Introduction to Next Generation Sequencing (NGS)

Anders K. Krabberød
Department of Biosciences/ Norwegian Sequencing center
University of Oslo
a.k.krabberod@ibv.uio.no

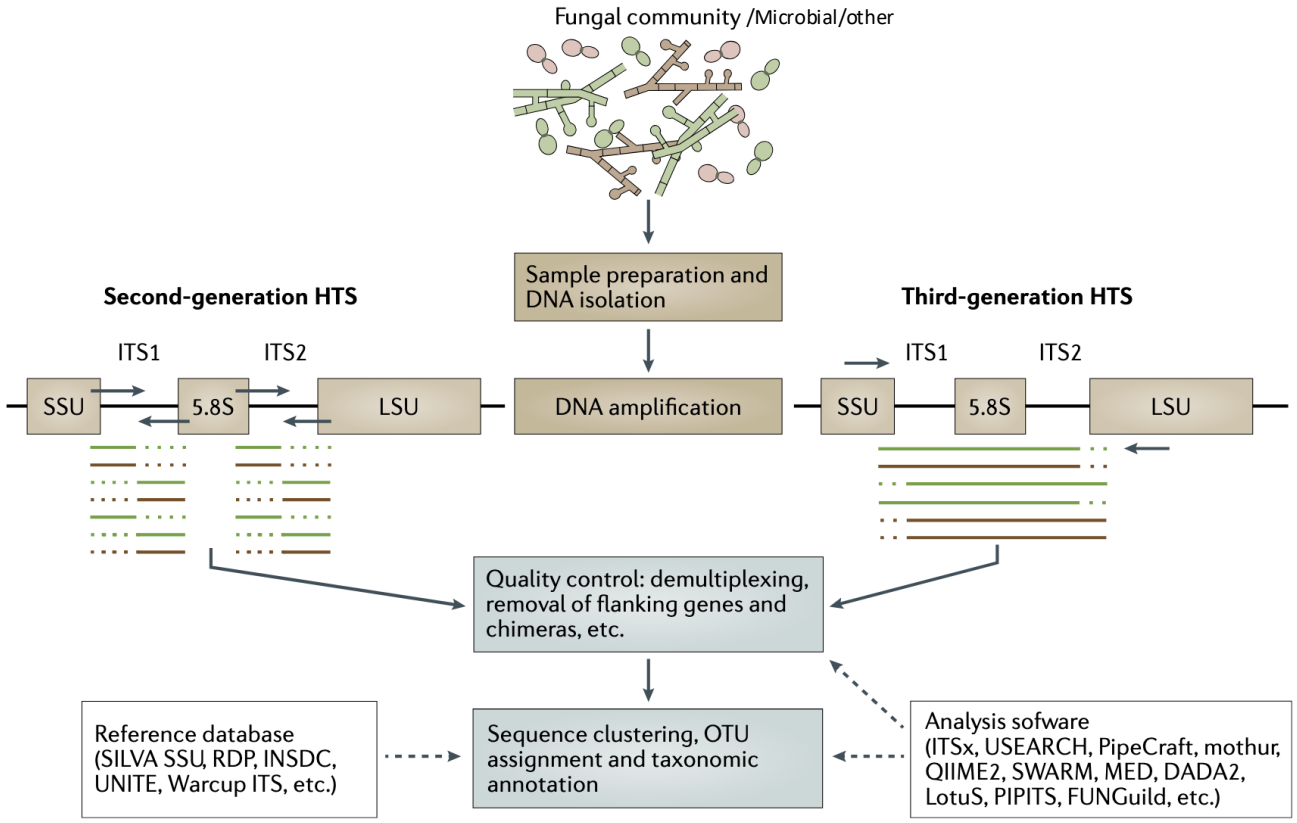
AB332 - 2024

UNIVERSITY OF OSLO

Some important terms

- **Metabarcoding** – The use of DNA markers to analyze biodiversity and species identification from environmental samples.
- **Amplicons** - DNA fragments generated through polymerase chain reaction (PCR)
- **OTUs (and ASVs)** - Groupings of closely related sequences used as proxies for species (Operational Taxonomic Unit, Amplicon Sequence Variants)
- **High-throughput Sequencing (HTS)** - A broad range of technologies that enable the rapid sequencing of large volumes of DNA or RNA
- **Next Generation Sequencing (NGS)** - A specific type of HTS technology that allows for massively parallel sequencing of DNA (or RNA)
- **Third-generation Sequencing** - A newer generation of sequencing technologies (e.g., PacBio, Oxford Nanopore) that read long, single molecules of DNA or RNA in real-time

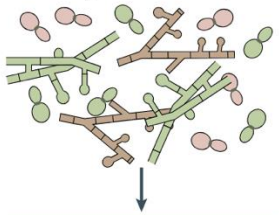
Metabarcoding



Nillson et al. 2019

Metabarcoding

Fungal community /Microbial/other



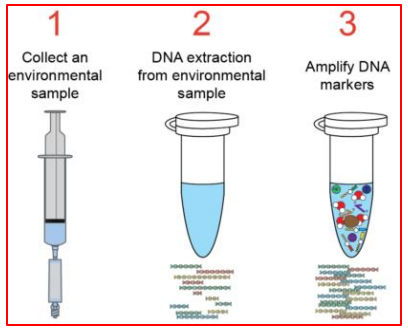
Sample preparation and DNA isolation

Second-generation HTS

Third-generation HTS

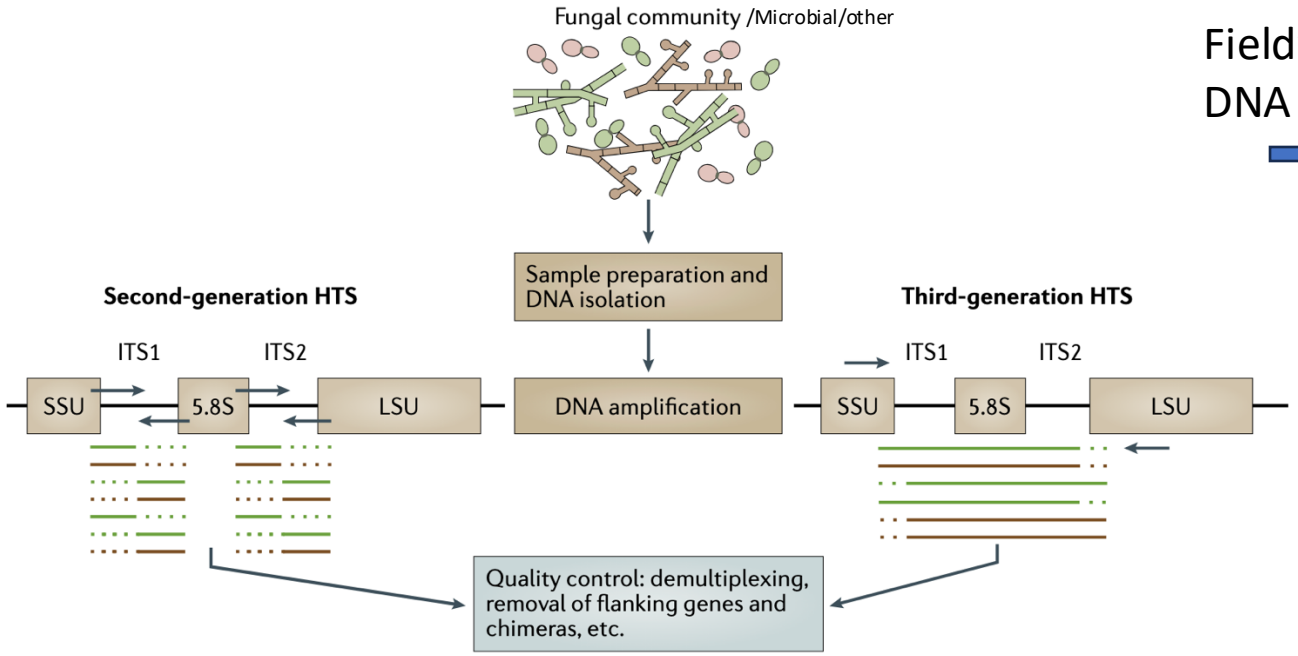
Fieldwork/labwork/
DNA extraction etc

Abdelfattah et al.,2018



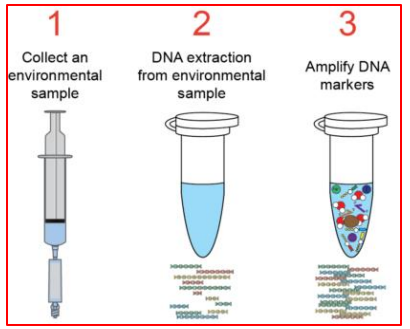
Nillson et al. 2019

Metabarcoding

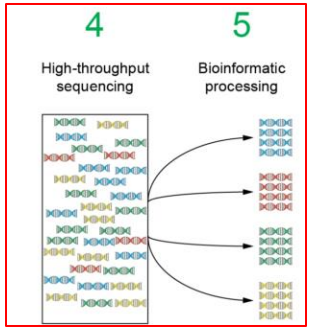


Fieldwork/labwork/
DNA extraction etc

Abdelfattah et al.,2018

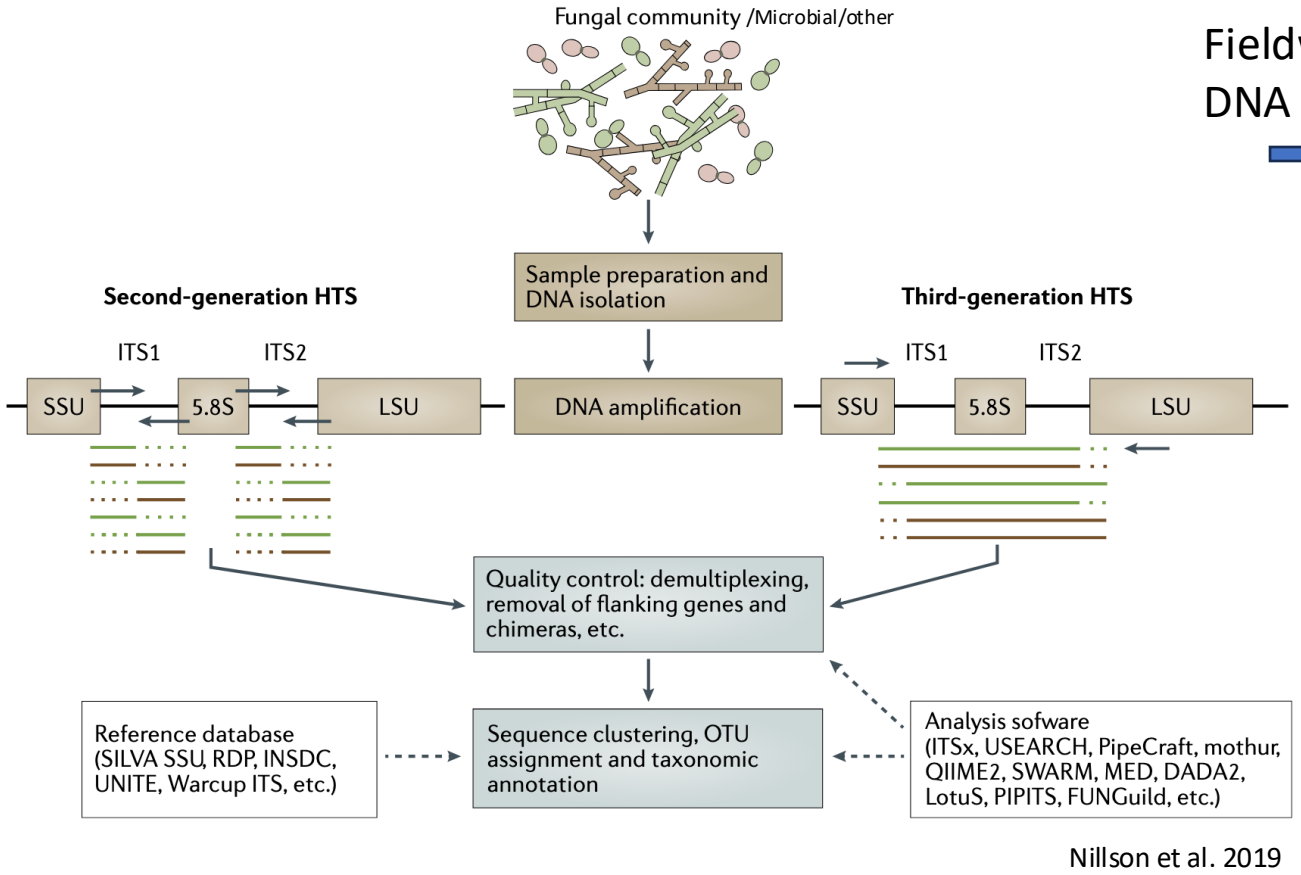


Sequencing /
Bioinformatics



Nillson et al. 2019

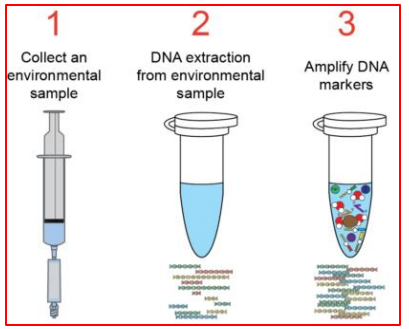
Metabarcoding



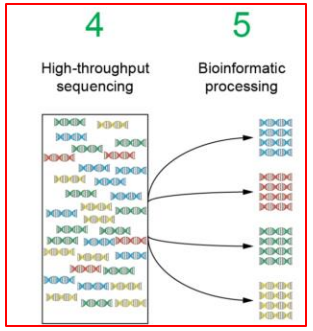
Nillson et al. 2019

Fieldwork/labwork/
DNA extraction etc

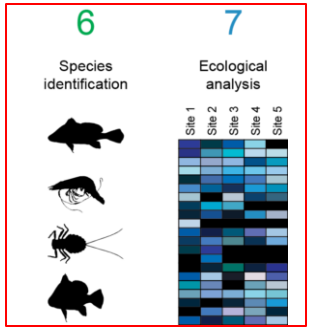
Abdelfattah et al.,2018



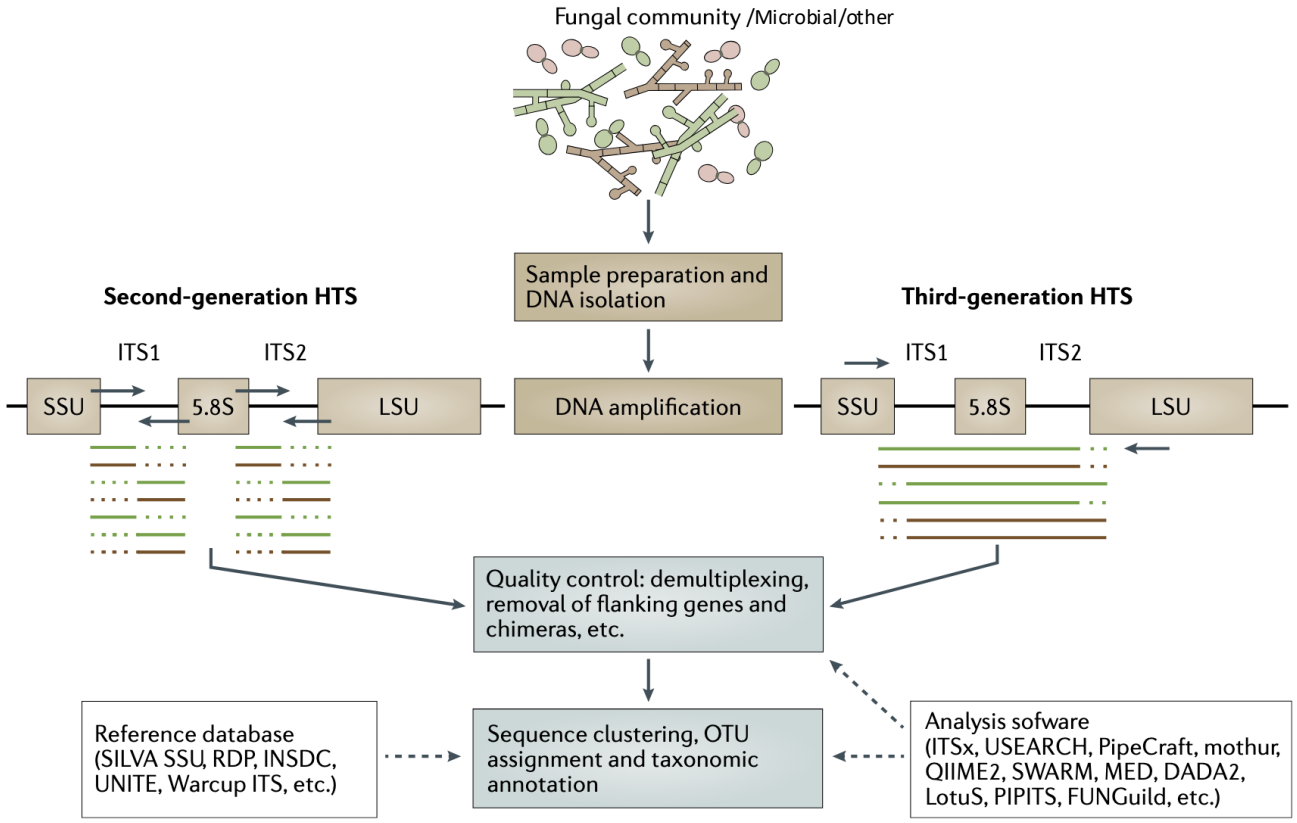
Sequencing /
Bioinformatics



Ecological
analysis

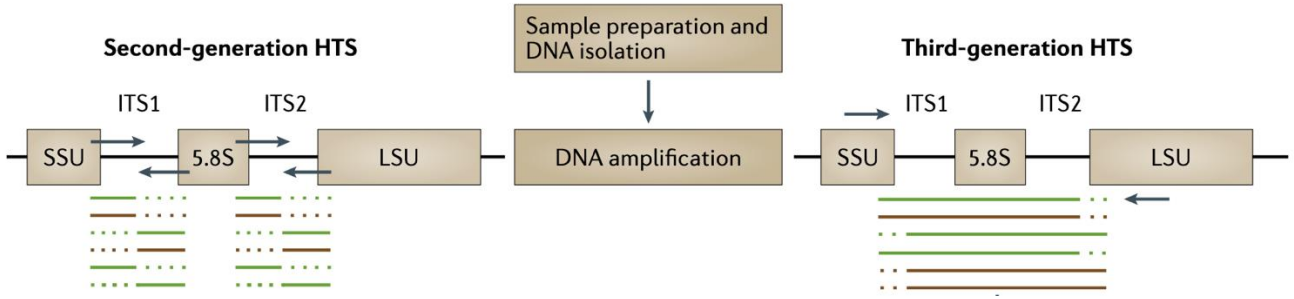


Metabarcoding

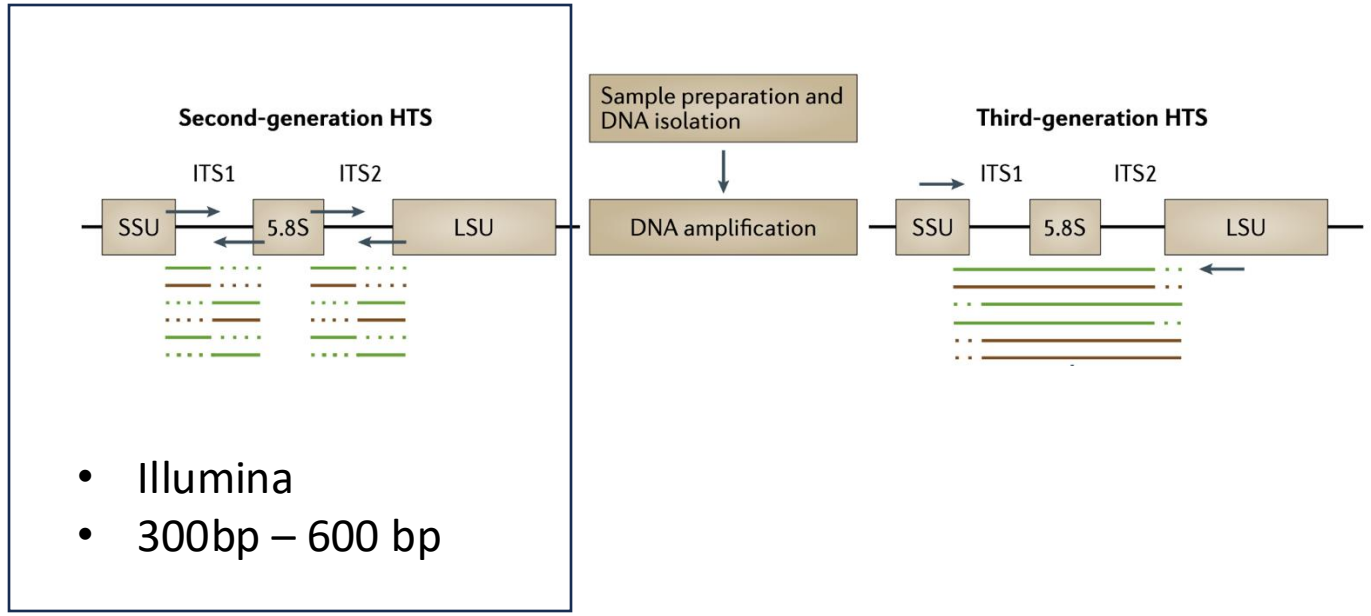


Nillson et al. 2019

Sequencing technology

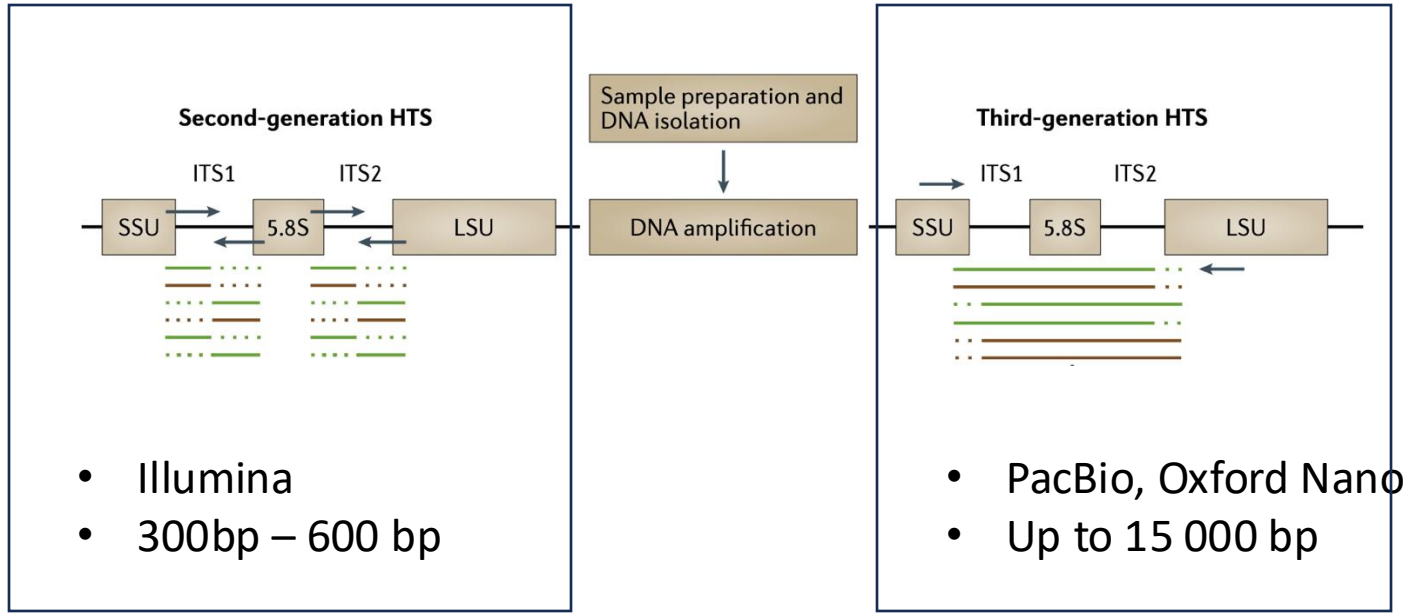


Sequencing technology



- Illumina
- 300bp – 600 bp

Sequencing technology

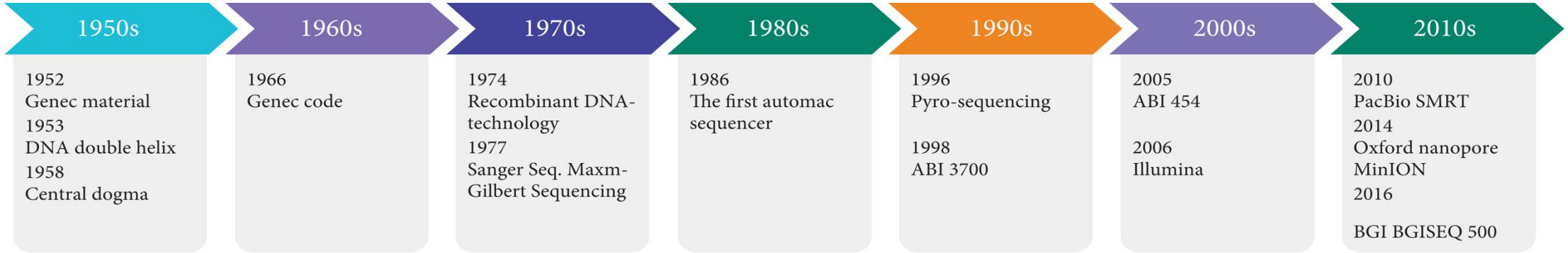


- Illumina
- 300bp – 600 bp

- PacBio, Oxford Nanopore
- Up to 15 000 bp

A short history of Sequencing

- Sequencing: determining the order of nucleotide bases in a string of DNA (or RNA)
- The development started in the 1950's after Watson and Crick described the structure of DNA and with the formulation of the Central dogma.
- The early methods were cumbersome (and dangerous) using radioactive material and adding individual nucleotides to a reaction one by one.
- The last few years the development has been phenomenal!



A short history of Sequencing

- F. Sanger et al. 1977
  - Short fragments
  - 15-200 nucleotides
  - Slooooow process

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

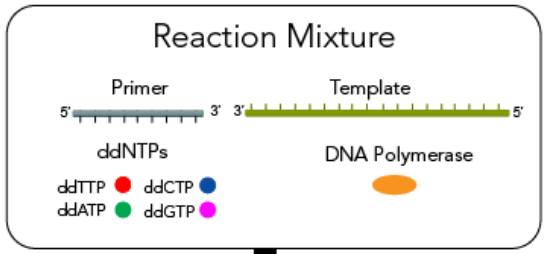
DNA sequencing with chain-terminating inhibitors

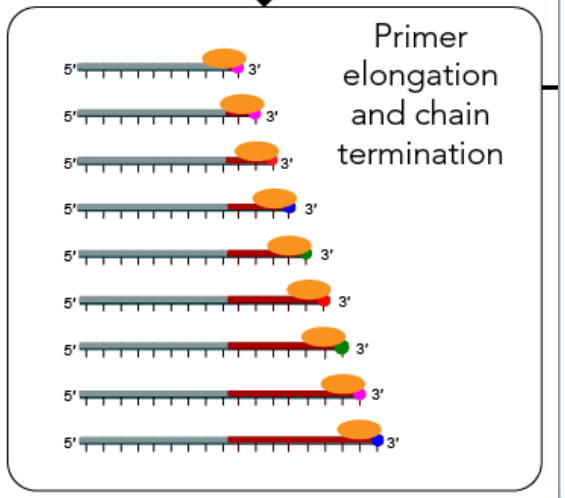
(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

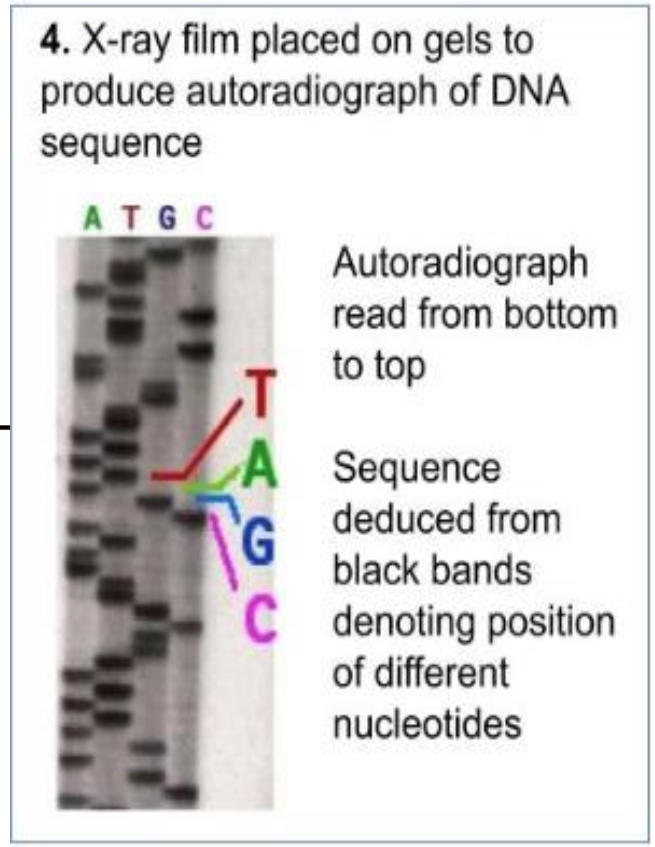
F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

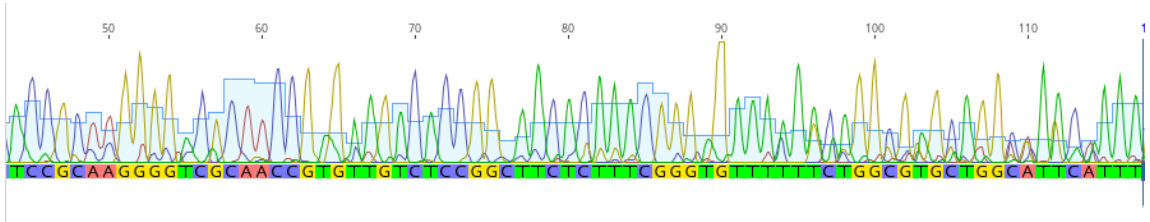


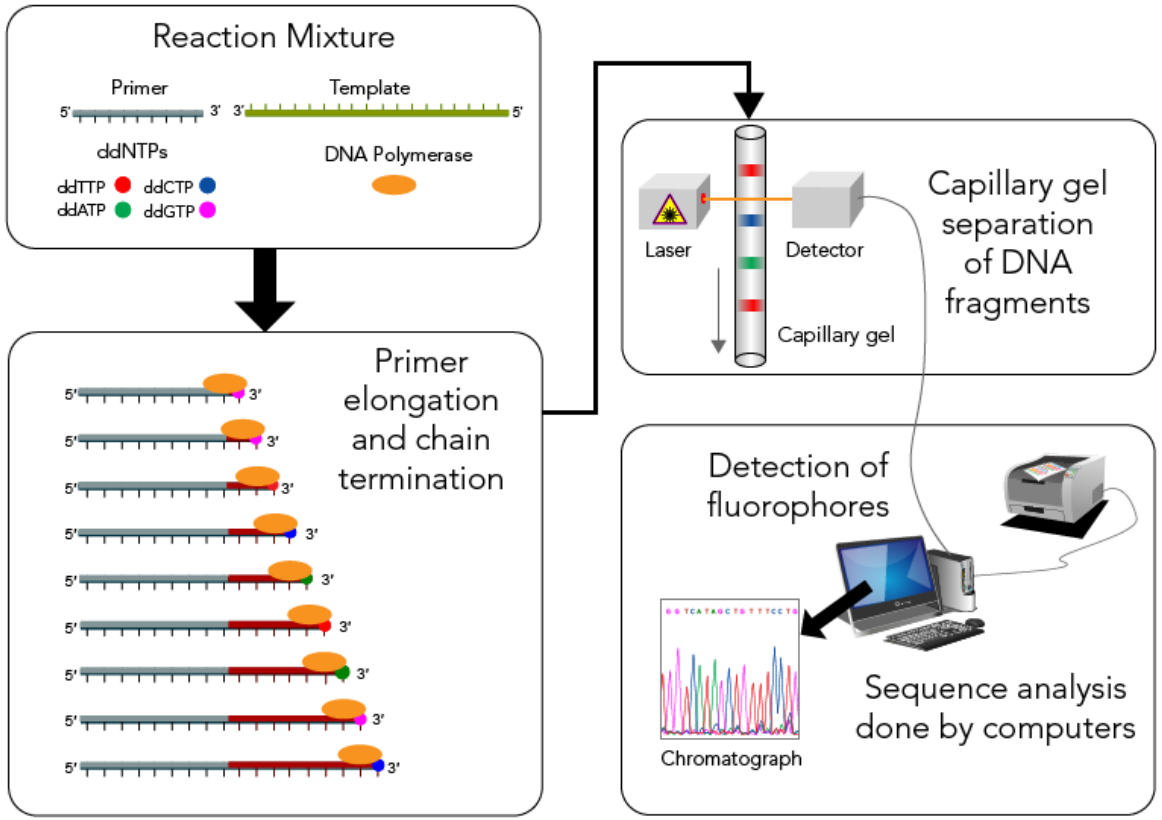




A short history of Sequencing

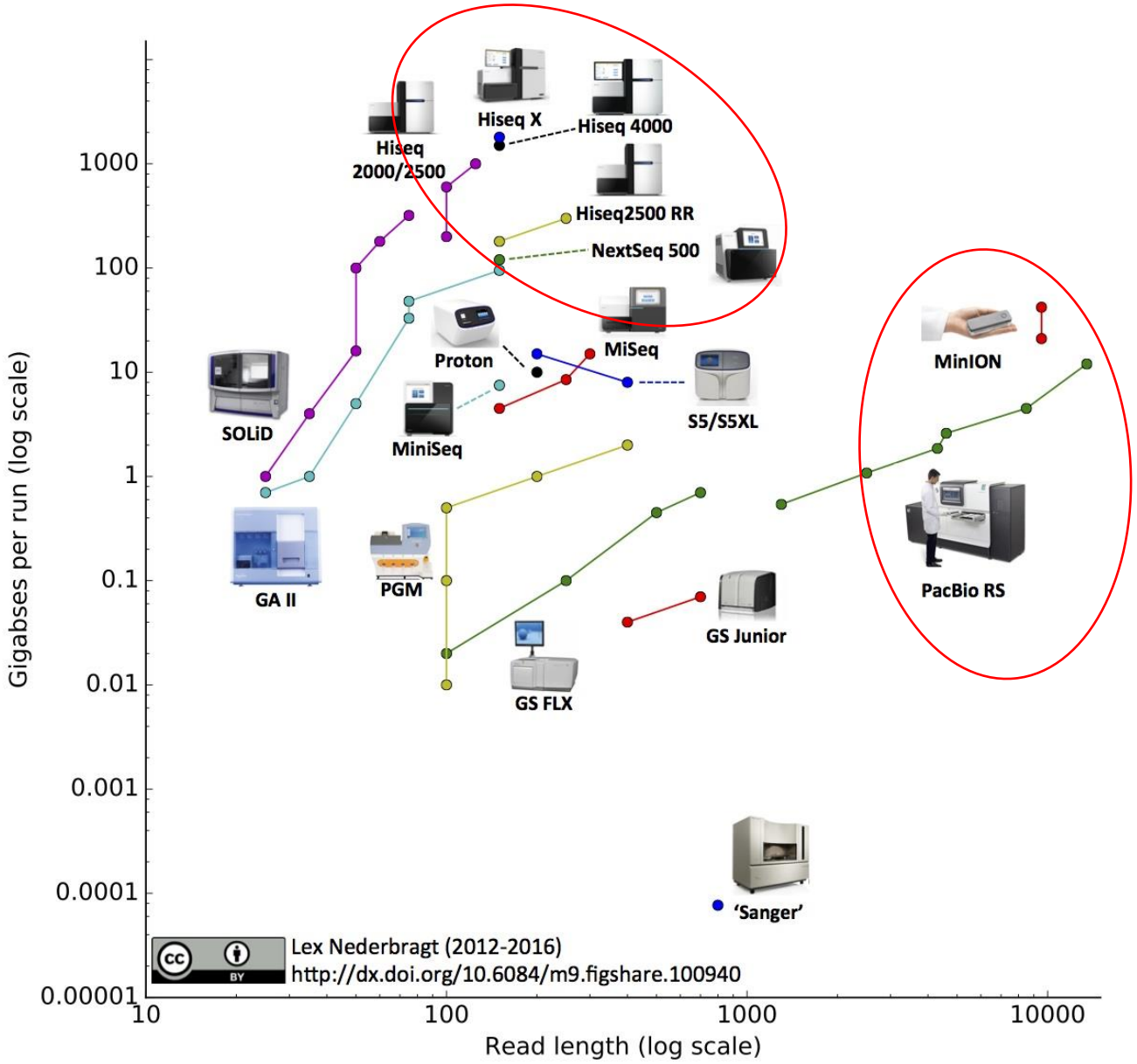
- F. Sanger et al. 1977
  - Short fragments
  - 15-200 nucleotides
  - Slooooow process
- Applied Biosystems automating the process in the late 80's early 90' with capillary electrophoresis, fluorescent dyes, and lasers.
- "First generation sequencing"
  - 500—1000 nucleotides
  - Slow, but much faster than manual
  - The main technique for the human genome project
  - Sequencing 3 gigabases took 10 years
  - Still used, since it is very high quality and cheap (if you only want to look at a handful of sequences)





High Throughput Sequencing (HTS)

- Early 2000's and onwards
- High Throughput Sequencing (HTS) is a collective term
  - Next generation sequencing (NGS)
    - Short reads, (100-300bp)
    - but generates a huge amount of reads (in the billions)
    - 454 Roche
    - **Illumina (HiSeq, MiSeq, NovaSeq, NextSeq)**
    - Ion torrent
  - Third generation sequencing
    - Longer reads, (1000-100kbp)
    - not so many as NGS, but still in the 100k or millions
    - Oxford Nanopore (MinIon, GridIon,PromethIon, etc)
    - **PacBio (Sequel, Revio)**


Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Hiseq 2000/2500
Hiseq X
Hiseq 4000
Hiseq2500 RR
NextSeq 500
MiSeq
Proton
SOLiD
MiniSeq
S5/S5XL
MinION
GA II
PGM
PacBio RS
GS Junior
GS FLX
'Sanger'
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

Illumina

- “Sequencing by synthesis”
- Short fragments
  - 150-300 bp in pairs
- Low error rate (0.1% - 0.5%)
- MiSeq output (2*300bp):
  - 25 million reads (15Gb)
- NextSeq (2*300 bp):
  - 1.2 billion reads (360Gb)
- NovaSeq 6000
  - 20 billion reads (6Tb)
- Other platforms exist

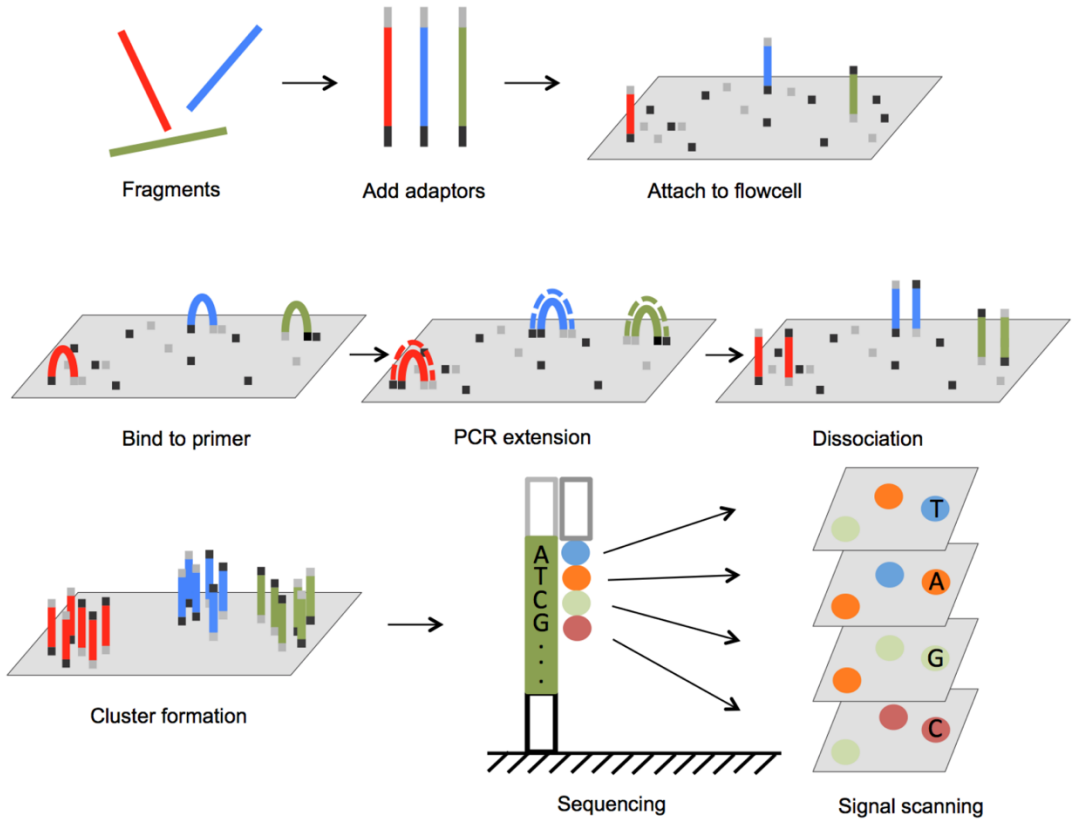


Figure 1: Principle of the illumina sequencing by synthesis (SBS) technology (Lu et al., 2016)

https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=1s

Illumina

Images

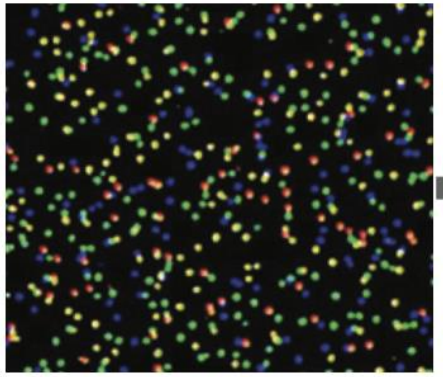


Image Analysis

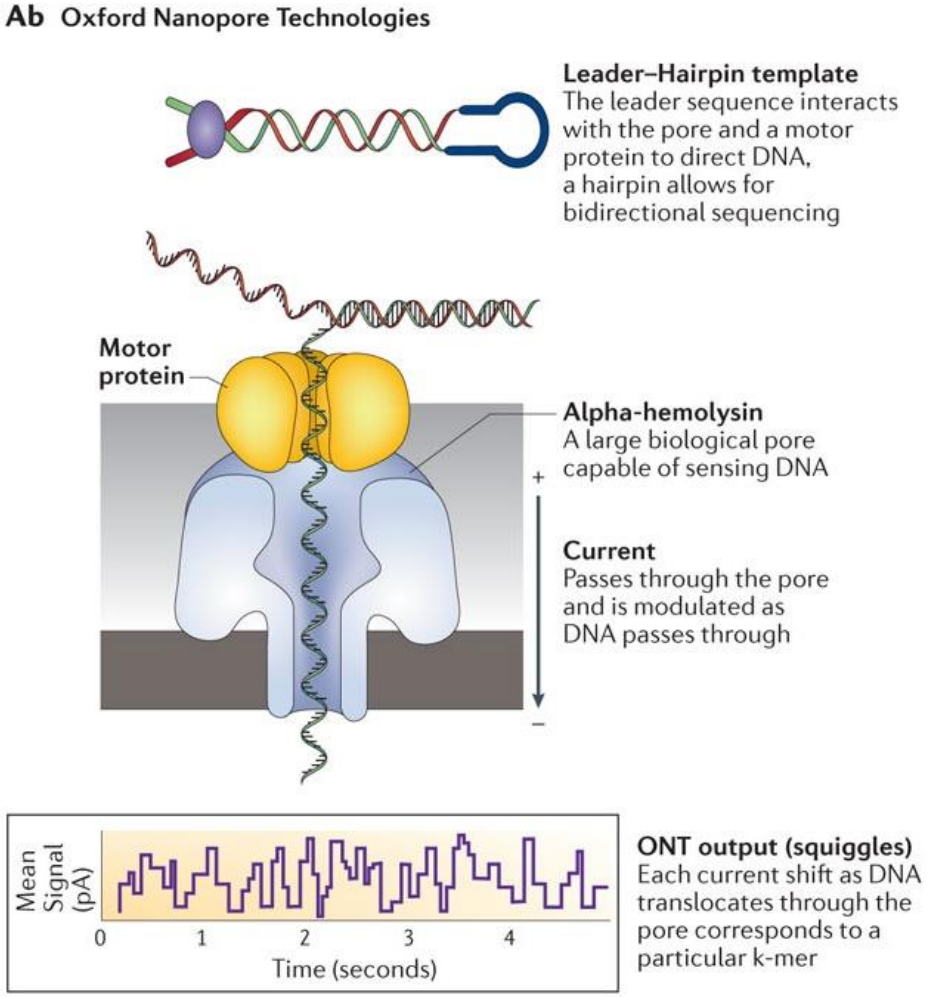
Lane	Tile	X	Y	Cycle 1 - A C G T	Cycle 2 - A C G T
5	12	924	1560	493.1 388.9 3626.7 2359.4	185.6 122.3 380.4 307.8
5	12	773	395	85.5 113.0 2327.5 1158.0	156.3 166.9 113.5 909.6
5	12	165	786	1243.8 741.1 45.8 67.4	318.4 692.6 48.3 41.7
5	12	598	690	1342.6 760.0 60.6 716.6	423.6 505.7 1919.1 959.3
5	12	1107	1207	58.9 63.0 957.5 818.2	98.6 230.5 815.1 512.1
5	12	1074	466	254.7 664.4 47.2 45.1	38.4 41.8 64.9 1102.9
5	12	887	356	743.1 486.4 42.2 305.0	230.3 603.6 -63.1 -20.1
5	12	642	1769	63.2 54.3 861.7 595.7	81.5 86.0 54.9 385.4
5	12	599	314	845.5 533.2 45.2 581.0	260.9 560.9 13.0 78.4
5	12	839	1103	372.0 812.6 16.7 70.5	59.4 69.4 35.4 1394.9
5	12	347	1792	343.8 706.9 108.4 638.5	73.2 43.9 121.6 1882.2
5	12	807	1114	63.9 63.8 828.3 1369.0	1074.4 714.3 -39.9 29.4

Base Calling

Aligned Reads

Oxford Nanopore

- Long to very long reads (10kb -100kb)
- Higher error rate, but it is improving
- Lower output than Illumina
  - MinION (50Gb)
  - PromethION (290Gb)
- Realtime sequencing
- Portable



https://www.youtube.com/watch?v=RcP85JHLmnI

Goodwin, et al 2016

Oxford Nanopore

- Long to very long reads (10kb -100kb)
- Higher error rate, but it is improving
- Lower output than Illumina
  - MinION (50Gb)
  - PromethION (290Gb)
  - GridION (?)
- Realtime sequencing
- Portable

https://www.youtube.com/watch?v=RcP85JHLmnI

nature › letters › article

Letter | Published: 03 February 2016

Real-time, portable genome sequencing for Ebola surveillance

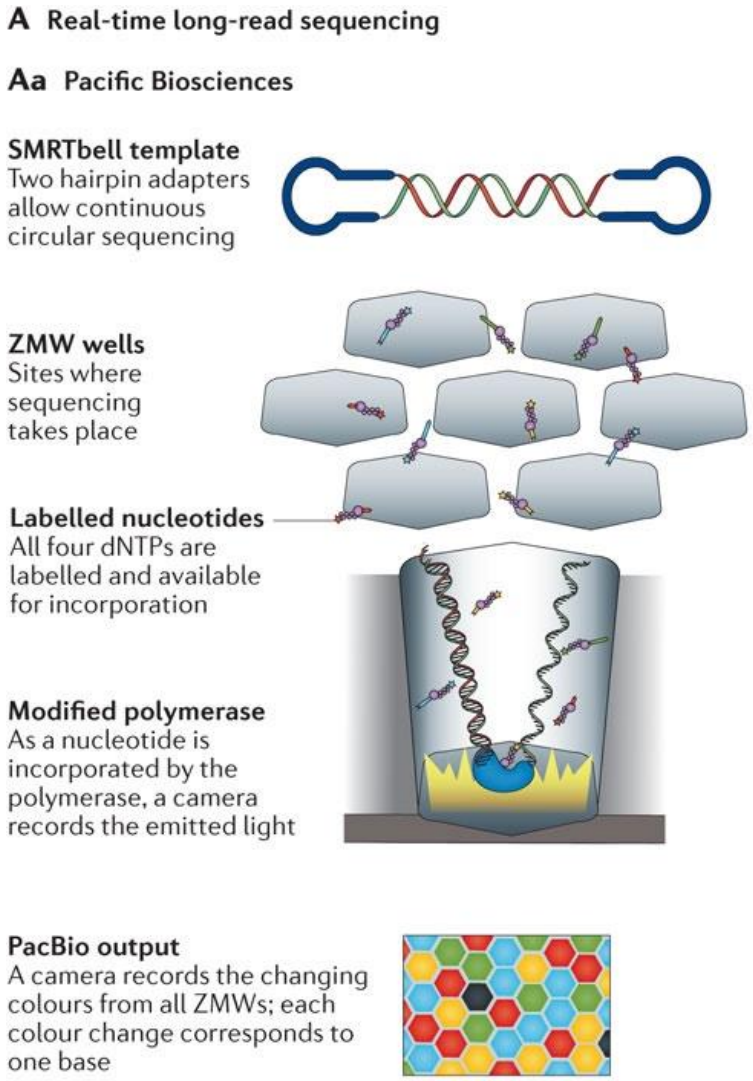
Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan H. J. Baum, Beate Becker-Ziaja, Jan Peter Boettcher, Mar Cabeza-Cabrerizo, Álvaro Camino-Sánchez, Lisa L. Carter, Juliane Doerrbecker, Theresa Enkirch, Isabel García- Dorival, Nicole Hetzelt, Julia Hinzmann, ... Miles W. Carroll + Show authors

Nature **530**, 228–232 (2016) | Cite this article

Figure 1: Deployment of the portable genome surveillance system in Guinea.

PacBio

- SMRT-sequencing
  - Single-molecule Real Time
- Long reads (~15kb)
- Low error rate (0.1%)
- High output
  - Theoretical output:
  - Sequel II (up to 8M reads, 120 Gb)
  - Revio (up to 23M reads, 3Tb)
  - (Real output is ~75% of this)

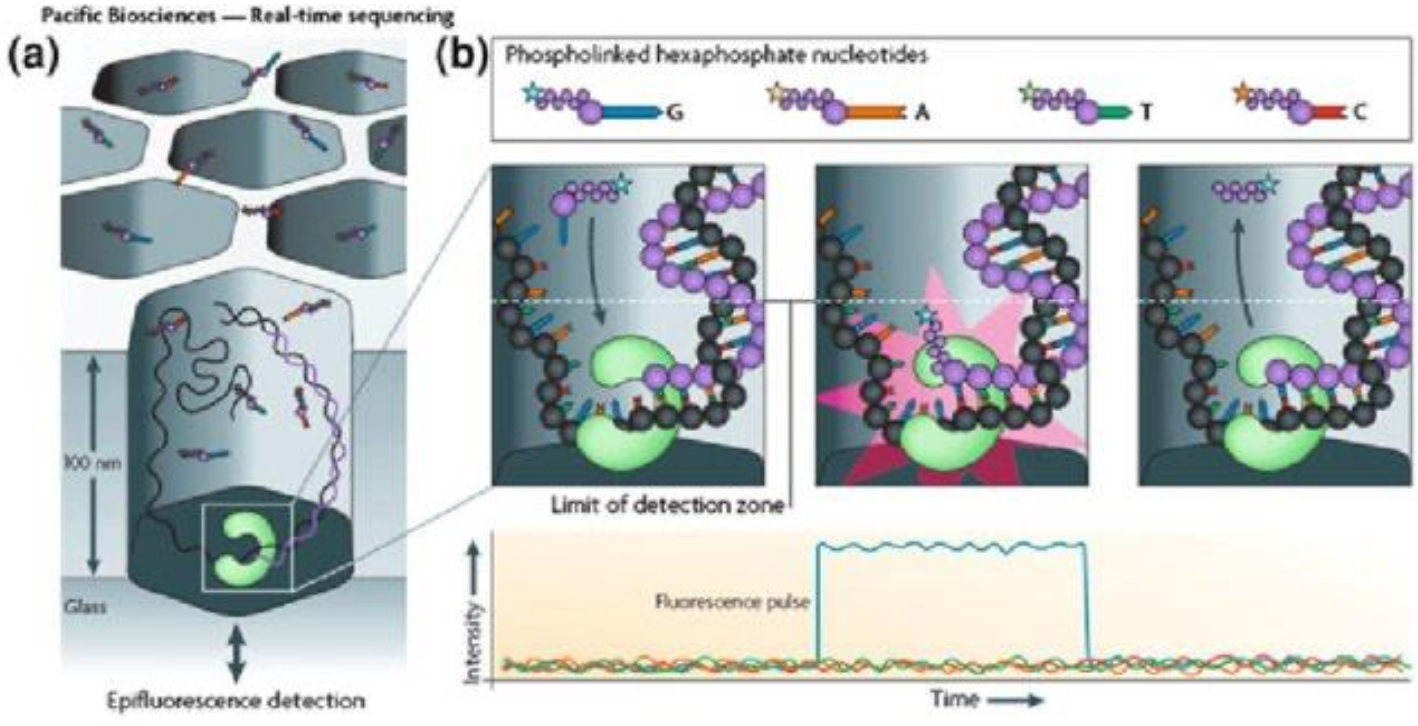


https://www.youtube.com/watch?v=NHCJ8PtYCFc

Goodwin, et al 2016

PacBio

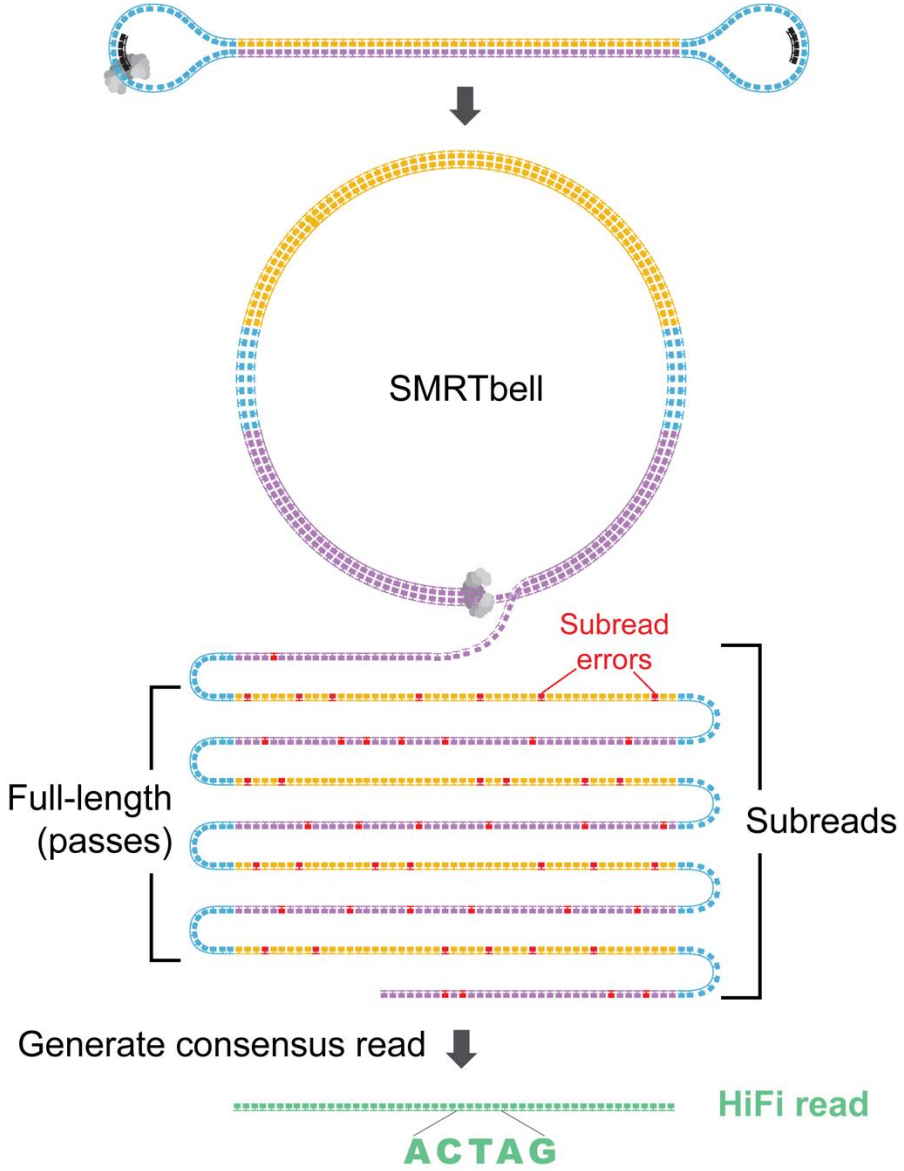
- SMRT-sequencing
  - Single-molecule Real Time
- Long reads (~15kb)
- Low error rate (0.1%)
- High output
  - Theoretical output:
  - Sequel II (up to 8M reads, 120 Gb)
  - Revio (up to 23M reads, 3Tb)
  - (Real output is ~75% of this)



https://www.youtube.com/watch?v=NHCJ8PtYCFc

PacBio

- The same molecule is sequenced multiple times, since it is circular
- CCS «HiFi» reads,
- circular consensus sequence with very high quality (99.9% accuracy)¨

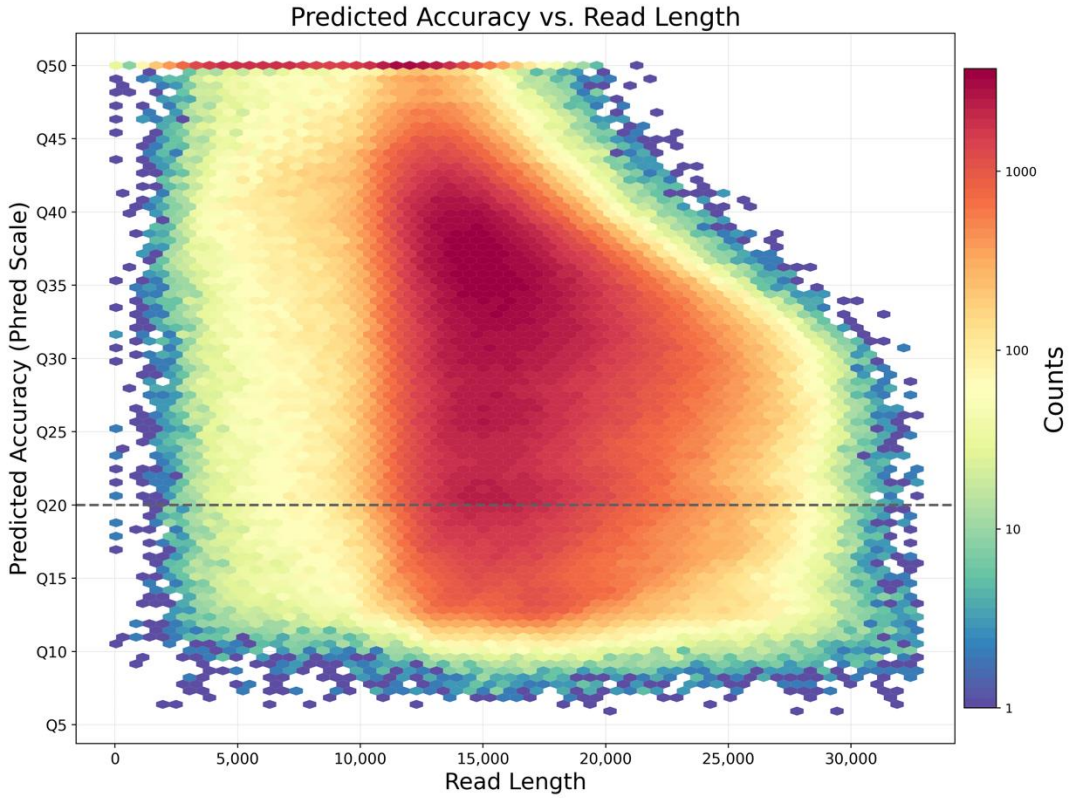


https://ccs.how/how-does-ccs-work.html

PacBio

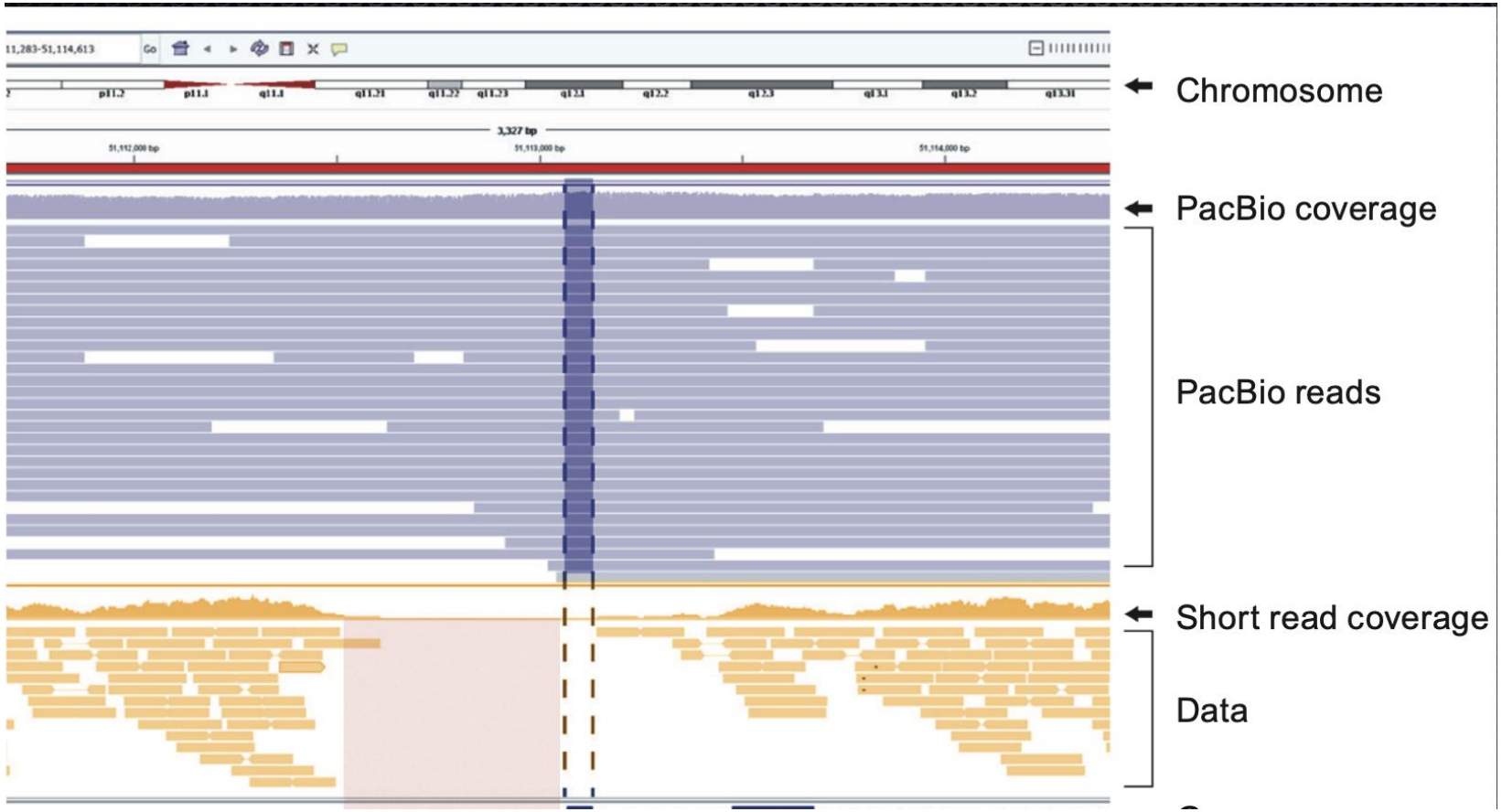
- The same molecule is sequenced multiple times, since it is circular.

- «HiFi» reads
- CCS «HiFi» reads,
- circular consensus sequence with very high quality (99.99% accuracy)



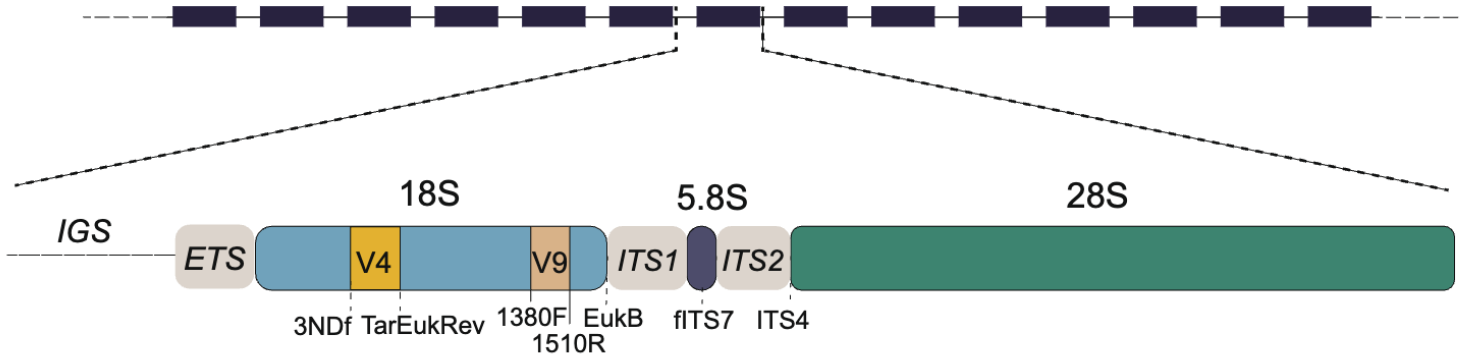
PacBio

- Long-read advantage
- In genome projects:
- Spans repeats better
- Uncovering long structural variation



Ribosomal operon as example

- Often shortened rRNA or rDNA
- 16S – 5S - 23S in prokaryotes
- 18S – 5.8S -28S in eukaryotes
  - S stands for Svedberg units; a unit of molecular size determined by centrifugation.
- Present across the Tree of Life!
- The full length is in the range of 5000-7000 bp (but with a lot of variation)
- Typically, for Illumina sequencing, a region of 300-450bp is used (e.g., V4).



Long-read metabarcoding

- Sequencing the full operon is possible with the development of sequencing technologies
- Stronger phylogenetic signal

- Comes with extra challenges
  - Harder to amplify longer regions
  - More chimeric sequences
  - Lower sequencing depth

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity

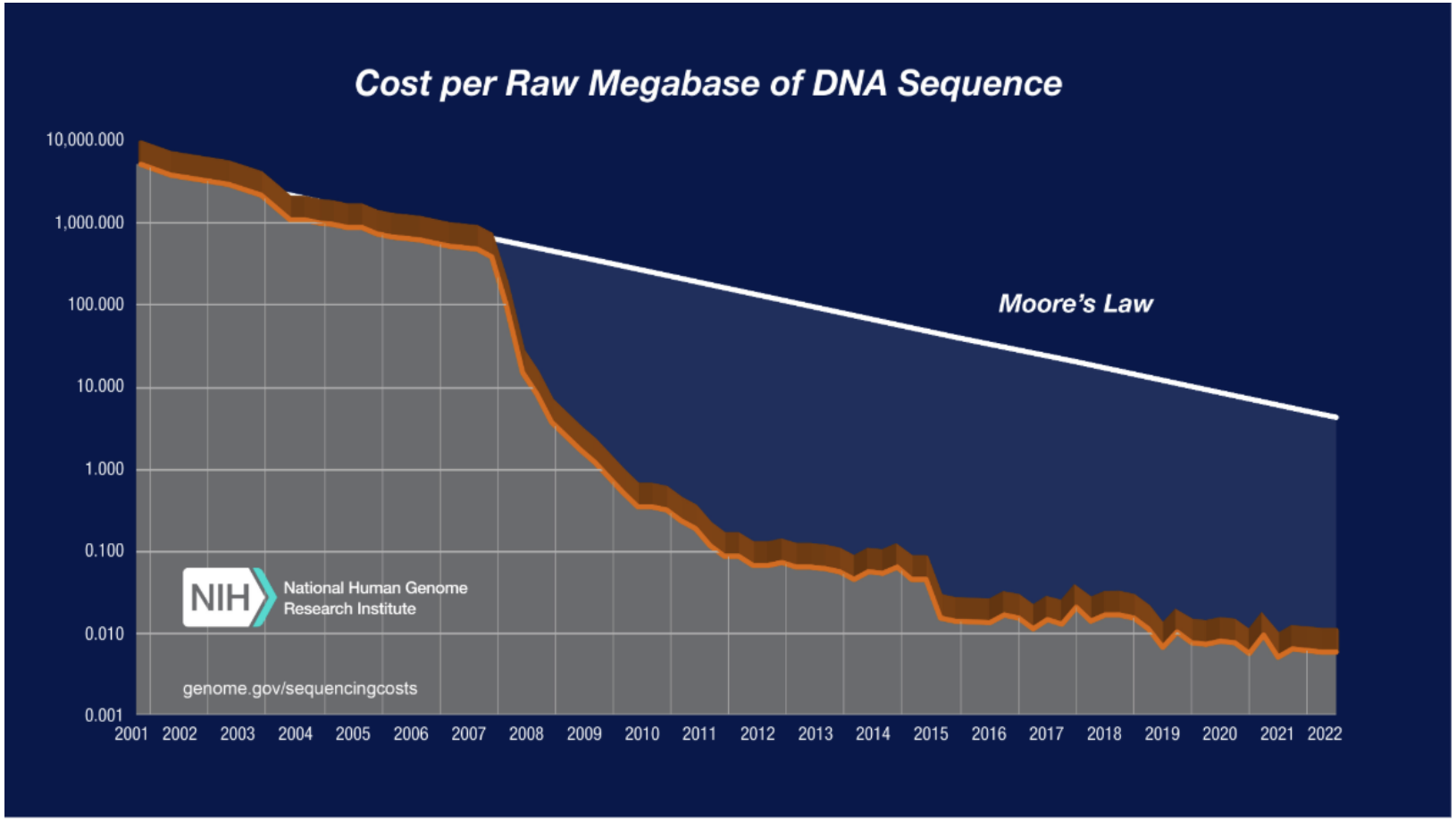
Mahwash Jamy[1] | Rachel Foster[2] | Pierre Barbera[3] | Lucas Czech[3] | Alexey Kozlov[3] | Alexandros Stamatakis[3,4] | Gary Bending[5] | Sally Hilton[5] | David Bass[2,6] | Fabien Burki[1]

Cost and output

Output of some of the machines in NorSeq

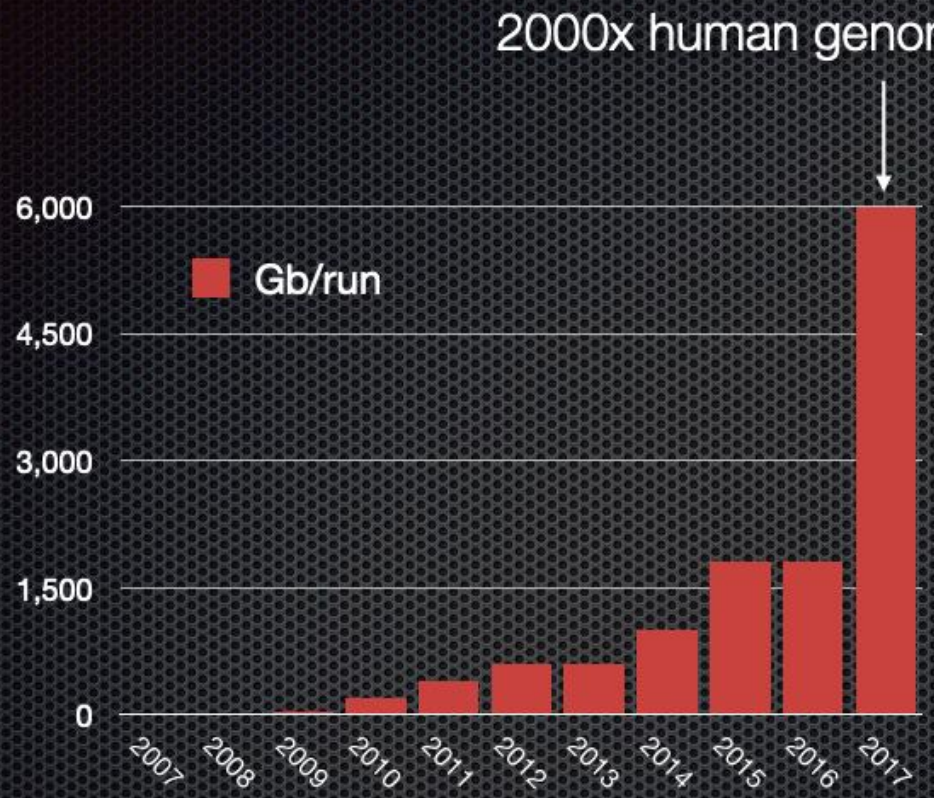
Platform	NovaSeq	NextSeq	MiSeq	PacBio REVIO
@NorSeq	2	5	4	1
Run time	1-2 days	29 hours	29 hours	0.5-30 hours
Read accuracy	99%	99%	99%	99%
Read number	20x10^9	400,000,000	20,000,000	18,000,000
Read length	2x150 bp	2x150 bp	2x300 bp	~20 kb
Output	6000 Gb	129 Gb	12 Gb	360Gb

Cost and output



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

HTS throughput: data per run



Illumina NovaSeq6000

- 6 Tb (6000 Gb)
- 2000x human genome

Really, so what?

Parameter	ABI 3100	ABI 3730	NovaSeq6000
Read length	~700	~700	150 (x2)
Reads per run	16	96	20000000000
Run time	2 hours	30 minutes	2 days
Time for 1x human genome (3 Gb)	120 years	15 years	~90 seconds

Typical sequence formats

- Raw data: **FASTQ** (reads + quality scores).
- Reference data: **FASTA** (genomes, transcripts).
- Alignments and mapping files: SAM/BAM (sequencing reads mapped to a reference).
- Annotations: GTF/GFF (gene structures).
- Variants: VCF (SNPs, indels).
- Different compression formats (e.g. .gz, .zip) and indexing (e.g. .bai, .fai).
- **Stockholm** (.sto)
  - Rich format for MSAs, supports annotations (e.g., secondary structure, RNA families).
  - Used by tools like Infernal, Rfam, Pfam.
- **Clustal** (.aln): Older format, human-readable.
- **Phylip**: Compact, used in phylogenetics.

Sequence formats - fastq

- FASTQ
- Four lines per sequence
- 1) seq. name
- 2) DNA sequence
- 3) linebreak
- 4) Quality in phred score

Header Sequence Quality

```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?######
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?
```

Quality scores as ASCII characters:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJK
Q:        0    5        15        30    40
P_error:  1.0  0.32     0.032     0.001 0.0001
```

$$Q = -10\log_{10} P_{error}$$

https://gencoded.com/index.php/2020/05/20/fastq-format-an-overview/

Sequence formats – Phred score???

- Phred scores represent the quality of base calls from sequencing data, originally for Sanger sequencing, and later adapted for NGS. The score is logarithmically related to the probability of a base call being wrong.
- PHRED stands for Phil's Read Editor (University of Washington Genome Center) and was developed in the 1990's.
- $Q = -10 \log_{10} (P)$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Bin	Emoji
N	🚫
2-9	💀
10–19	💩
20–24	⚠️
25–29	😄
30–34	😆
35–39	😎
≥ 40	😍

https://fastqe.com/

Sequence formats – Phred score???

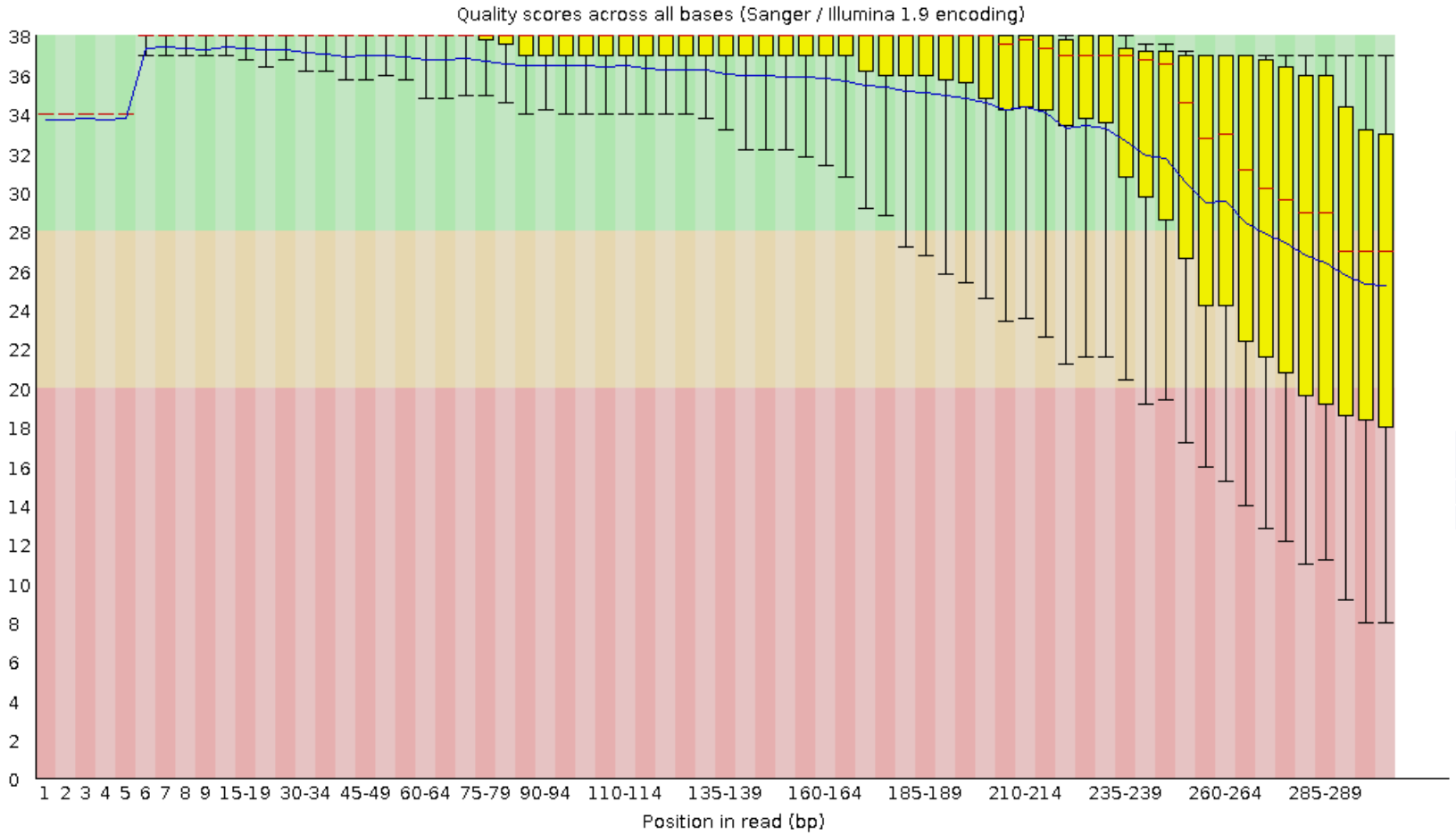
- Phred scores represent the quality of base calls from sequencing data, originally for Sanger sequencing, and later adapted for NGS. The score is logarithmically related to the probability of a base call being wrong.
- PHRED stands for Phil's Read Editor (University of Washington Genome Center) and was developed in the 1990's.
- $Q = -10 \log_{10} (P)$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Relating the Q-values to some typical Illumina data

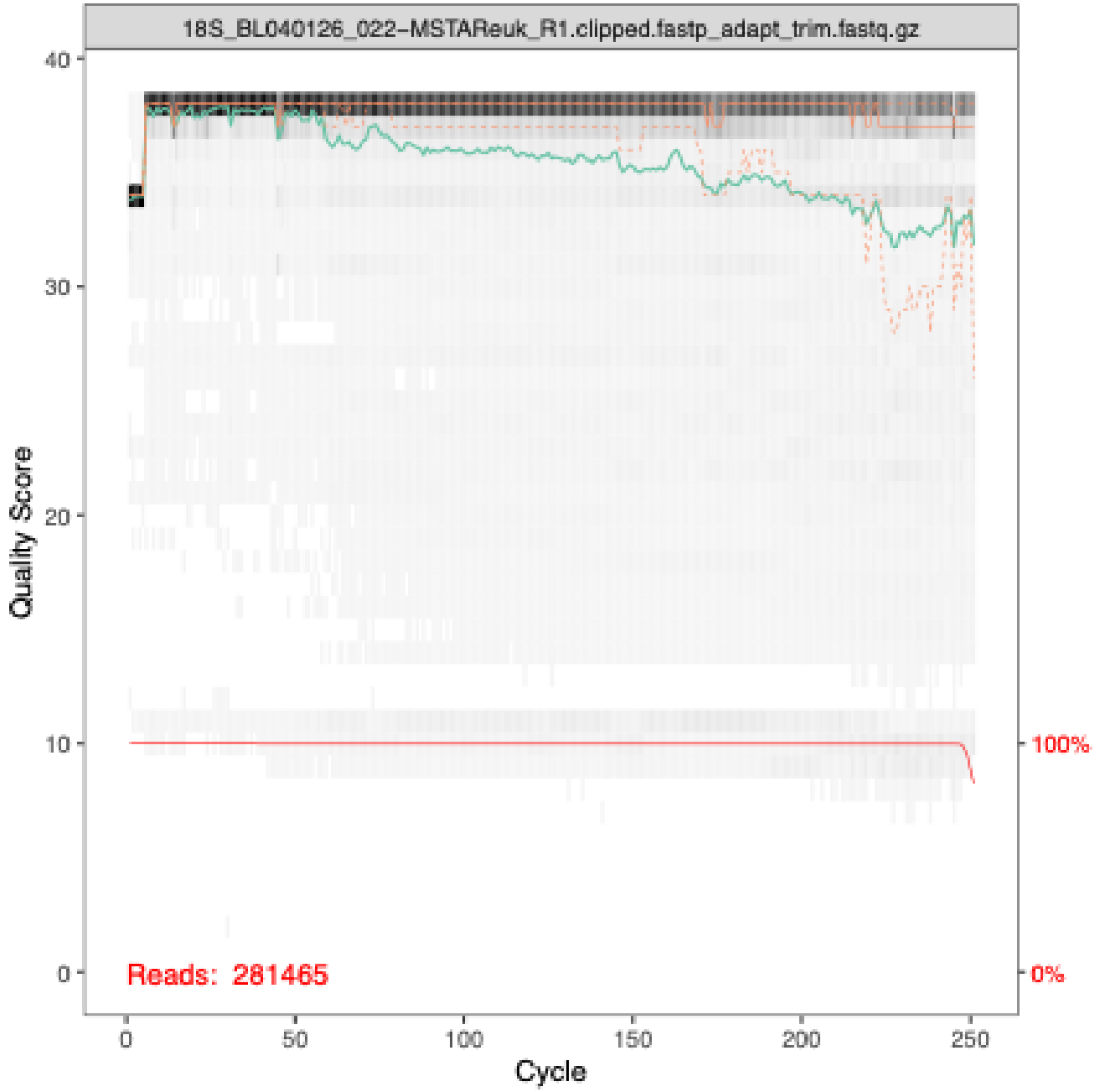
Metric	MiSeq Read 1	MiSeq Read 2	HiSeq Read 1	HiSeq Read 2
%Bases Q>30	91.9	87.5	89.3	86.1
%Total Bases Q >30	89.7		87.7	

Illumina sequence

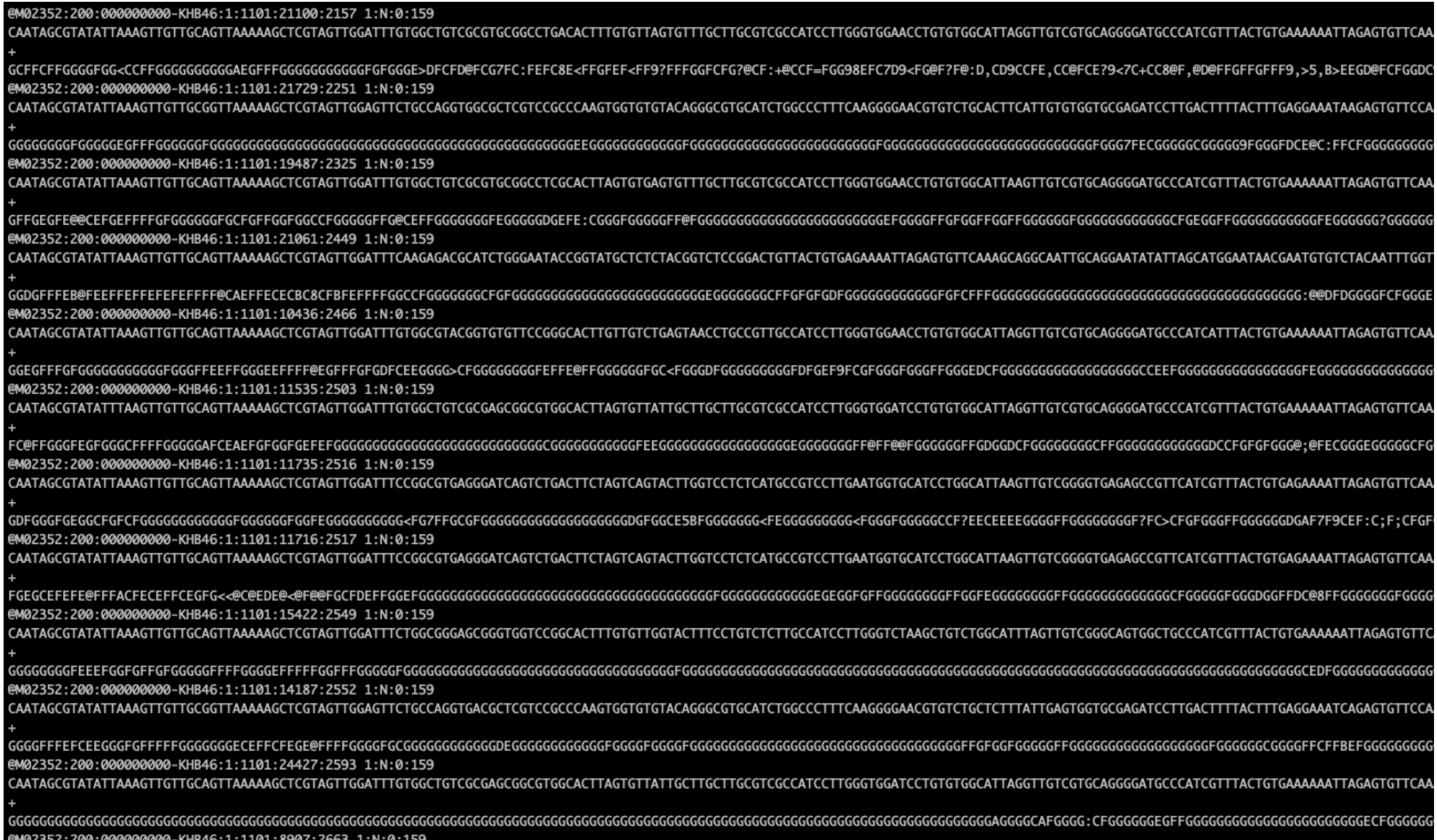


Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,00	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Metric	MiSeq Read 1	MiSeq Read 2	HiSeq Read 1	HiSeq Read 2
%Bases Q>30	91.9	87.5	89.3	86.1
%Total Bases Q >30	89.7		87.7	



Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,00	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%



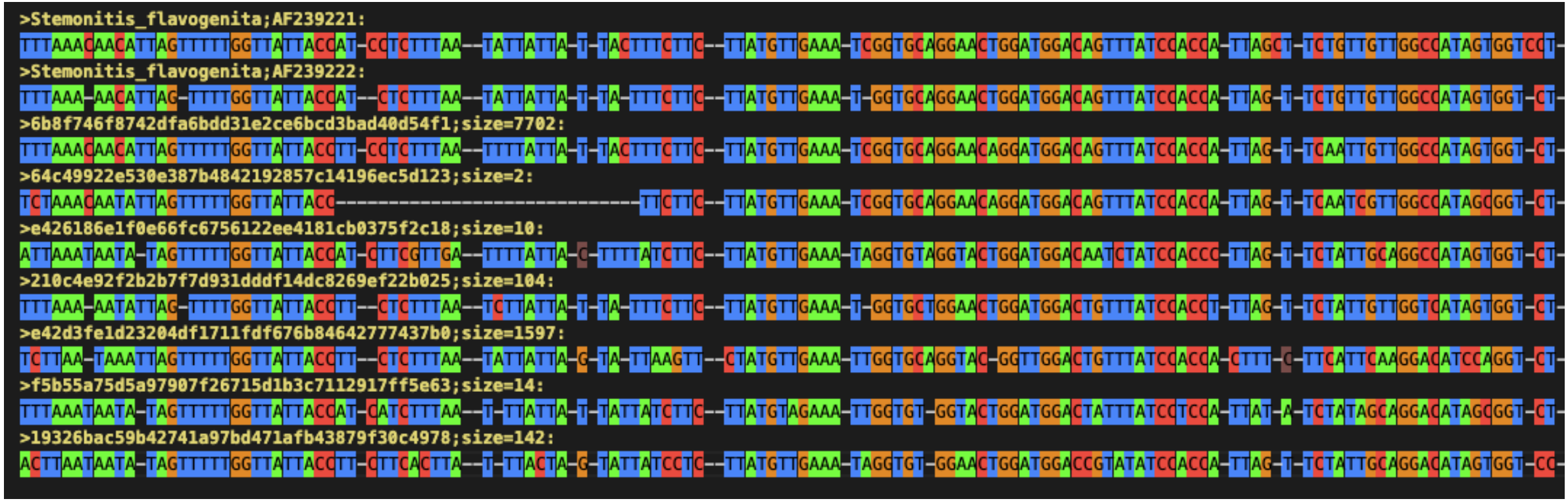
Sequence formats - fasta

- FASTA stands for "FAST-All", because it works with any alphabet, an extension of the original "FAST-P" (protein) and "FAST-N" (nucleotide) alignment tools.
- Initial release 1985
- A sequence (DNA, RNA or Proteins) is represented with two lines:
- A sequence starts with a greater-than sign “>” (often called the header) followed by a description. The next lines represent the sequence, one letter per amino acid or nucleic acid.

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLT
DEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPG
LTSDSDKIPFHPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTL
TMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGXIENY
```

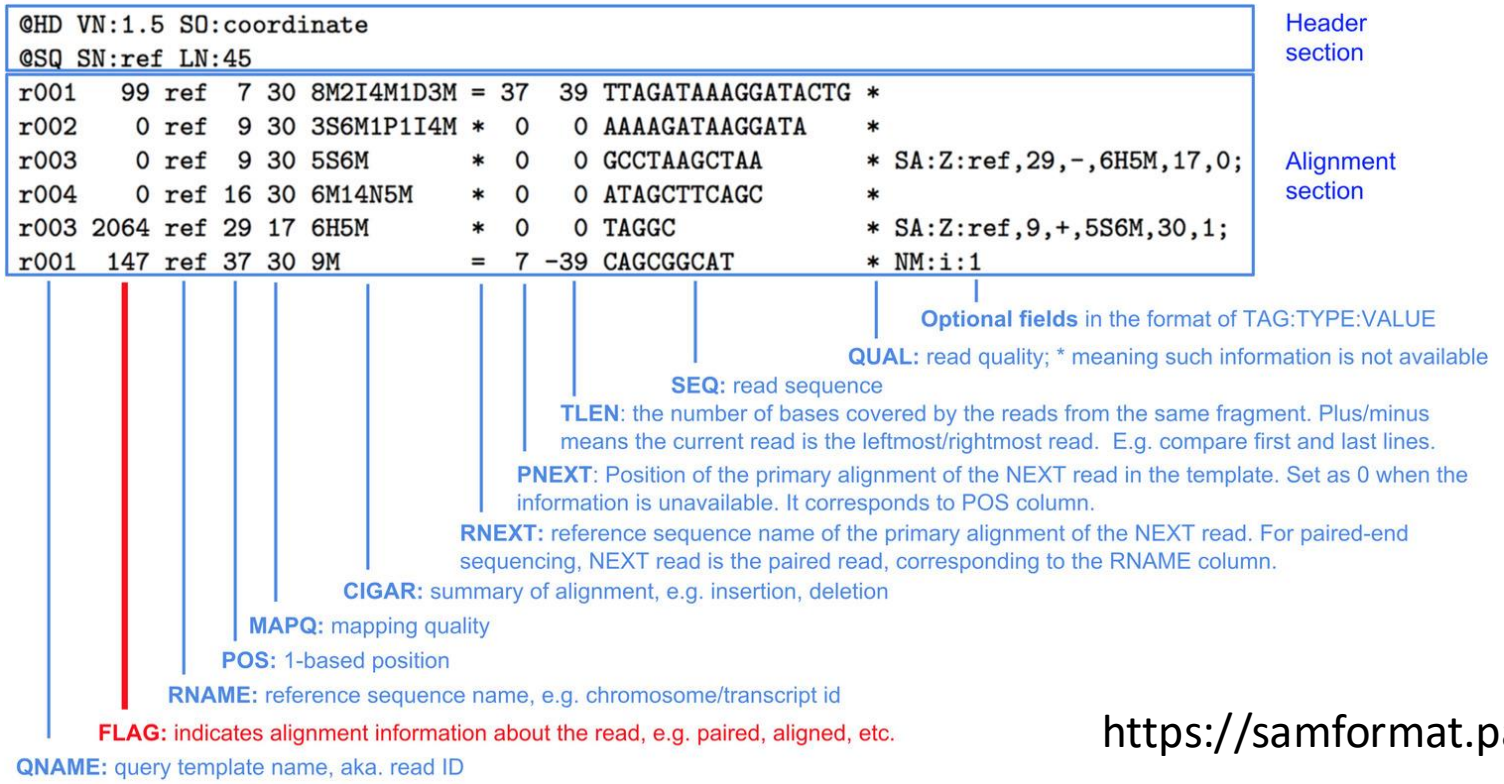
Sequence formats - fasta

- Fasta files can represent alignments. Displayed with a different color for each base (or amino acid) it looks like this (gaps in the alignment are marked with -)



Sequence formats – SAM/BAM

- SAM (Sequence Alignment/Map) is a text-based file format that stores alignments of sequencing reads to a reference genome. It includes read names, alignment positions, mapping quality, and optional tags. (BAM is a binary, compressed version)



https://samformat.pages.dev/sam-format-flag

Sequence formats – other

- Various other formats you might encounter.
  - **.csv (Comma-separated values): columns are separated by the comma character, and rows are** terminated by newlines.
  - **.tsv (Tab-Separated Values):** Columns are separated by **tabs**, rows by newlines.
  - **Newick files (.nwk):** Format for storing phylogenetic trees; clades are grouped using parentheses.
  - **Nexus files (.nex, .nexus):** Used to store phylogenetic trees, alignments, metadata, and more. Structured into blocks, each beginning with BEGIN and ending with END;.
  - **GTF, GFF, GFF3:** Text formats for describing gene features and annotations.
  - **VCF (Variant Call Format):** Stores variants such as SNPs and indels, along with metadata and sample genotype info.

Sequence formats – other

- More general formats:
  - **.md (Markdown):** a lightweight markup language used for formatting plain text. Common in documentation, READMEs, notebooks (e.g., Jupyter), and reports.
  - **JSON (JavaScript Object Notation)**: Data is structured as key-value pairs, arrays, or nested objects.
  - **XML (eXtensible Markup Language):** Structured with tags, supports attributes and nested elements.
  - **HTML (HyperText Markup Language):** A markup language used to structure and display content on the web.
  - Etc.

Sequence formats – Don’t panick

- Broadly speaking , there are only two formats.
  - **Binary formats**: Designed for efficient storage and processing by computers. They are compact and fast to parse but generally unreadable to humans (e.g., BAM, BCF, HDF5).
  - **Text-based** (ASCII) formats: Human-readable and easy to inspect or edit with a text editor. However, some can be structured in ways that are complex or verbose, making them difficult to interpret without prior knowledge (e.g., XML, GTF, Newick).