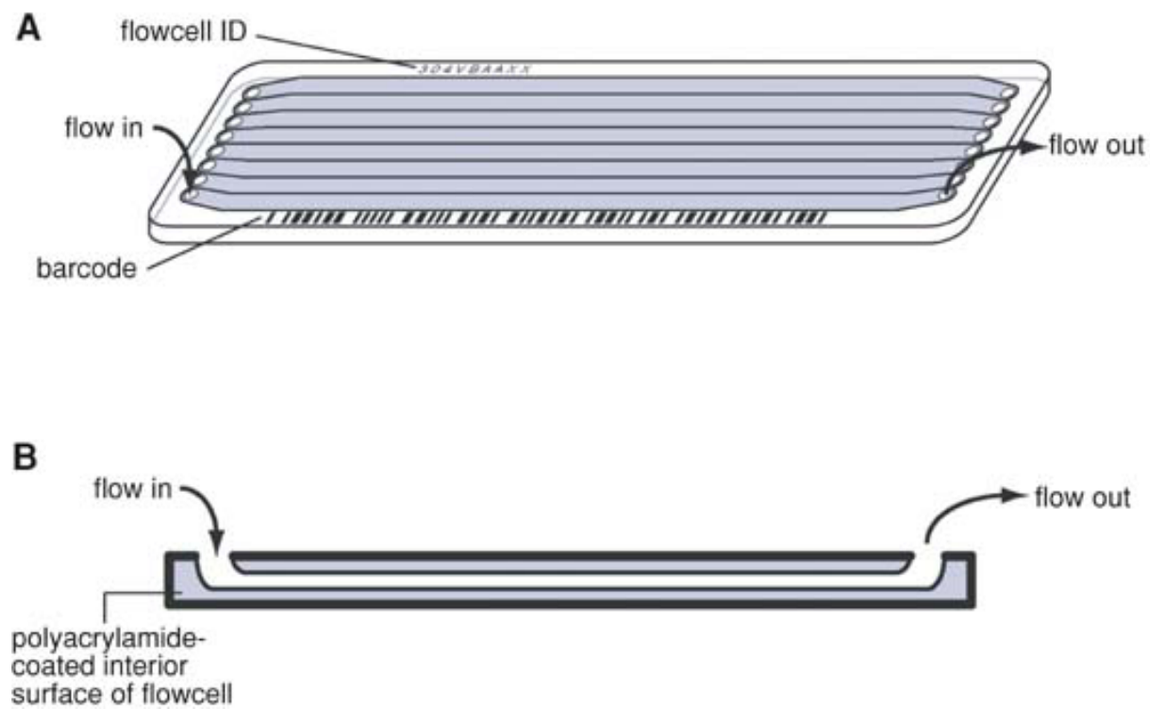
- Large-scale sequencer
- 4 billion reads/run
- Reads up to 150nt

The Illumina systems accumulate errors towards the end of the read sequence.

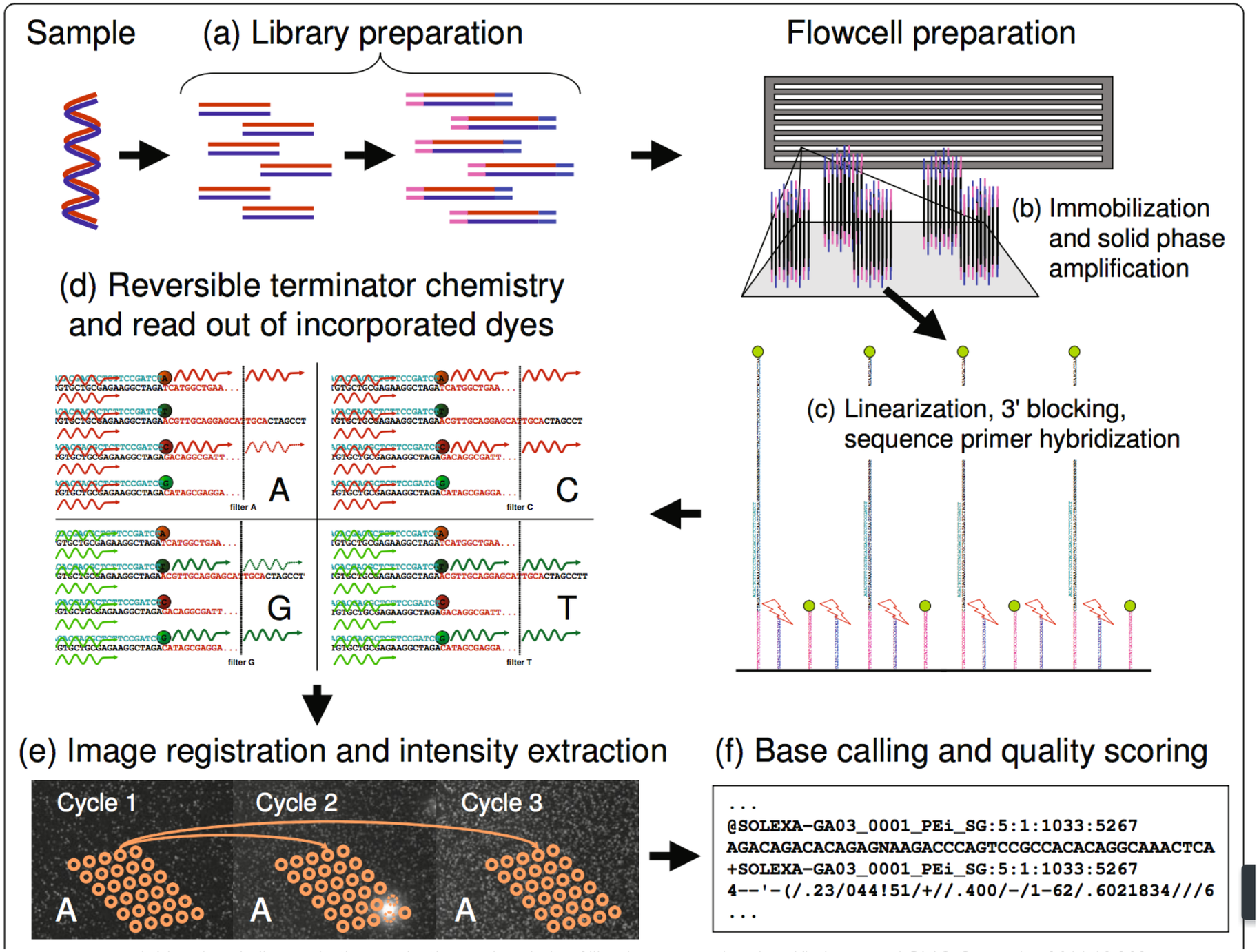
Illumina sequencing (2)



- An Illumina flowcell is a surface to which seq. adaptors are covalently attached.
- DNA with complementary adaptors is attached, clonally amplified, and then sequenced by synthesis
- Each flowcell is subdivided into hundreds of tiles



Illumina sequencing (3)



Addressing challenges in the production and analysis of illumina sequencing data. Kirchner et al. BMC Genomics, 2011,12:382

The FASTQ-file format

The file-format that you will encounter soon is called FASTQ

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTC
+
‘’*((((***+))%%%++)(%%%%).1***-+*’’))**55CCF>>>>>>
```

The FASTQ-file format (2)

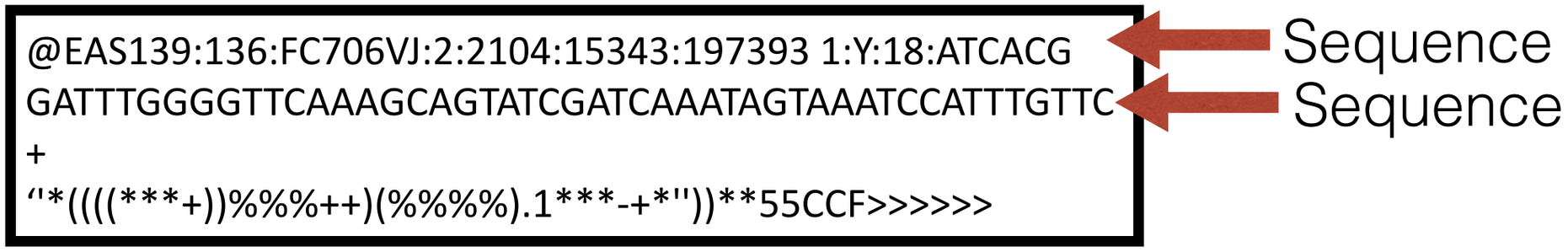
The file-format that you will encounter soon is called FASTQ

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTC
+
‘’*((((***+))%%%++)(%%%%).1***-+*’’))**55CCF>>>>>>
```

Sequence ID

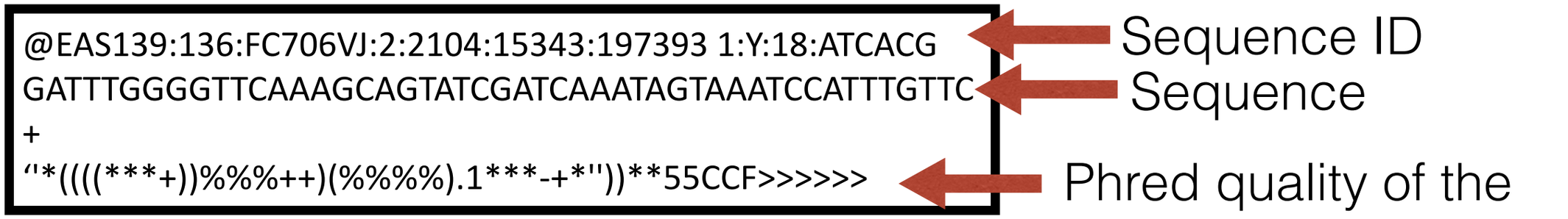
The FASTQ-file format (3)

The file-format that you will encounter soon is called FASTQ



The FASTQ-file format (4)

The file-format that you will encounter soon is called FASTQ



The FASTQ-file format (5)

The file-format that you will encounter soon is called FASTQ

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTC
+
‘’*((((***+))%%%++)(%%%%).1***-+*’’))**55CCF>>>>>>
```

Sequence ID

Casava 1.8 the format

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Sequencing errors

- Sequencing errors or mis-called bases occur when a sequencing method calls one or more bases incorrectly leading to an incorrect read.
- The chance of a sequencing error is generally known and quantifiable, thanks to extensive testing and calibration of the sequencing machines
- Each base in a read is assigned a quality score, indicating confidence that the base has been called correctly.

https://www.broadinstitute.org/crd/wiki/index.php/Sequencing_error

The FASTQ-file format (6)

- FastQ: Phred base quality

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTC
+
''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>
```

Phred quality of the corresponding nucleotide (ASCII code)

- One ASCII character per nucleotide.
- Encodes for a quality $Q = -10*\log_{10}(P)$, where P is the error probability

$-10*\log_{10}(0.1) =$

The Relationship Between Quality Score and Base Call Accuracy

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Q	ASCII	P
1	"	0.79433
2	#	0.63096
3	$	0.50119
4	%	0.39811
5	&	0.31623
6	'	0.25119
7	(	0.19953
8	)	0.15849
9	*	0.12589
10	+	0.10000
11	,	0.07943

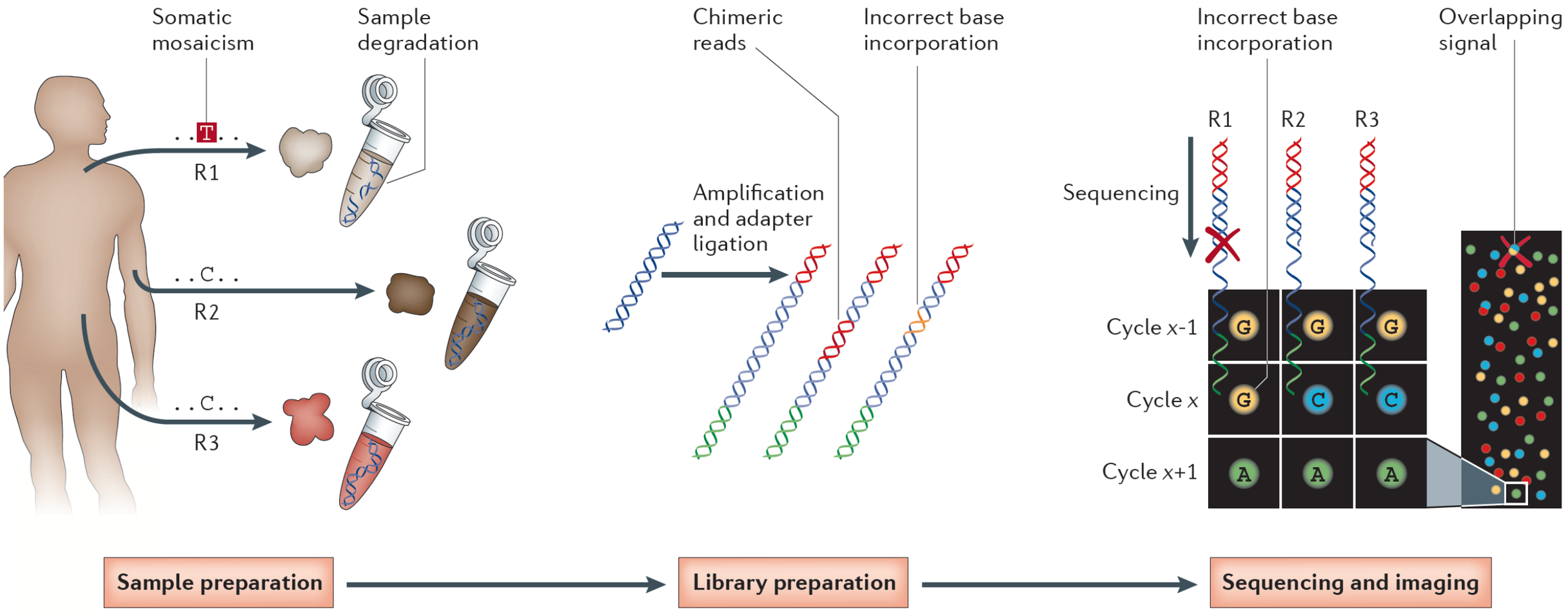
Sources of sequencing errors

- The importance and the relative effect of each error source on downstream applications depend on many factors, such as:

  - sample acquisition
  - reagents
  - tissue type
  - protocol
  - instrumentation
  - experimental conditions
  - analytical application
  - the ultimate goal of the study.

The role of replicates for error mitigation in next-generation sequencing. Robasky et al. Nature Reviews Genetics, 2014, 15, 56-62

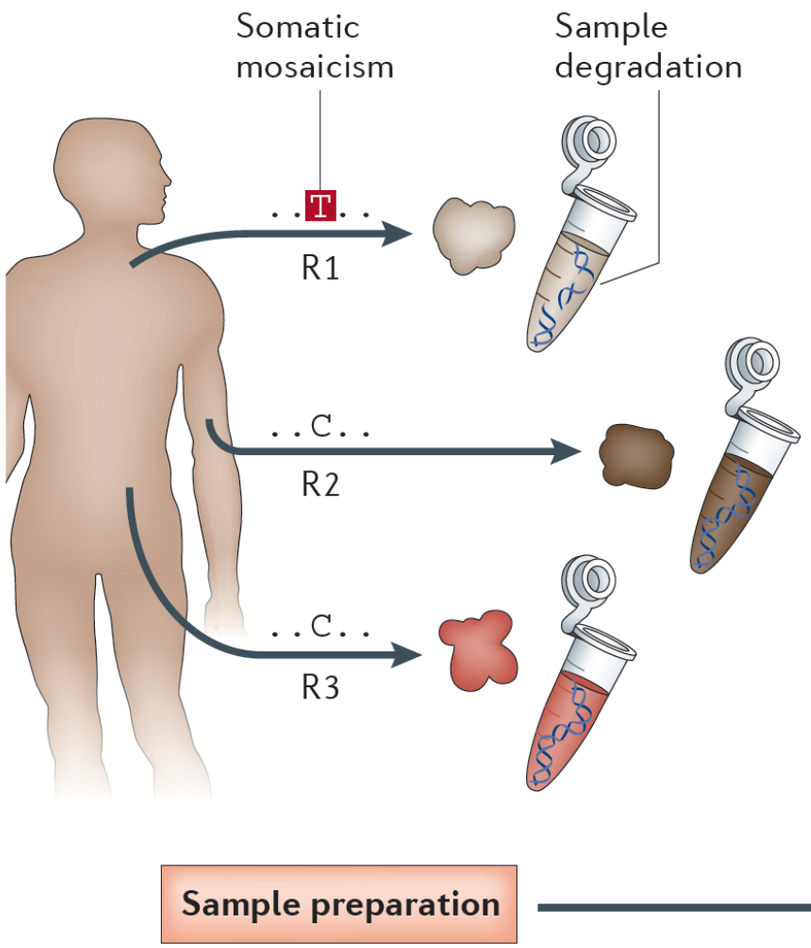
Sources of sequencing errors (2)



Sequencing errors can stem from any time point throughout the experimental workflow, including initial sequence preparation, library preparation and sequencing.

The role of replicates for error mitigation in next-generation sequencing. Robasky et al. Nature Reviews Genetics, 2014, 15, 56-62

Sources of sequencing errors (3)

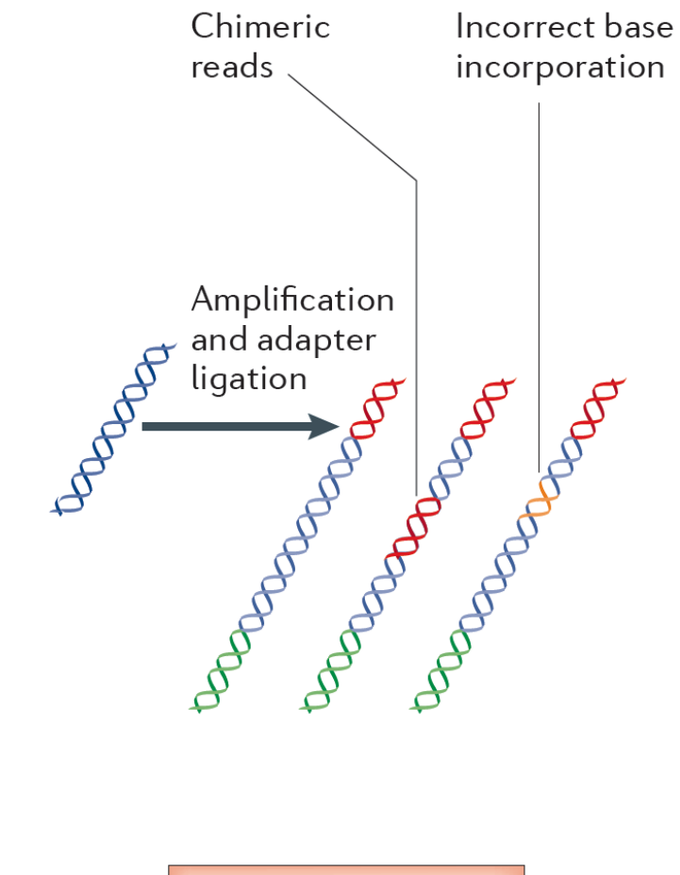


Sample preparation

- User errors; for example, mislabelling
- Degradation of DNA and/or RNA from preservation methods; for example, tissue autolysis, nucleic acid degradation and crosslinking during the preparation of formalin-fixed, paraffin-embedded (FFPE) tissues
- Alien sequence contamination; for example, those of mycoplasma and xenograft hosts
- Low DNA input

The role of replicates for error mitigation in next-generation sequencing. Robasky et al. Nature Reviews Genetics, 2014, 15, 56-62

Sources of sequencing errors (4)

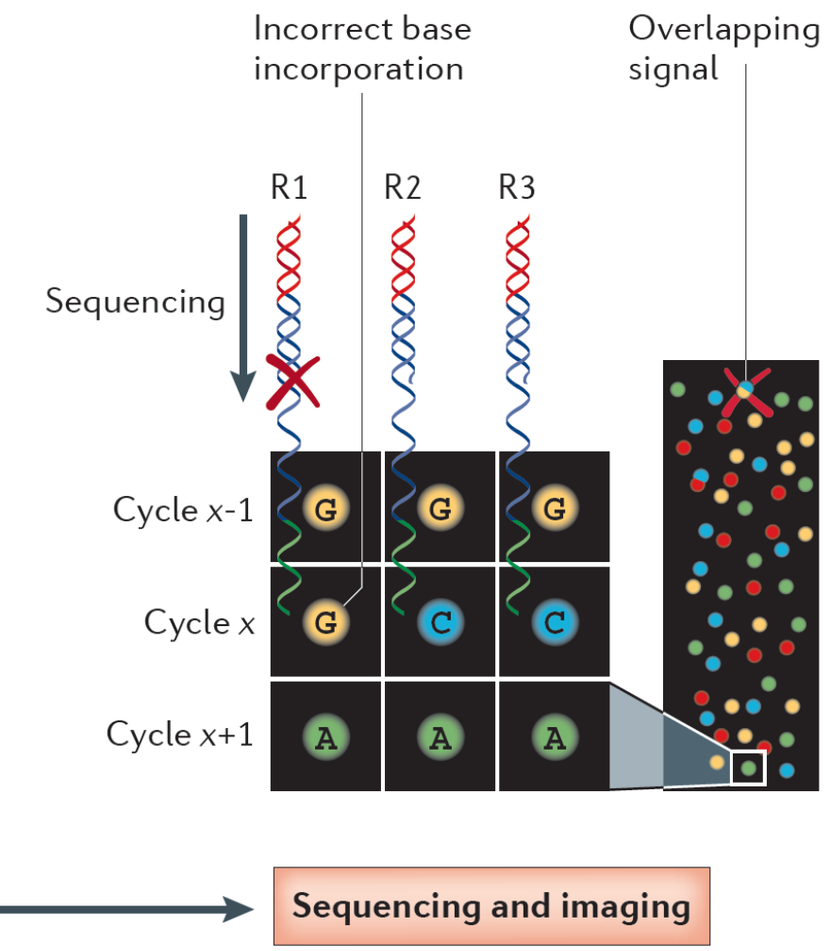


Library preparation

- User errors; for example, carry-over of DNA from one sample to the next and contamination from previous reactions
- PCR amplification errors
- Primer biases; for example, binding bias, methylation bias, biases that result from mispriming, nonspecific binding and the formation of primer dimers, hairpins and interfering pairs, and biases that are introduced by having a melting temperature that is too high or too low
- 3′–end capture bias that is introduced during poly(A) enrichment in high-throughput RNA sequencing
- Private mutations; for example, those introduced by repeat regions and mispriming over private variation
- Machine failure; for example, incorrect PCR cycling temperatures
- Chimeric reads
- Barcode and/or adaptor errors; for example, adaptor contamination, lack of barcode diversity and incompatible barcodes

The role of replicates for error mitigation in next-generation sequencing. Robasky et al. Nature Reviews Genetics, 2014, 15, 56-62

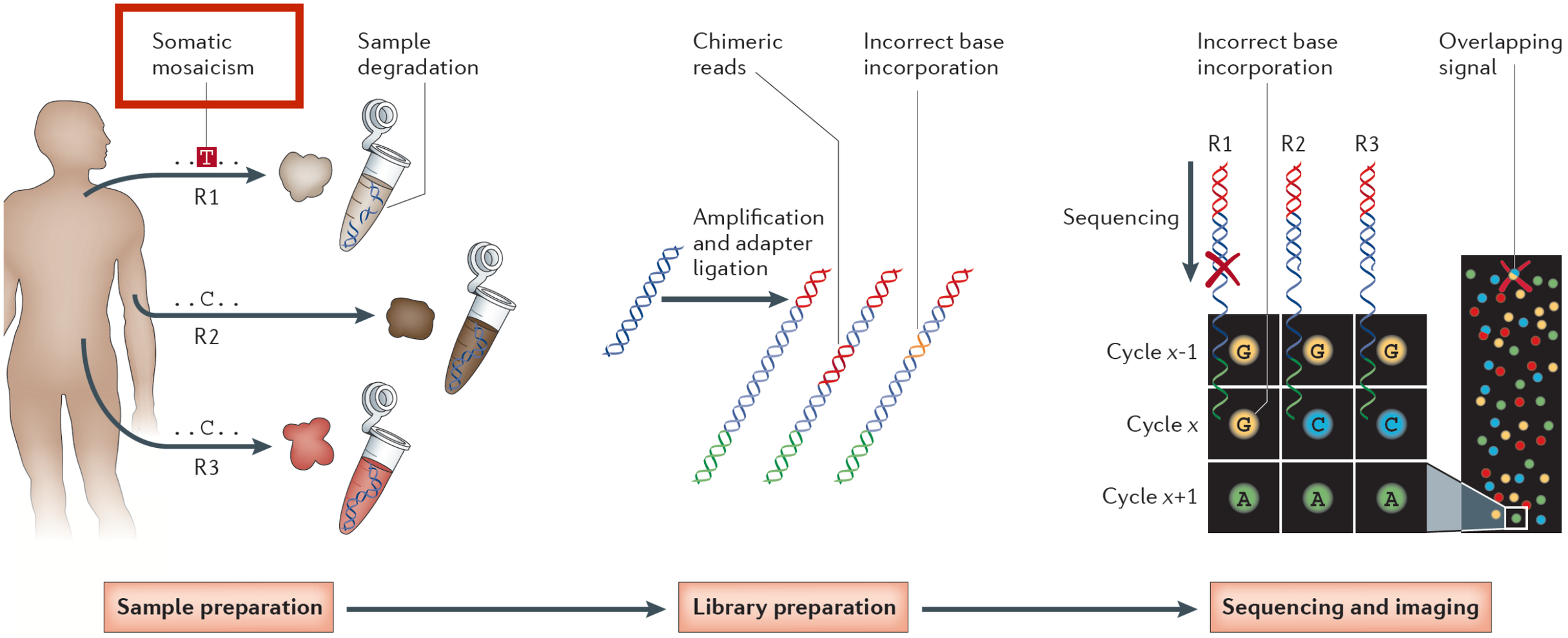
Sources of sequencing errors (5)



Sequencing and imaging

- User errors; for example, cluster crosstalk caused by overloading the flow cell
- Dephasing; for example, incomplete extension and addition of multiple nucleotides instead of a single nucleotide
- 'Dead' fluorophores, damaged nucleotides and overlapping signals
- Sequence context; for example, GC richness, homologous and low-complexity regions, and homopolymers
- Machine failure; for example, failure of laser, hard drive, software and fluidics
- Strand biases

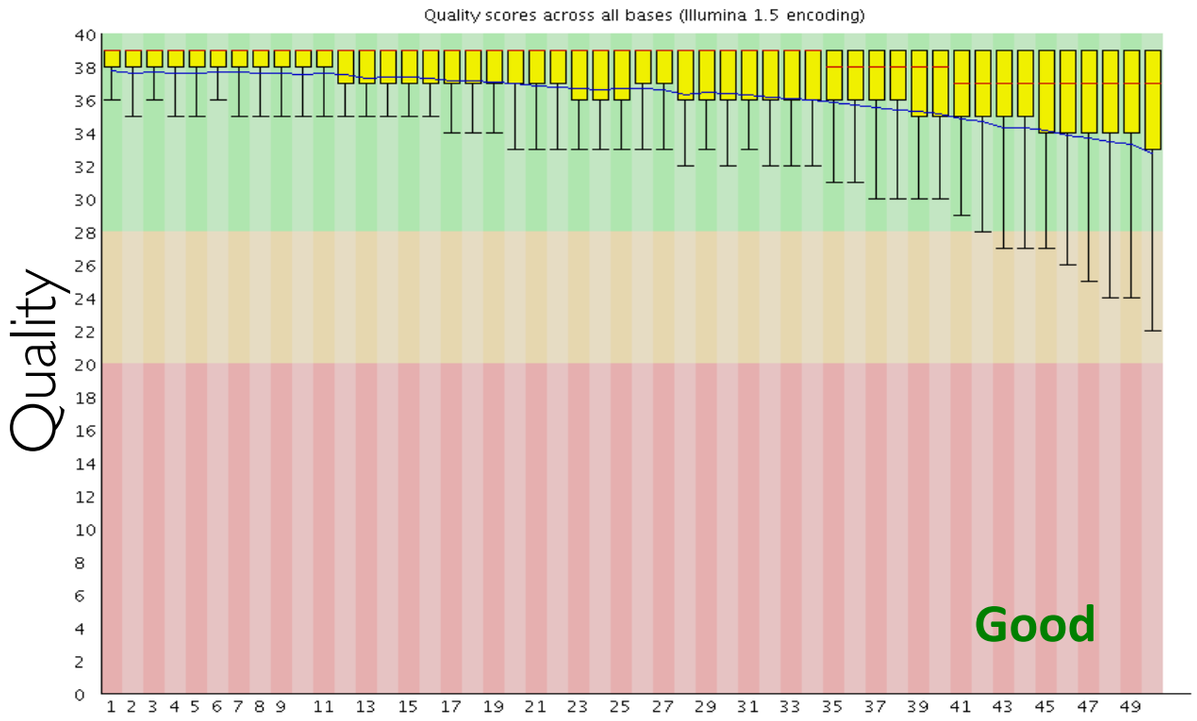
Sources of sequencing errors (6)

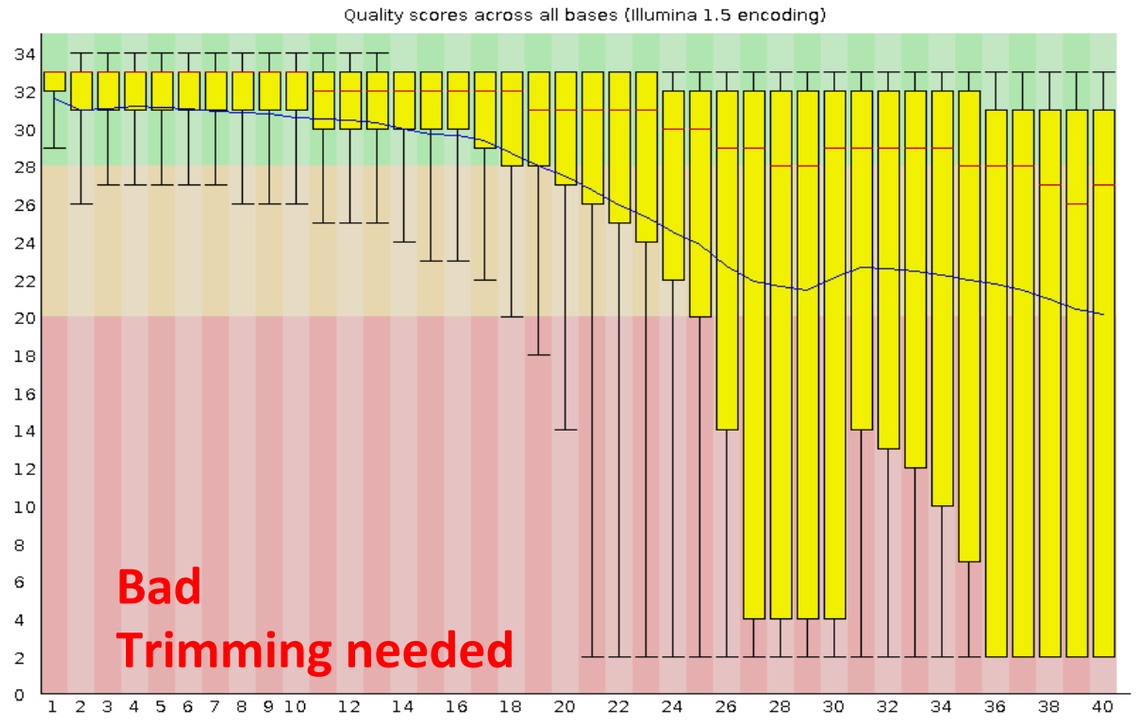


It is important to remember that one get somatic (acquired) mutations as well. These are not sequencing error but can be mistaken for them.

The role of replicates for error mitigation in next-generation sequencing. Robasky et al. Nature Reviews Genetics, 2014, 15, 56-62

Assessing quality: Reads

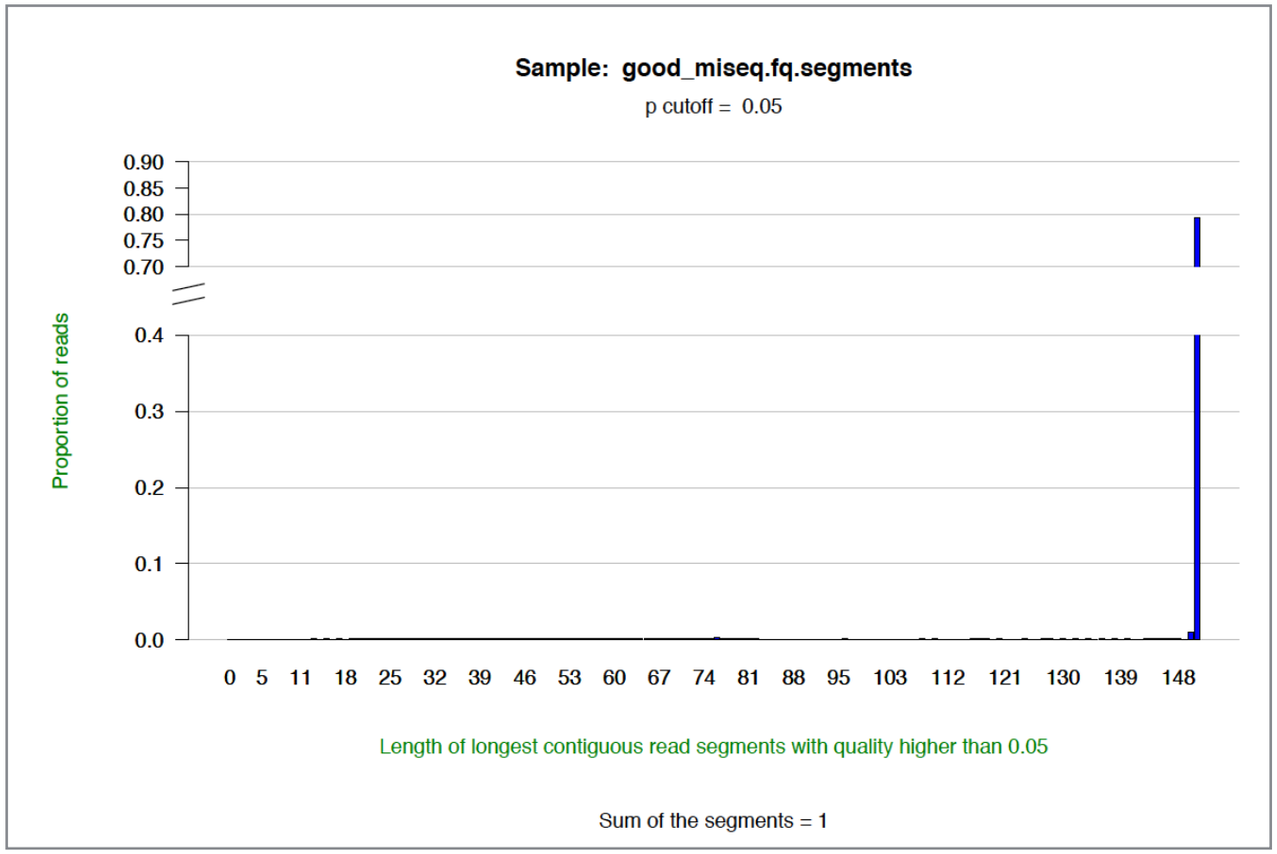


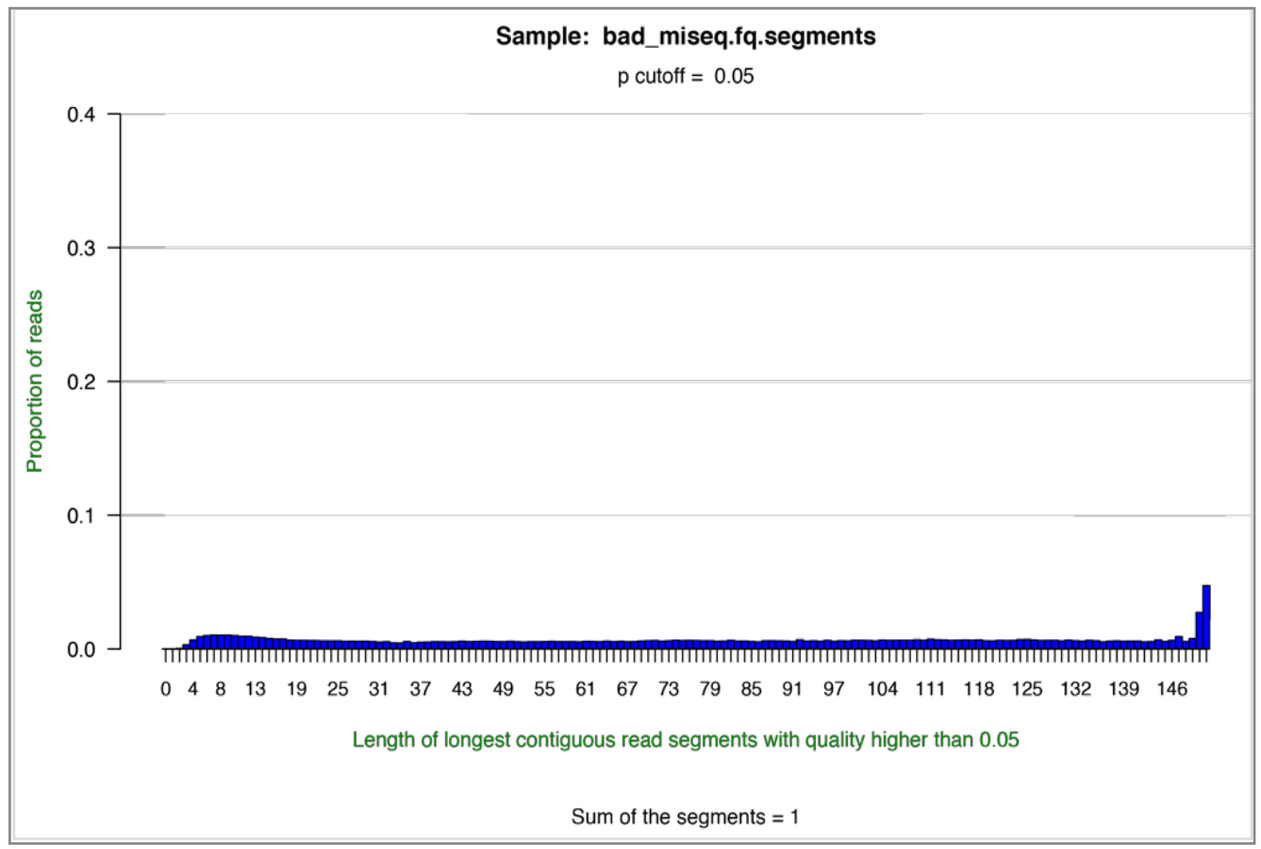


- If the quality of the reads is bad we can trim the nucleotides that are bad of the end of the reads
- Not trimming the end has a huge influence on downstream processes, e.g. assemblies

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

Assessing quality: Reads (2)





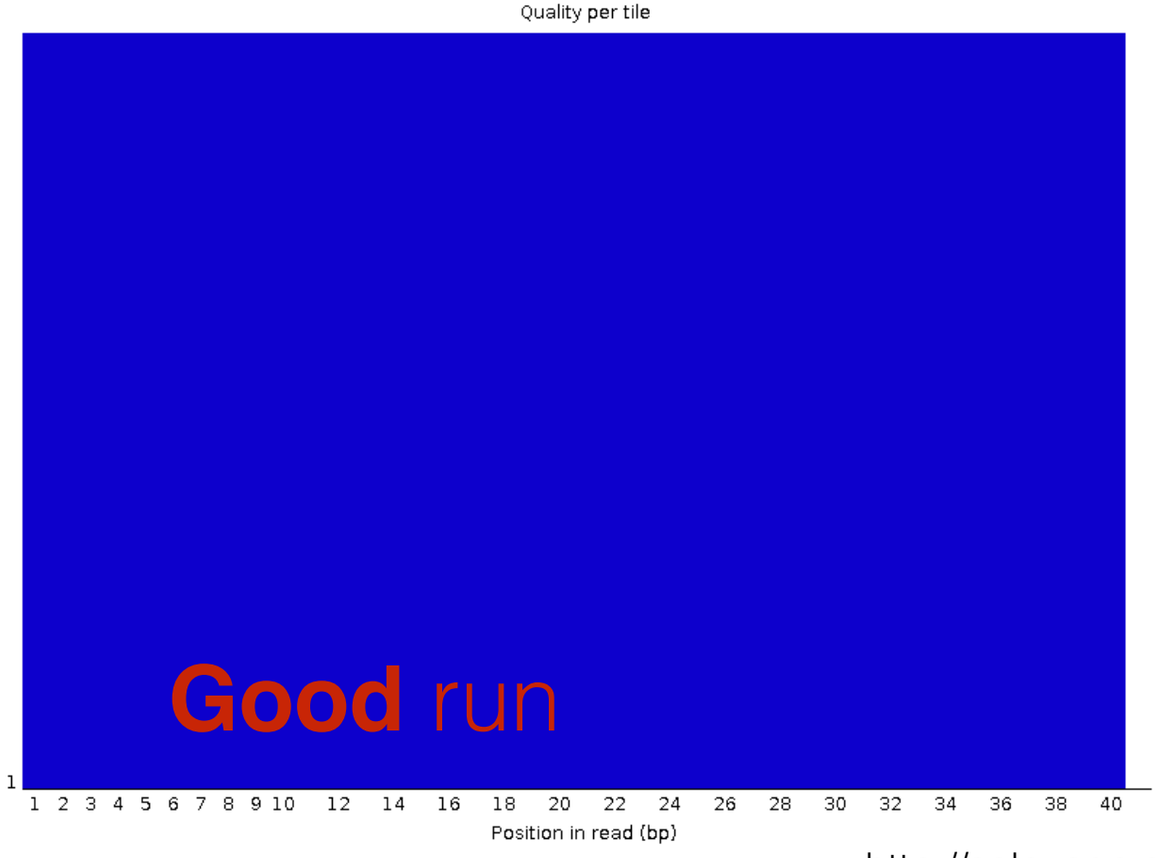
http://solexaqa.sourceforge.net/

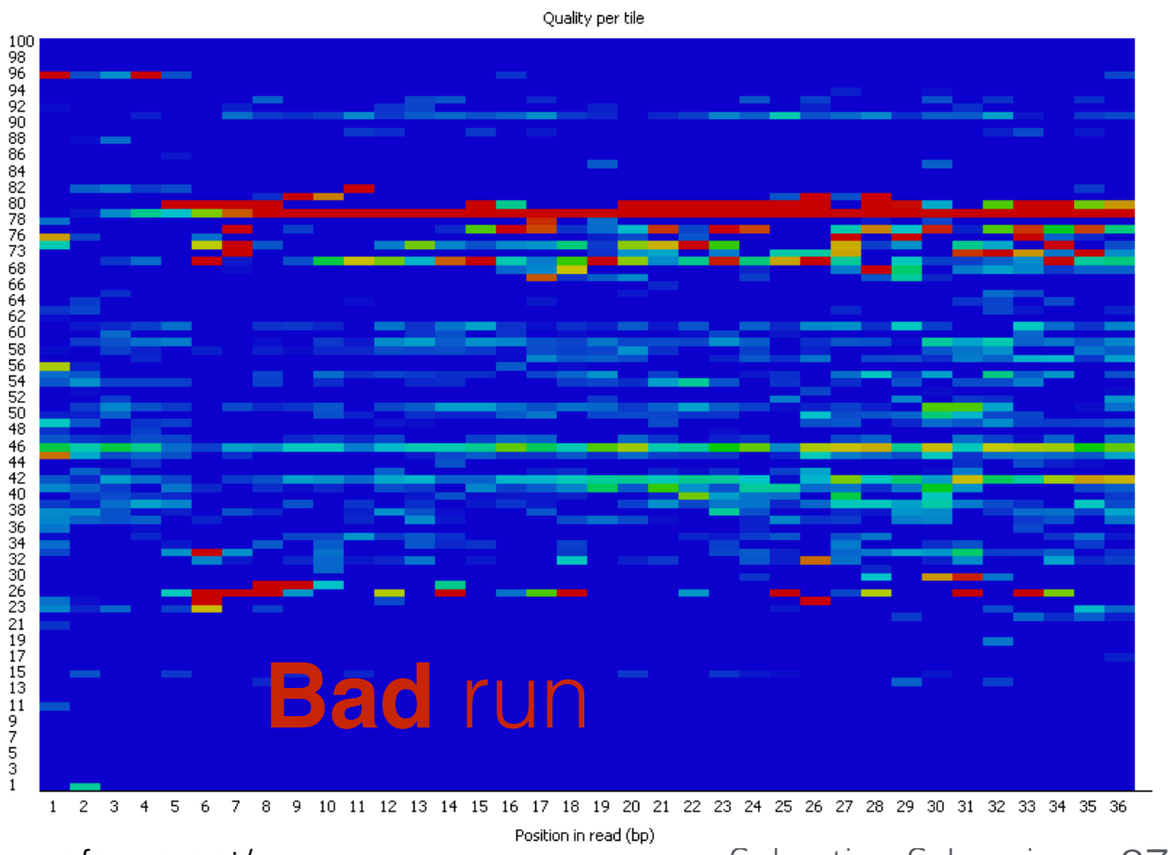
Assessing quality: Tiles

- One can assess also the quality of a run based on the tiles of a lane of a flowcell
  - ➢ spot problems with a particular tile on a lane, e.g. Bubbles in the reagents
- The homogeneity of the Illumina process ensures that the relative frequencies are similar from tile to tile and distributed uniformly across each tile
  - ➢ when the machine is functioning properly
- Major discrepancies in these conditions can be discerned by sight
- Many such discrepancies are small and their effects are limited to one, or a few, tiles.

Assessing quality: Tiles (2)

- Encoded in the FASTQ-file is the flowcell tile from which each read came.
- The graph allows you to look at the quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flowcell.

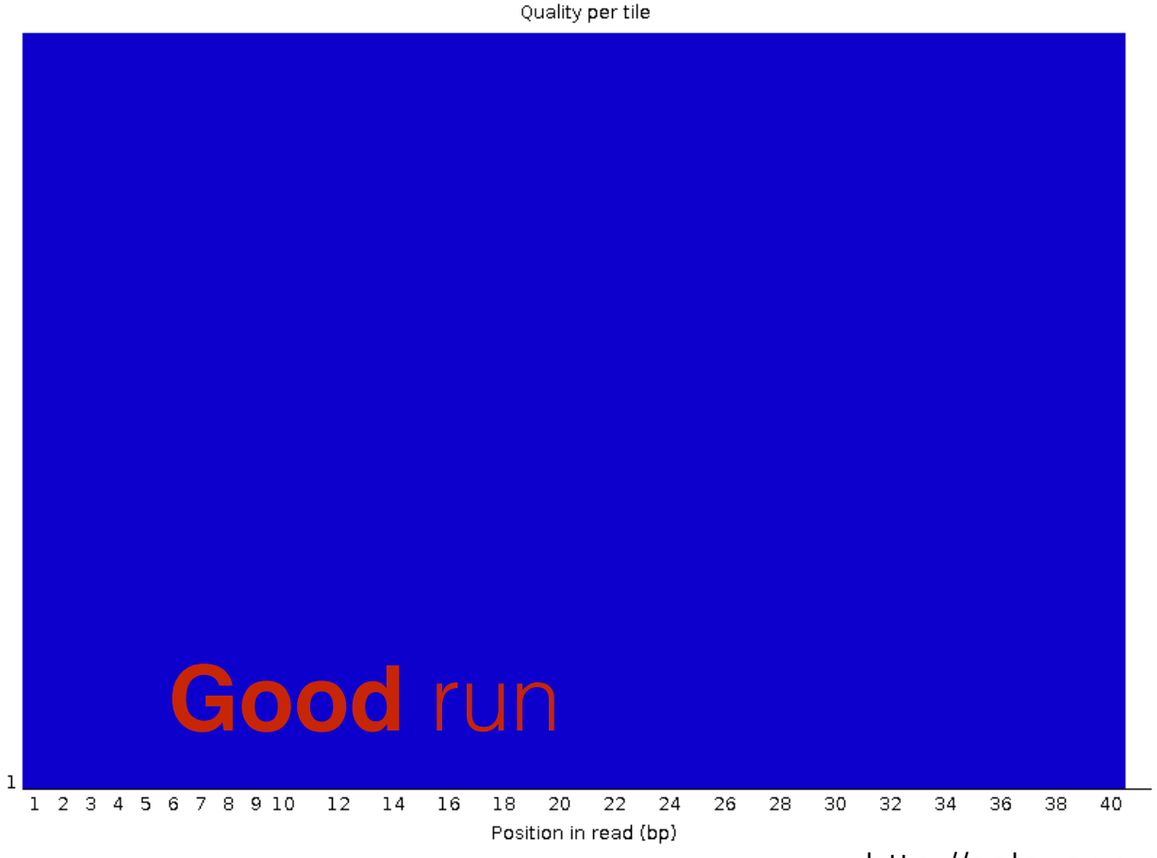


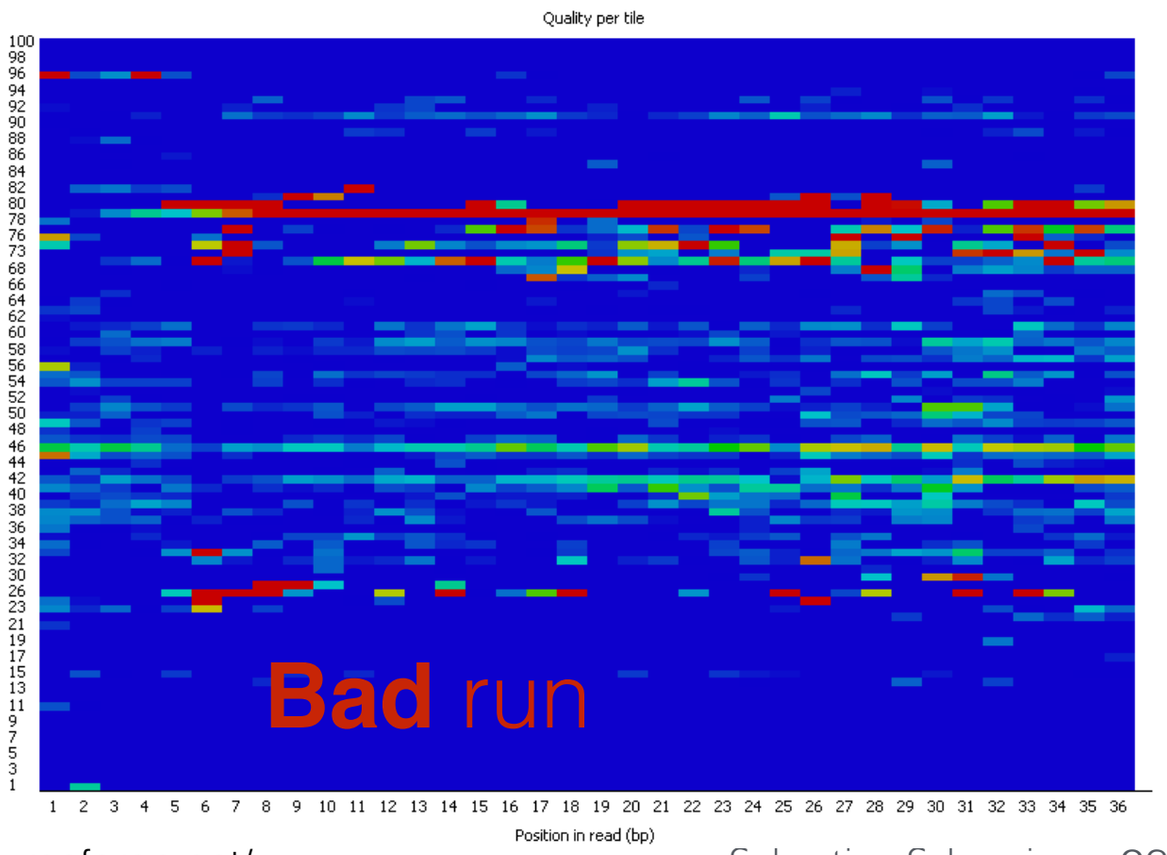


http://solexaqa.sourceforge.net/

Assessing quality: Tiles (3)

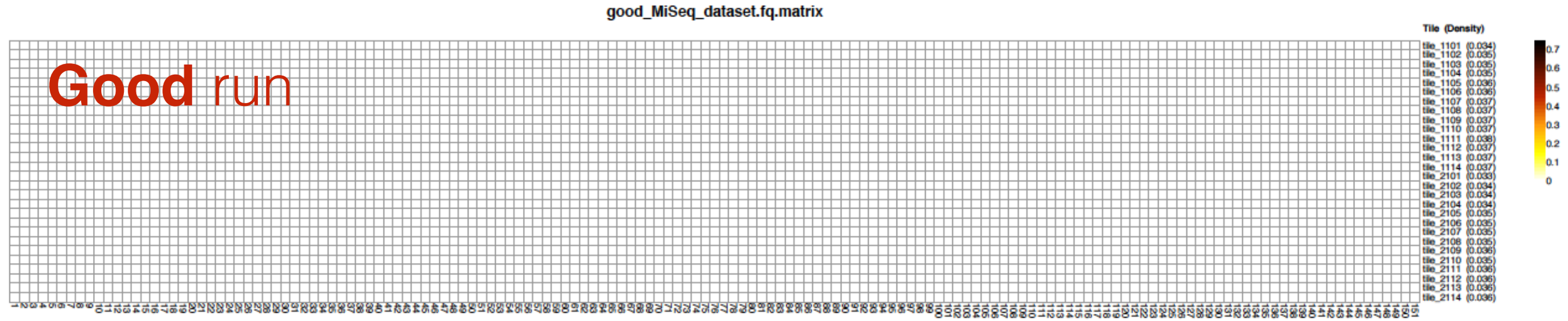
- The plot shows the deviation from the average quality for each tile.
- The colours are on a cold to hot scale
- Cold colours being positions where the quality was at or below the average for that base in the run
- Hotter colours indicate that a tile had worse qualities than other tiles for that base.

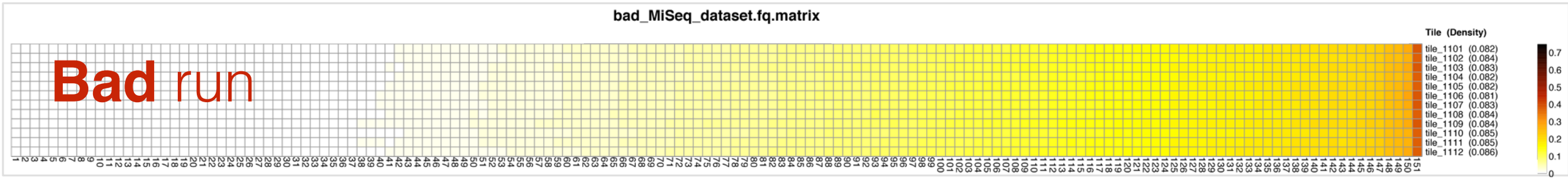




http://solexaqa.sourceforge.net/

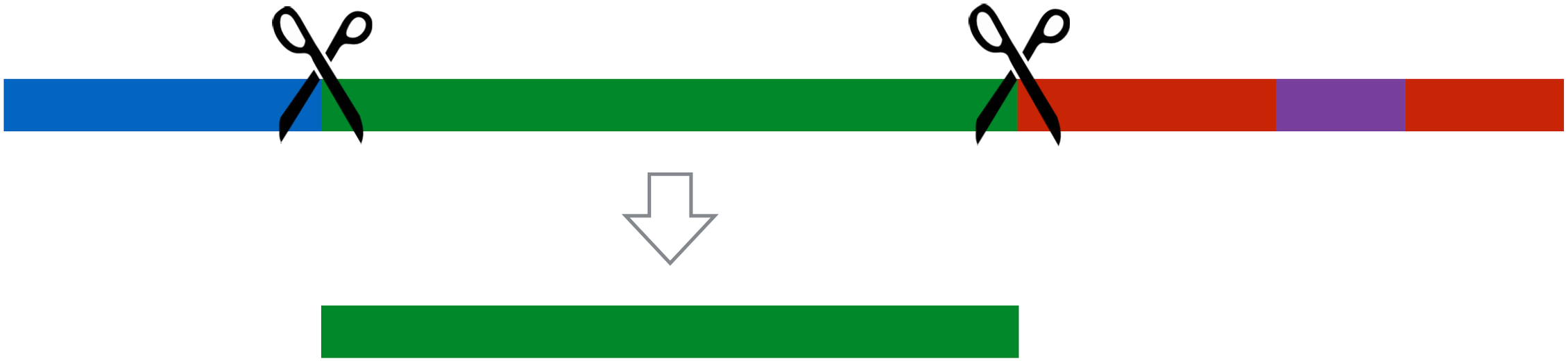
Assessing quality: Tiles (4)





Assessing quality: Data processing

- **Adapter trimming**: If not already done, we can remove the adapter used for sequencing.



Universal adapter
DNA sequence of interest
Indexed adapter
6 base index region

Example for Illumina TrueSeq

Assessing quality: Data processing (2)

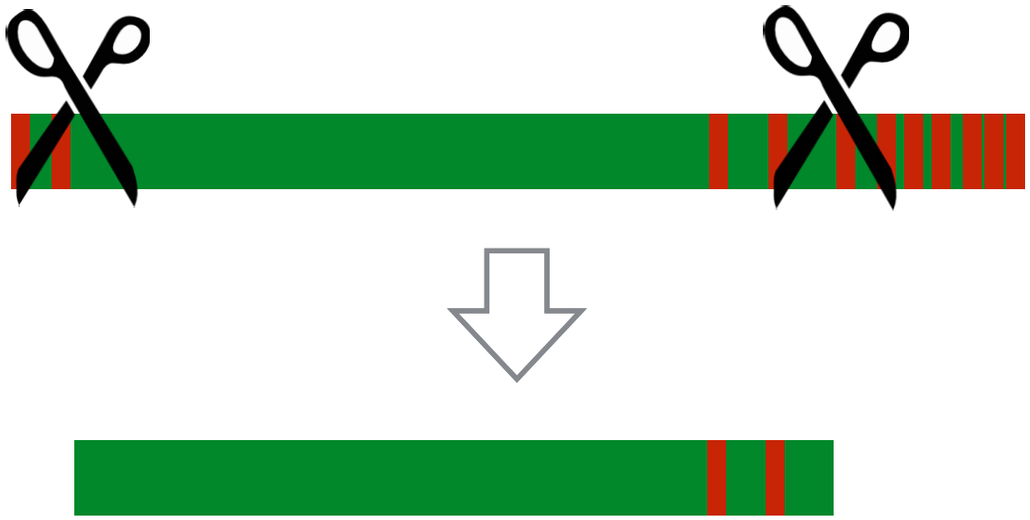
- **Filtering**: We can remove all reads that do not have a particular quality over the read length, e.*g. at least q20 for 80% of the read*

good quality

bad quality

Assessing quality: Data processing (3)

- **Cropping**: We can try to remove all *nt* from both ends that do not fulfil a certain quality

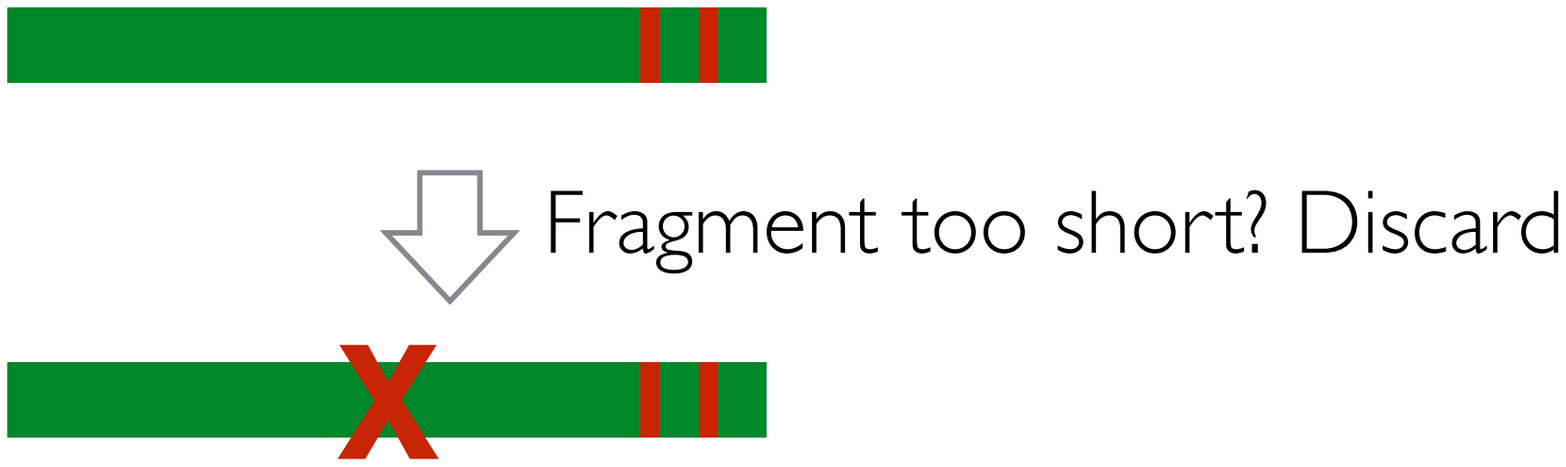


good quality

bad quality

Assessing quality: Data processing (4)

- **Removal:** We can remove reads that are too short after cropping.



good quality

bad quality

Assessing quality: Data processing (5)

- **Adapter trimming**: If not already done, we can remove the adapter used for sequencing
- **Filtering**: We can remove all reads that do not have a particular quality over the read length, e.*g. at least q20 for 80% of the read*
- **Cropping**: We can try to remove all *nt* from both ends that do not fulfil a certain quality
- **Removal:** We can remove reads that are too short after cropping.

In the end, we work with an adjusted set of sequencing reads for which we are more certain that they represent correct nt sequences from the genome

=> However filtering/trimming does not always improve things as we loose information

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

References

The role of replicates for error mitigation in next-generation sequencing. Robasky et al. Nature Reviews Genetics, 2014, 15, 56-62

Addressing challenges in the production and analysis of illumina sequencing data. Kirchner et al. BMC Genomics, 2011,12:382

Sebastian Schmeier
s.schmeier@massey.ac.nz
http://sschmeier.com/bioinf-workshop/

THE ENGINE
OF THE NEW
NEW ZEALAND