UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

WORKSHOP SIB 2025

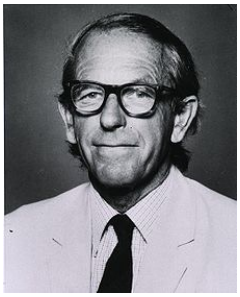
NEXT GENERATION SEQUENCING

Laurent Falquet

Swiss Institute of Bioinformatics

Contents

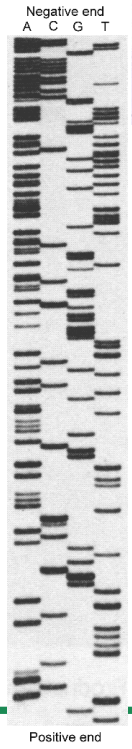
DNA Sequencing: Frederick Sanger (1918-2013)



1977 First genome sequenced: phage ϕX174…
…with the concept of shotgun sequencing!

Other method published the same year by Allan Maxam and Walter Gilbert.

F. Sanger also developed protein sequencing in 1955, he was awarded 2 Nobel prizes, one for each of the 2 methods!



Reminder of the sequencing methods

1986 - 2001 | 2005 | 2006 | 2010 | 2011 | 2014 | 2018

ABI Sanger | Roche 454 | Illumina | Life SOLiD | Life Ion Torrent | Pacific Biosciences | Oxford Nanopore

First Generation

Second Generation

Third Generation

Reminder of the sequencing methods

2014 2018 2019 2022 2024 2025?

MGI Illumina

Oxford Nanopore

Pacific Biosciences

Second Generation

Third Generation

Illumina family of machines in 2025

MiniSeq

MAX OUTPUT 8 Gb
MAX READ NUMBER 25 million
MAX READ LENGTH 2x150 bp

$50K

MiSeq

MAX OUTPUT 15 Gb
MAX READ NUMBER 25 million
MAX READ LENGTH 2x300 bp

$100K

NextSeq500

MAX OUTPUT 120 Gb
MAX READ NUMBER 400 million
MAX READ LENGTH 2x150 bp

$250K

NextSeq2000

300 Gb
1 billion
2x150 bp

$335K

NovaSeq 6000

MAX OUTPUT 6000 Gb
MAX READ NUMBER 20 billion
MAX READ LENGTH 2x150 bp

$1 Million

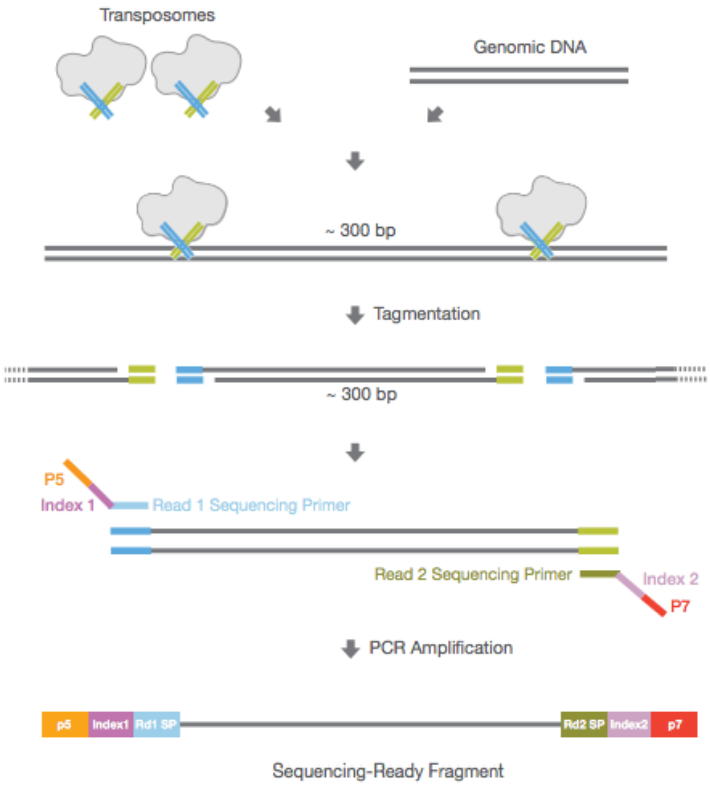
1.2 Gb
4 million
2x150 bp

$20K

MiSeq i100

MAX OUTPUT 30 Gb
MAX READ NUMBER 100 million
MAX READ LENGTH 2x300 bp

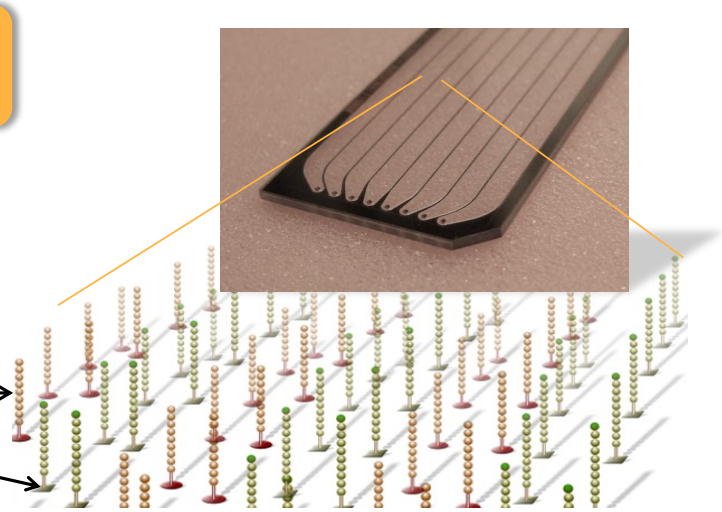
Illumina library preparation: tagmentation and barcoding

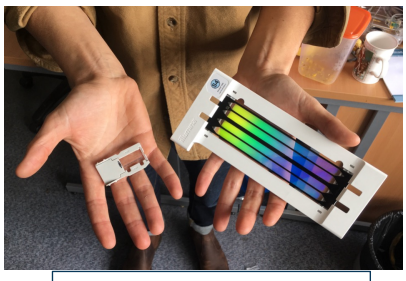


What is a Flow Cell?

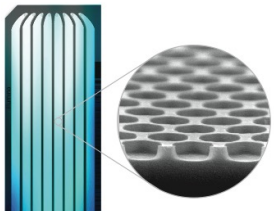
A flow cell is a thick glass slide with several channels or lanes

Each lane is randomly coated with a lawn of oligos that are complementary to library adapters or patterned



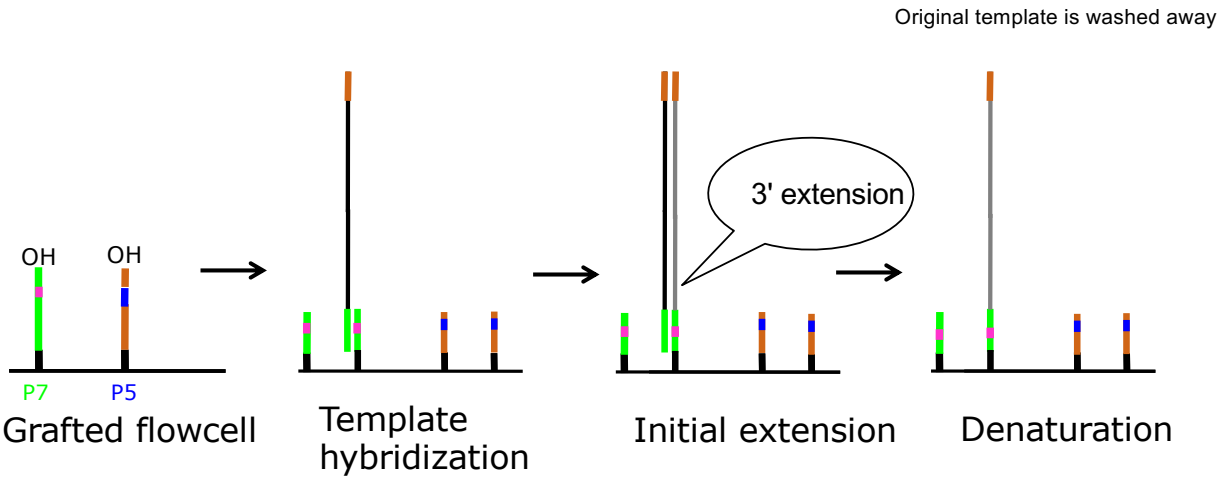


MiSeq vs NovaSeq

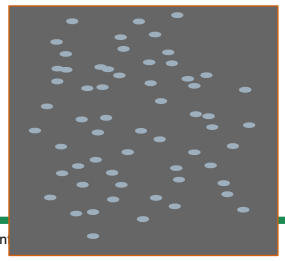


Patterned Flow Cells: higher number of clusters

Cluster Generation: Template hybridization and Initial Extension

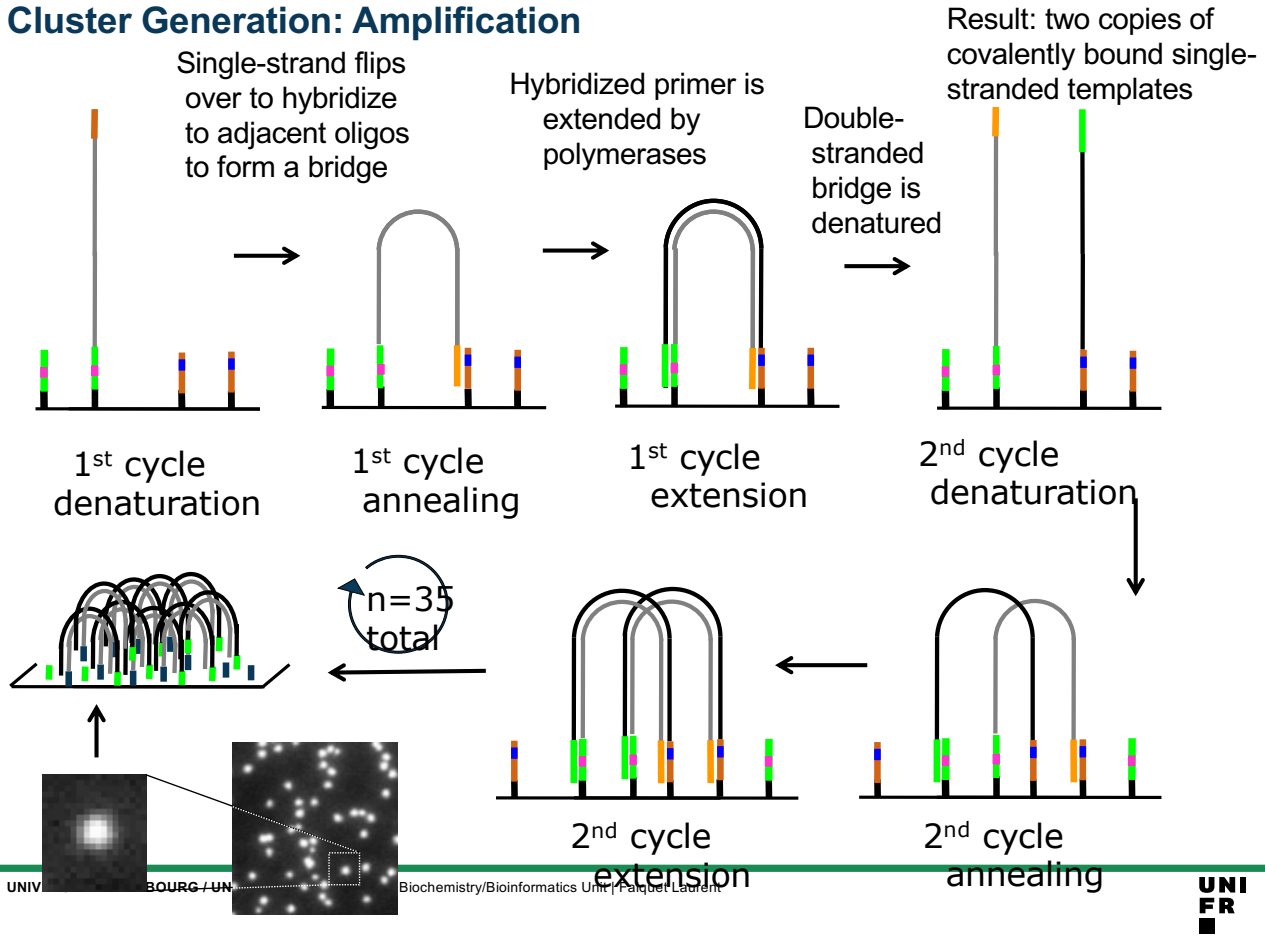


- >200 million single molecules hybridize to the lawn of primers



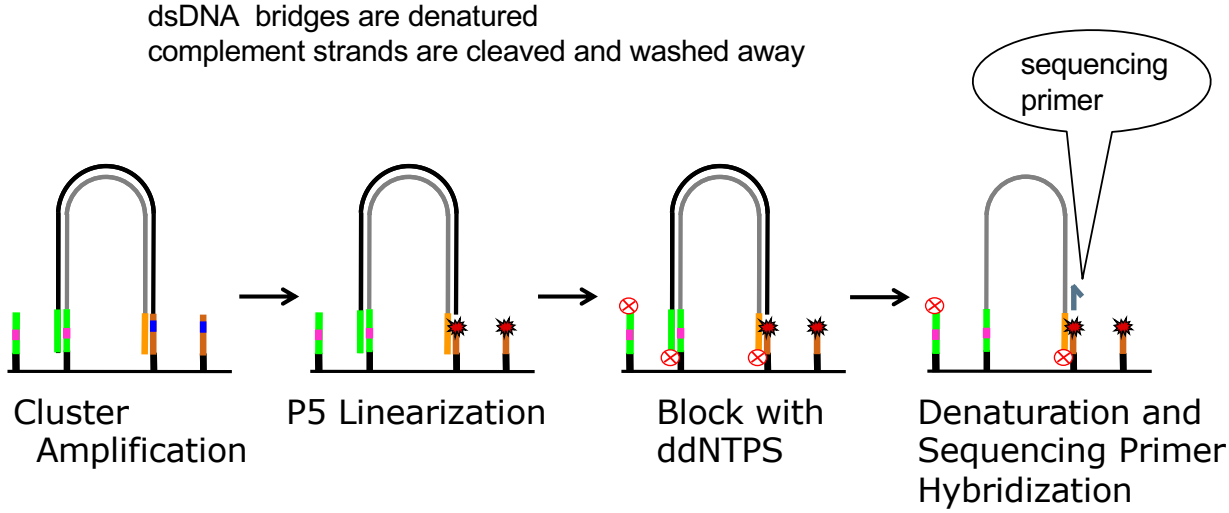
single molecules bound to flow cell in a random pattern

Cluster Generation: Amplification



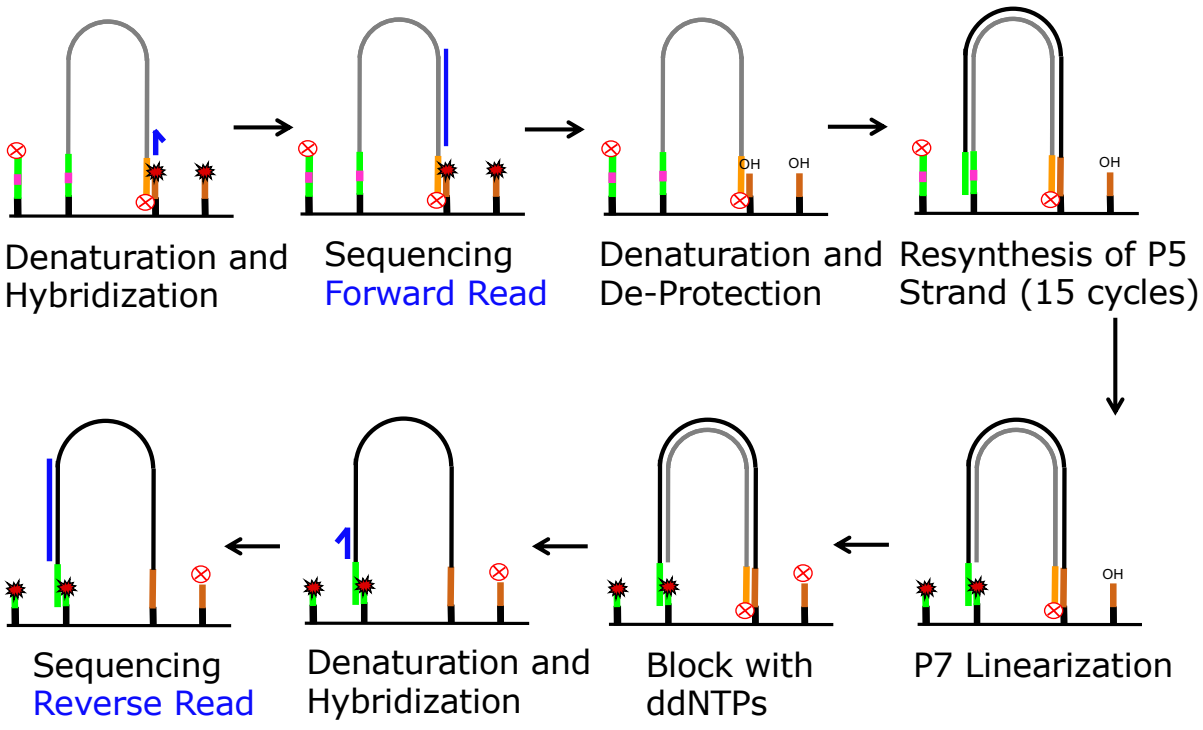
Cluster Generation: Linearization, Blocking and sequencing primer hybridization

dsDNA bridges are denatured
complement strands are cleaved and washed away

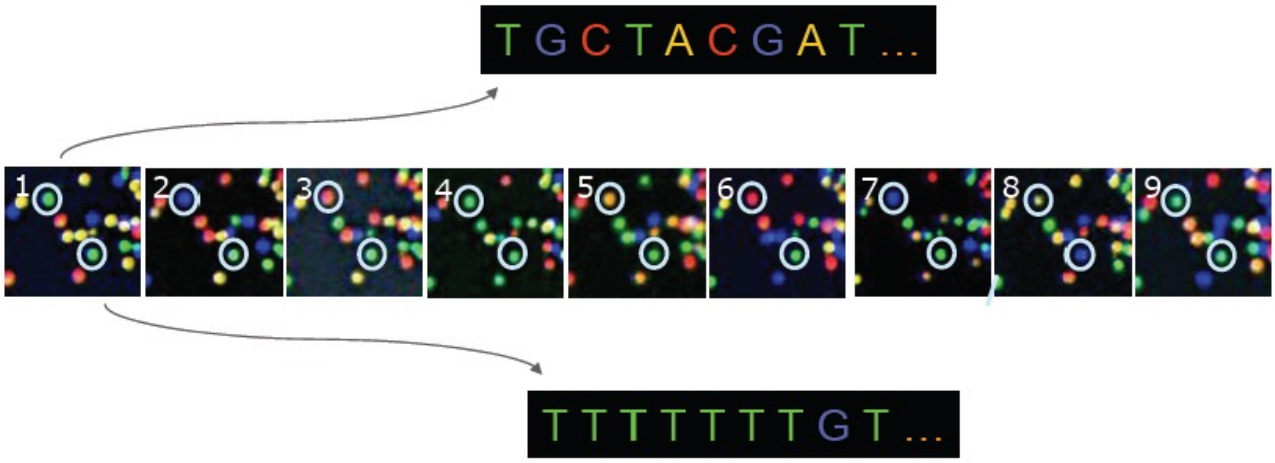


Free 3' ends are blocked to
prevent unwanted DNA priming

Sequencing forward and reverse reads



Base calling from raw data

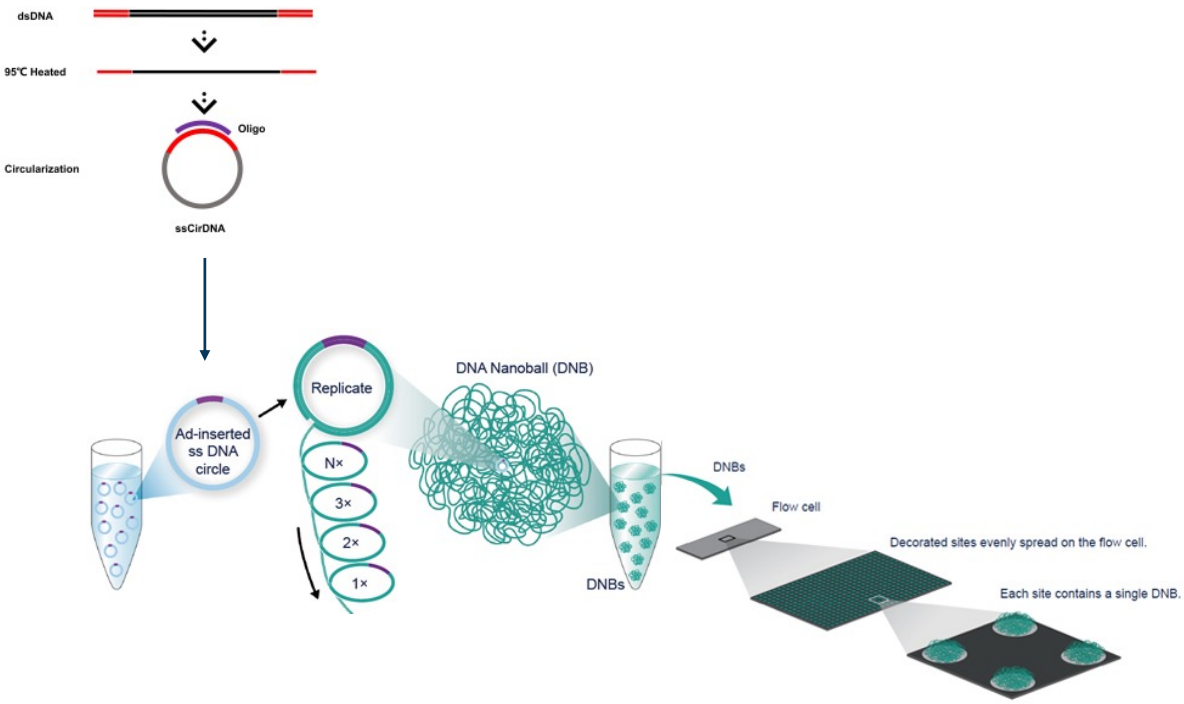


The identity of each base of a cluster is read off from sequential images

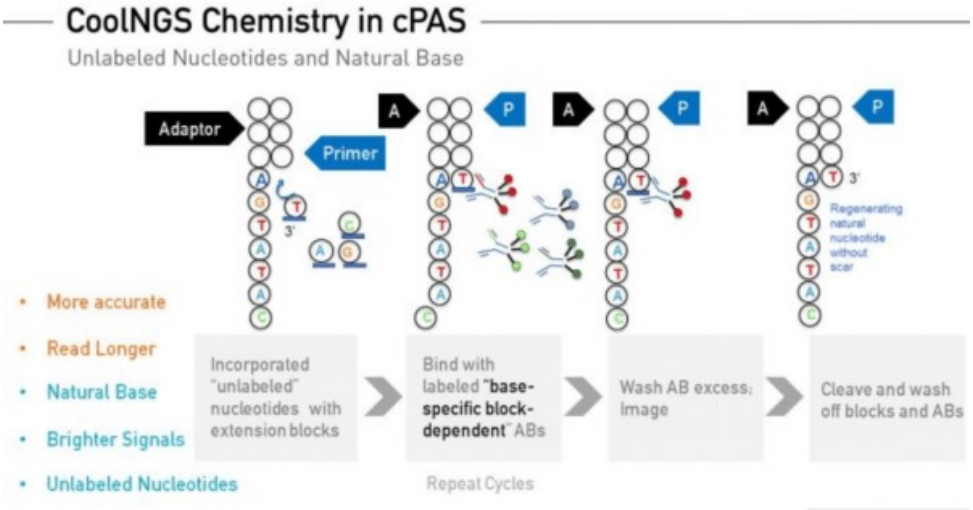
The MGI family of instruments and T7,T10,T20 machines

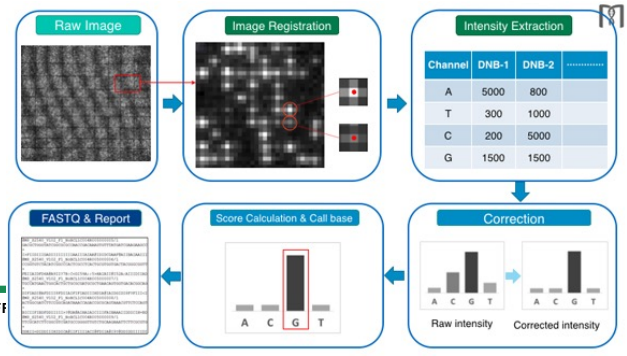
	Sequencers	Sequencers	Sequencers	Sequencers	Sequencers	Sequencers	Sequencers	Sequencers
Product Model	DNBSEQ-T20×2	DNBSEQ-T10×4RS	DNBSEQ-T7	DNBSEQ-T7* For HotMPS Only	DNBSEQ-G400	DNBSEQ-G400* For HotMPS Only	DNBSEQ-G50	DNBSEQ-G99
Features	Ultra-high Throughput	Whole Process Customization	Ultra-high Daily Throughput	Ultra-high Daily Throughput	Flexible Throughput, Diverse Reads	Flexible Throughput, Diverse Reads	Flexible Data Output	Fast
Applications	Ultra-high-depth Whole Genome Sequencing	Ultra-high-depth Whole Genome Sequencing	Deep Whole Genome Sequencing	Deep Whole Genome Sequencing	WGS, WES, Transcriptome sequencing, etc.	WGS, WES, Transcriptome sequencing, etc.	Small whole genome sequencing, targeted panels, low-pass whole genome sequencing	Targeted oncology panel sequencing, infectious disease sequencing, oncology methylation sequencing
Max. Flow Cell Run	6	8	4	4	2	2	1	2
Flow Cell Type	Slide	Slide	FC	FC	FCS & FCL	FCL	FCS & FCL	FC
Lane/Flow Cell++	1 Lane	1 Lane	1 Lane	1 Lane	4 or 2 Lanes	4 Lanes	1 Lane	1 Lane
Operation Mod	Ultra-high throughput	Ultra-high throughput	High and Ultra-high Throughput	High and Ultra-high Throughput	Medium and High Throughput	Medium and High Throughput	Medium Throughput	Low and Medium Throughput
Max. Throughput / Run	~72 Tb	76.8 Tb	6Tb	4 Tb	1440 Gb	720 Gb	150 Gb	48 Gb
Effective Reads / Flow Cell	35 B (PE100)	32 B (PE150)	5000M	5000M	300M, 550M, 1500-1800M	1500-1800M	100M, 500M	80M
Average run time	60-80 hours	96~106 hours	~24 hours	20~22 hours	CS: 13~37 hours FCL: 14~109 hours	15.5-50.5 hours	9~40 hours	5~12 hours
Min. Read Length	PE100	PE100	PE100	PE100	SE50	SE50	SE50	SE100
Max. Read Length	PE150	PE150	PE150	PE100	PE300	PE100	PE150	PE150

MGI comparable to Illumina, uses DNA nanoballs technology

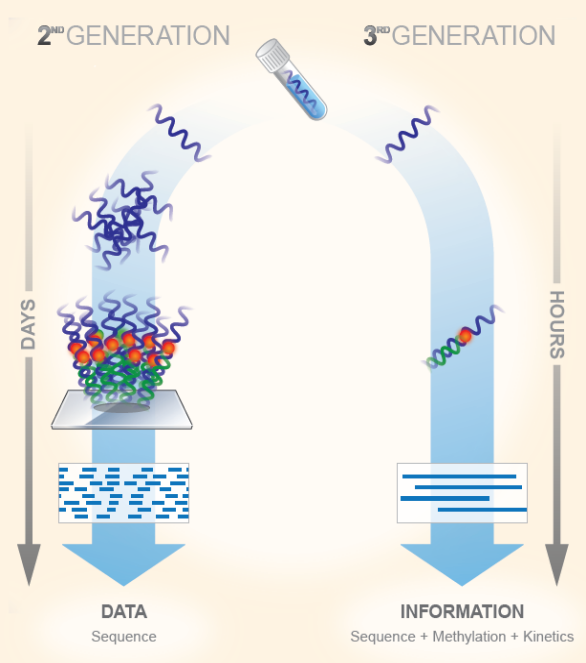


MGI coolNGS detection to avoid Illumina patents…

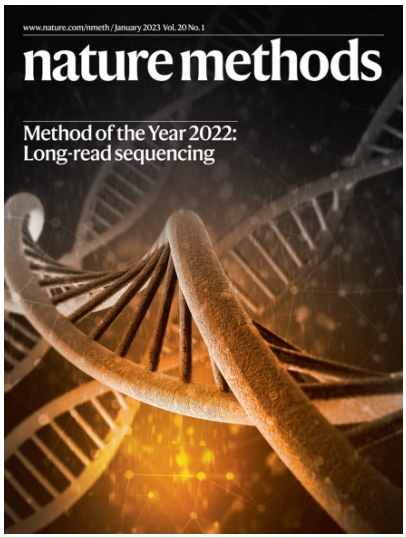




The Third Generation Sequencing Platforms: PacBio and Oxford Nanopore



Lower throughput, but longer reads and additional information (base modifications)



Pacific Biosciences

2010 First release PacBio RS
Single Molecule Real Time (SMRTcells)
Mean read length (~12-15Kbp)

2015 PacBio Sequel system
Increased throughput (7X)

2019 PacBio Sequel II system
Increased throughput (8X)
Mean length greater than 40Kbp

2022 PacBio Revio system
Increased throughput (15X) = 1’300 human genomes / year

2024 PacBio Vega system = small Revio (one SMRT cell vs 4)
New SPRQ chemistry (2X)

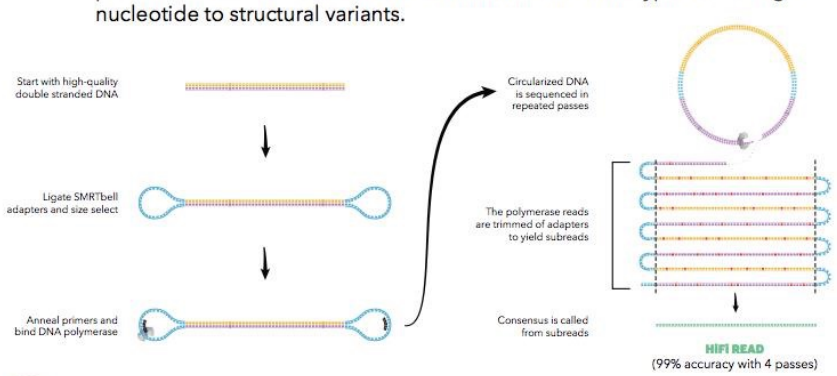
HiFi reads

PacBio

HIFI READS

Generate Highly Accurate Long Reads

Produce HiFi reads using the circular consensus sequencing (CCS) mode to provide base-level resolution for detection of all variant types from single nucleotide to structural variants.



10-20 KB

LONG READS

Optimize Your Run for Even Longer Reads

Sequence read lengths in the tens of kilobases using the continuous long read (CLR) sequencing mode to enable high-quality assembly of even the most complex genomes.

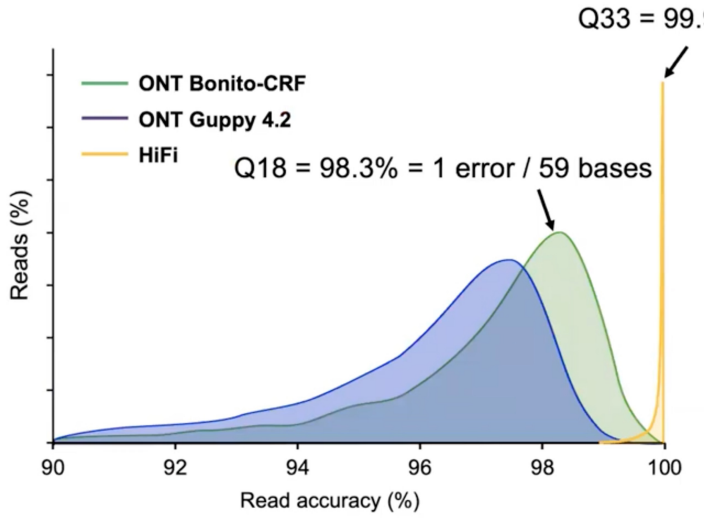
Half of Data in Reads	Longest Reads Up To
>50 kb	**175 kb**

HiFi reads compared to ONT

PacBio

HiFi READS

PacBio HiFi reads are much more accurate than ONT reads



HiFi read accuracy far exceeds the best reported from ONT basecallers.

HiFi: HG003 18 kb library, Sequel II System Chemistry 2.0, precisionFDA Truth Challenge V2
ONT: Bonito-CRF & Guppy 4.2 NCM Nanopore Tech Update Dec. 2020

Oxford Nanopore family of machines

MinION (available $1999 for 2 runs)

Minion

MinION

GridIONx5 ($67K)

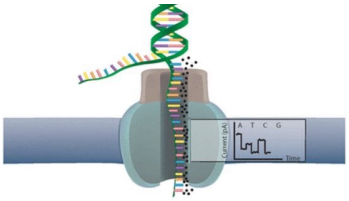
PromethION ($436K)

SmidgION (not yet)

First products announced in Jan 2012:

- **MinION – USB disposable sequencer has 512 nanopores (available 2014)**
- **GridIONx5 (available 2018)**
- **PromethION rack for 144'000 nanopores (available 2019)**
- **SmidgION (announced 2016), for mobile devices (in development)**

Oxford Nanopore

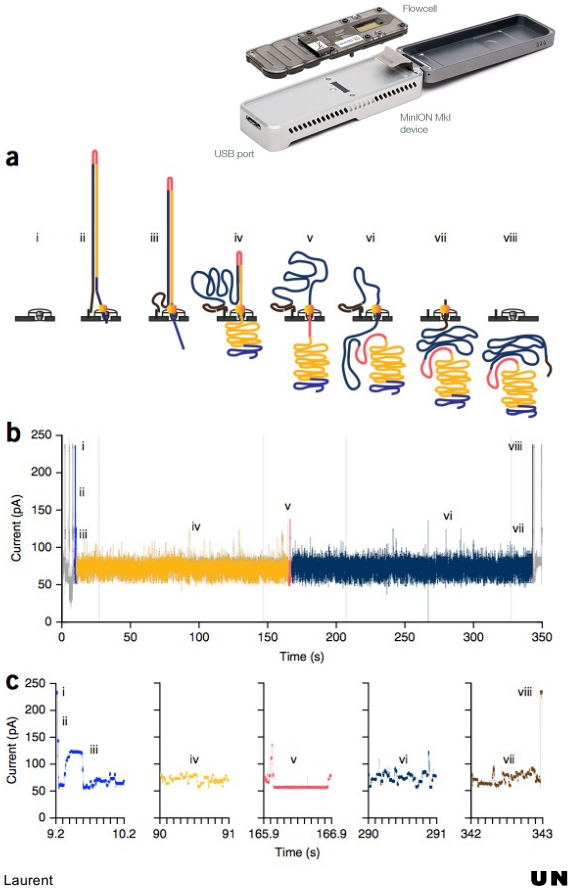


error rate still high!

indels 13%

substitution 5%

Recently improved <2% with R10.4



Short reads vs Long reads

Short reads	Long reads
Illumina or MGI	**PacBio (HiFi) or Nanopore**
Highly flexible usage	High quality genome assembly
	DNA base modifications detection
High quality reads of identical length	High quality HiFi reads (PacBio)
Enormous throughput, low price per base	Cheap apparatus (Nanopore)
Limitations	**Limitations**
Short (max 300bp PE)	Uneven length
Not suitable to solve repeats	High error rate in raw reads (mainly indels)

Advantages of long reads

High quality genome assembly

- structural variation (such as inversions or insertions)
- complex rearrangements
- repeat expansion length variations
- variants in repetitive or highly polymorphic regions
- telomere-2-telomere
- haplotigs separation or phasing

Other areas

- DNA base modifications detection
- transcripts isoforms identification

Sequencing advise

Best high throughput: Illumina NovaSeq (X), MGI

Small scale: Illumina MiSeq or MiniSeq or iSeq100 or Nanopore

Speed: Ion Torrent, Nanopore

Long reads and modifications: PacBio (HiFi for high quality)

Long reads: Nanopore

Multiplexing using barcoding allows reducing the costs per sample

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

METAGENOMICS QC & DATA CLEANING

Laurent Falquet
MA/MER @ UniFr
Group Leader @ SIB (Swiss Institute of Bioinformatics)

Swiss Institute of Bioinformatics

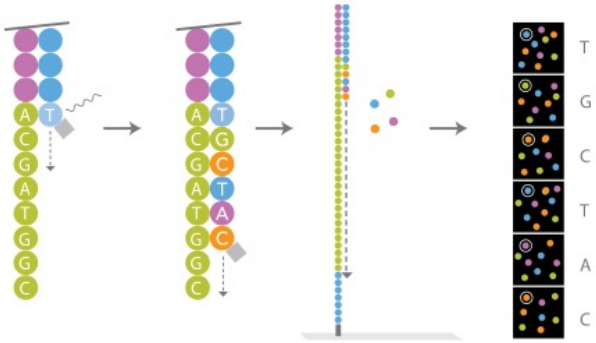
Quality Control and Cleaning of the raw reads

Important first step at the beginning of the analysis

What are reads?

The sequenced part of one DNA fragment.

These DNA sequences are given by the sequencing machine with a Phred quality score in so called FASTQ format




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

Phred quality score, a measure of base call quality

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

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect 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%

The quality score is ASCII encoded in the FASTQ format

FASTQ is a FASTA with score

Example of FASTA

```
>C3PO_0001:2:1:17:1499#0/1
TGAATTCATTGACCATAACAATCATATGCATGATGCAAATTATAATATCATTT
TTGTTTGAGCAAATGATTCATAATAATGTATTTCAATATTTTTAGGAATATCT
CCCAATATTGCGCGTGCTGAATTCCATCCGGAATTTTTGACGTCCCCCCCCGA
ANGGANGNGANNNNGNNGNNNTNTNNAAANGNNNNN
...
```

Example of FASTQ Illumina 1.8+

```
@M01867:115:000000000-ABF5V:1:1101:9268:1666 1:N:0:51
AACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGCGAACTGGTTGTTGGGTGCTTTTTG
+
--A-6@8CE,@<CEFGGFAFF9CEFF,C@CE@B<8@C:CC,,+,7@C<6,668C,,+8,6,,<9,+
@M01867:115:000000000-ABF5V:1:1101:9214:1685 1:N:0:51
AACCGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGTCTACTAGTTGTTGGTGGAGTAAAA
+
--AA@7:FF9C9C@FEFE<CF9FEFF,C@FE:B8,6C:+C6CFD9CE,<C6<C@,,8,,,,;,,,-
@M01867:115:000000000-ABF5V:1:1101:18344:1708 1:N:0:51
AACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCAACTAGCCGTTGGGAGCCTTGAG
+
--A99E8CE<C9CFFGGG8FF9@CFF9ECFF@F;,CFC7C,CF,,CF@@EE@@@,,+,,6BE@,,-
...
```

read 1

read 2

read 3

Quality Control of the data

First step after receiving the data

Sometimes already done by the sequencing center (e.g., chastity)

Objective:

- **Remove bad quality reads**
- **Remove contaminants**
- **Trim ends of reads**
- **Remove orphans (if possible or desirable)**
- **Correct errors**

FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/)

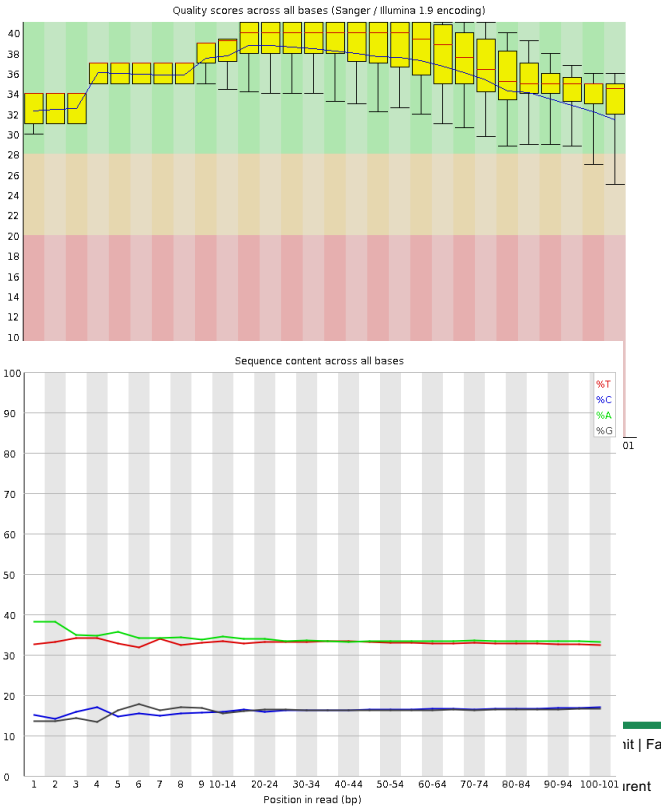
MultiQC (http://multiqc.info)

FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)

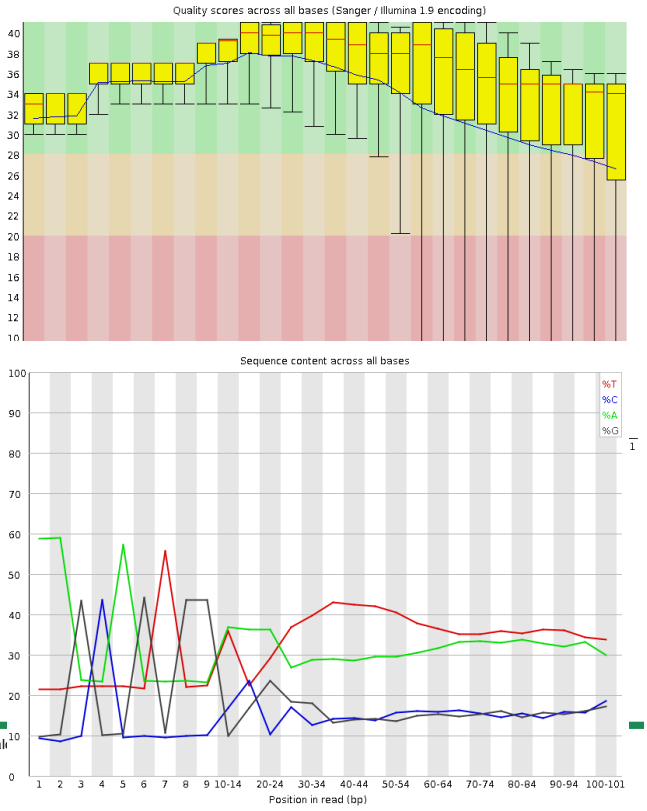
PrinSeq (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi)

Quality control examples

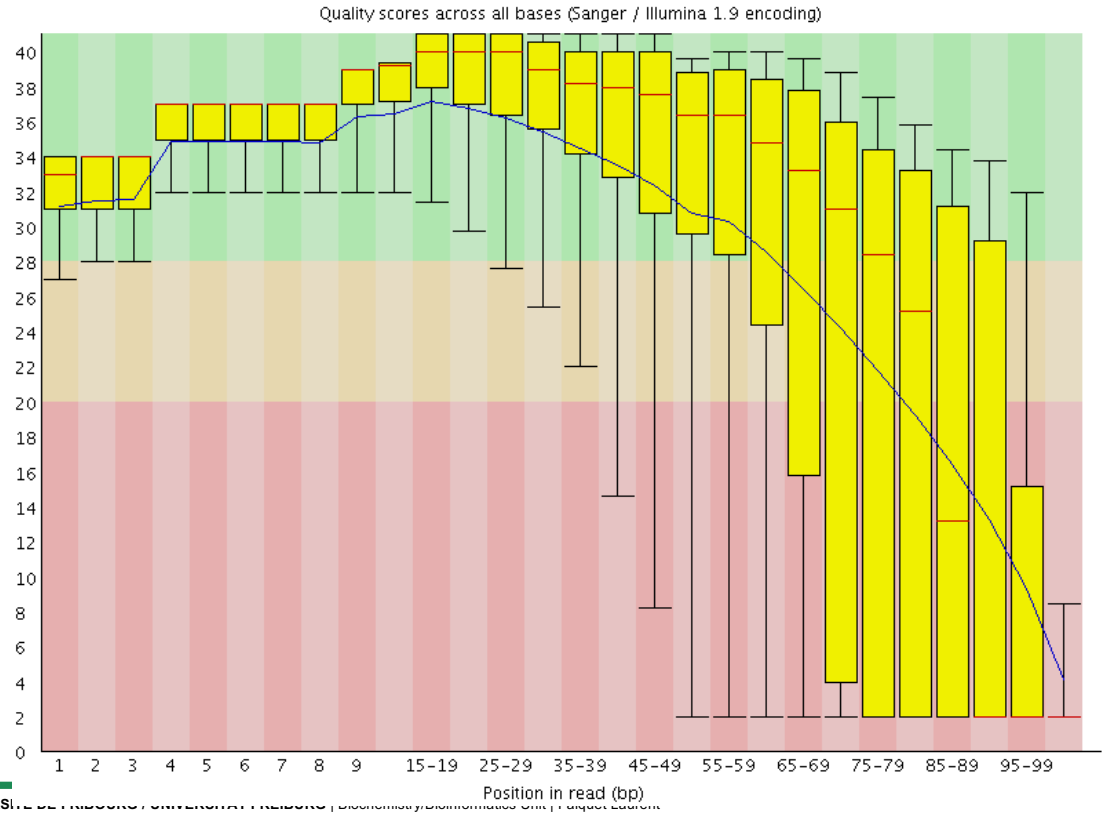
Forward



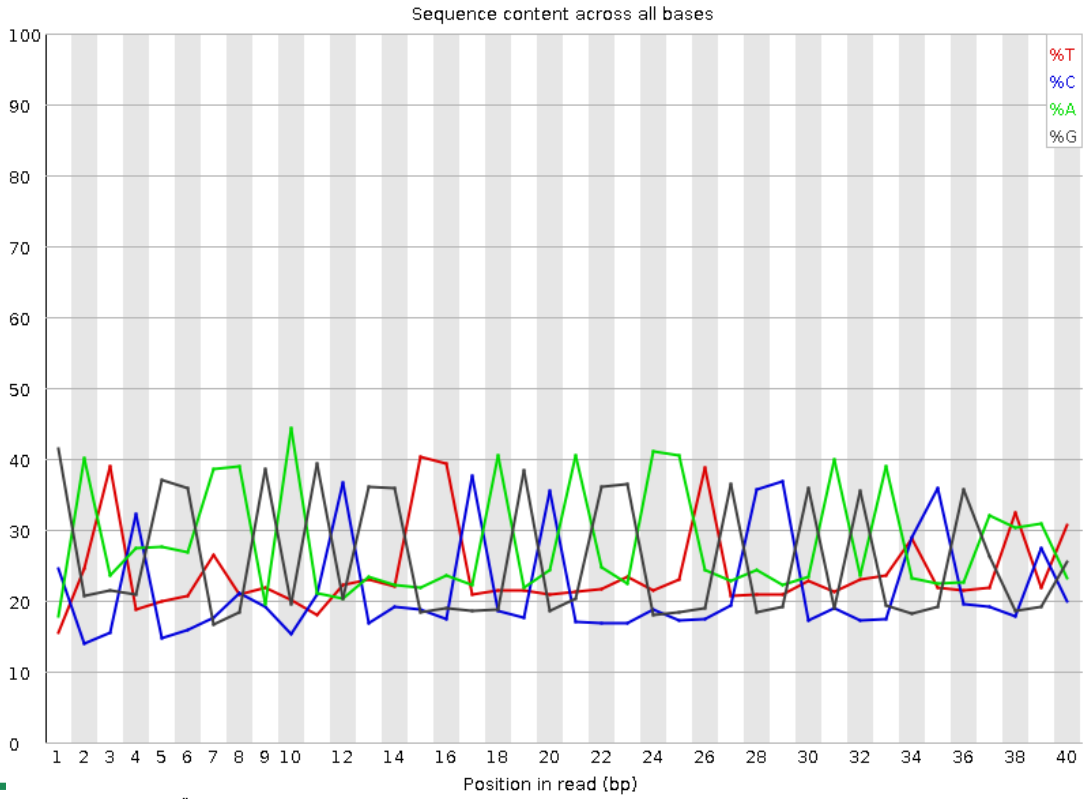
Reverse



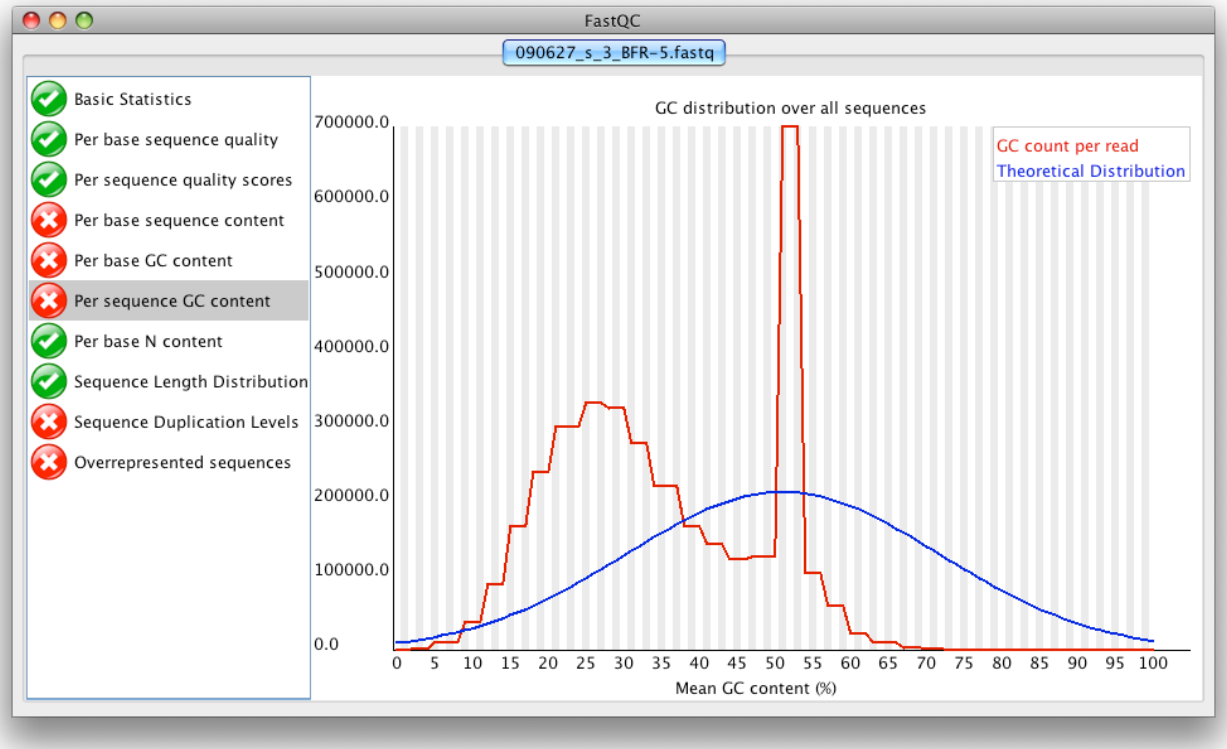
Quality Control example: bad



Quality Control example: very bad??



Quality Control example: contamination

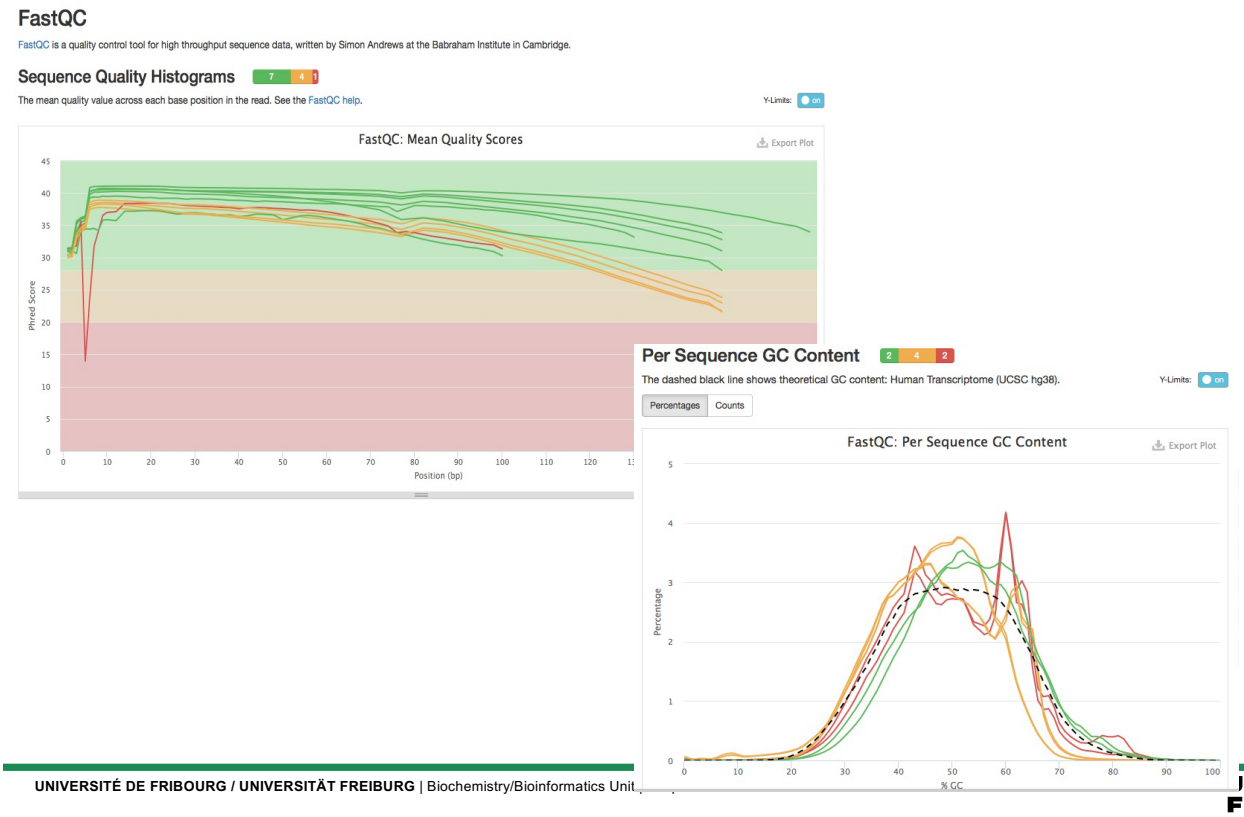


Quality Control example: contamination

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Use MultiQC to look at multiple QC results jointly



Read trimming or filtering

Trimming remove 5' and/or 3' ends of reads (bad quality or adapter)
Filtering remove full reads (e.g., contaminants)

Tools:

FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)
PrinSeq (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi)
Sickle (https://github.com/najoshi/sickle)
ea-utils (https://code.google.com/p/ea-utils/)
Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)
cutadapt (https://cutadapt.readthedocs.org/)
Fastp (https://github.com/OpenGene/fastp)

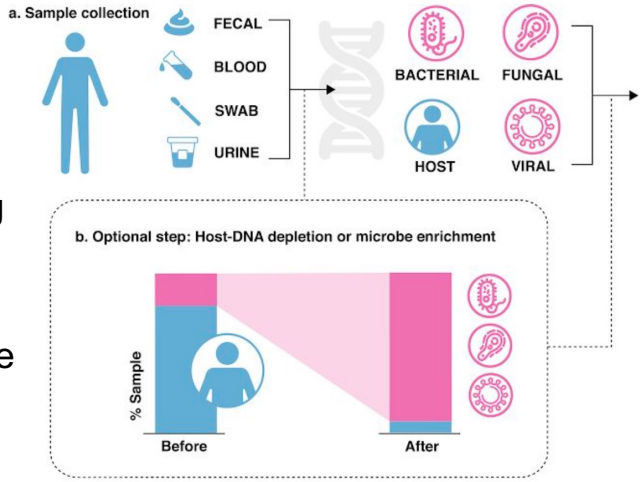
...

Host removal or depletion

Host DNA sequences are very likely to contaminate your metagenomics data. It is critical to remove those sequences either before sequencing or before analysis.

E.g., PCR amplified 16S RNA can be contaminated by mitochondria or chloroplast 16S of the host.

E.g., WGS can be contaminated by host genomic DNA.



https://doi.org/10.48550/arXiv.1911.11304

Summary of the topic

NGS methods can generate short or long reads

FASTQ is the universal format of reads (with also BAM for PacBio, FAST5 for ONT)

QC and cleaning is necessary

Host removal is needed for WGS

Analysing this type of data has many limits

Thank you for your attention. Questions?

UNIVERSITÉ DE FRIBOURG / UNIVERSITÄT FREIBURG | Biochemistry/Bioinformatics Unit | Falquet Laurent