2001: A Base Odyssey

The era of genomics and massive parallel sequencing

Matthias Zepper, PhD

February 24, 2025

NGI Stockholm
https://ngisweden.scilifelab.se

2001: Draft assemblies of the human genome are published

Figure 1: The private company Celera [Venter et al., 2001] and the International Human Genome Sequencing Consortium [Lander et al., 2001] both publish a draft sequence of the euchromatic portion of the human genome.

doi: 10.1126/science.1058040; doi: 10.1038/35057062

The overture to the genomic era

A remake of the opening scene by SumoSebi, CC-BY-SA on Wikimedia Commons

Stanley Kubrick's *2001- A Space Odyssey* premieres 2 April 1968

1968: Nobel prize for the interpretation of the genetic code

Nobel Prize in Physiology or Medicine 1968

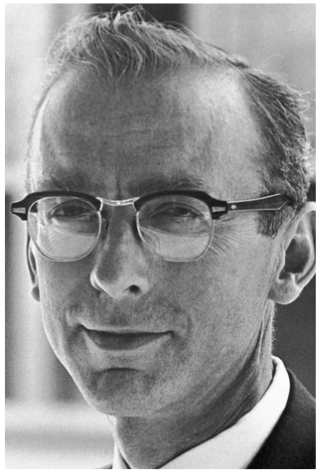


Photo from the Nobel Foundation archive.

Robert W. Holley

Prize share: 1/3

Photo from the Nobel Foundation archive.

Har Gobind Khorana

Prize share: 1/3

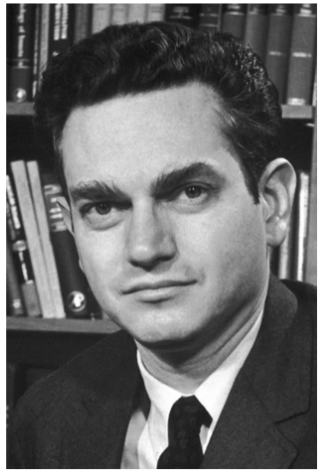
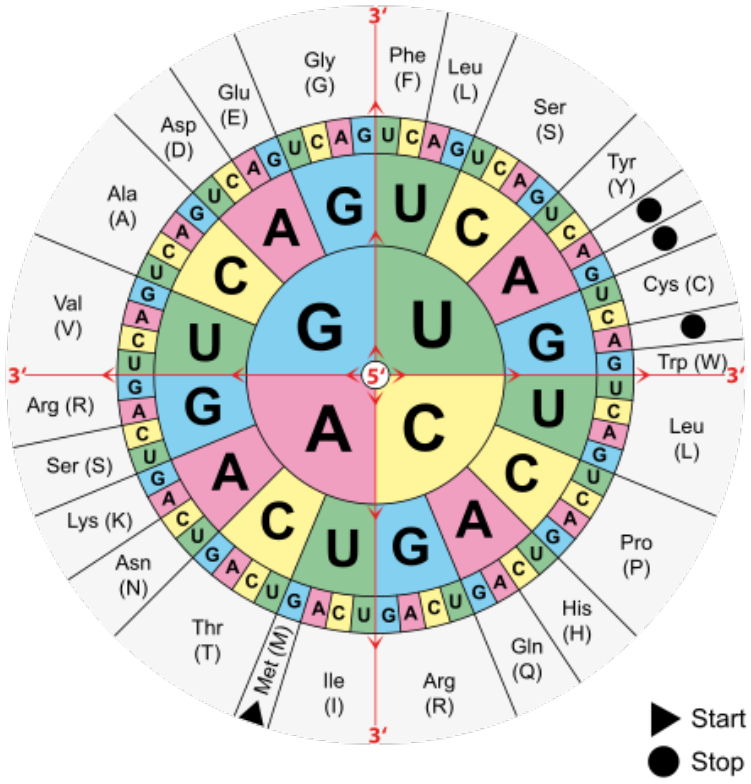


Photo from the Nobel Foundation archive.

Marshall W. Nirenberg

Prize share: 1/3

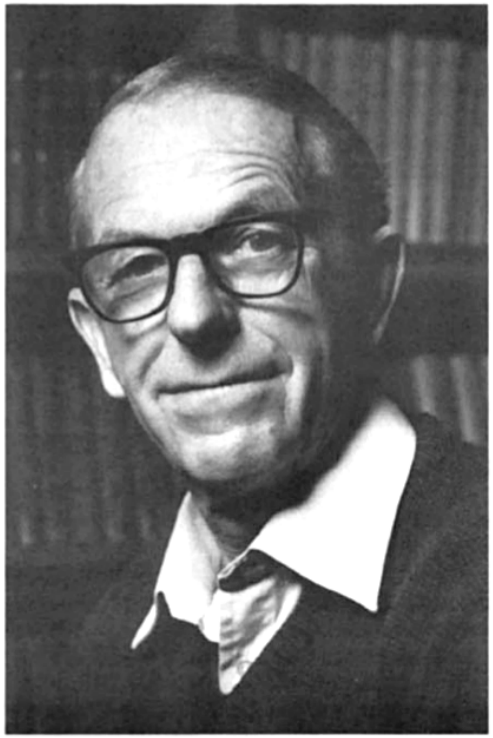


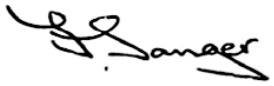
- The genetic code is (almost) universal[1]
- It was resolved entirely using synthetic sequences.

[1] http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=tgencodes

https://www.nobelprize.org/prizes/medicine/1968/summary/ | User:Mouagip, Wikimedia Commons, PD

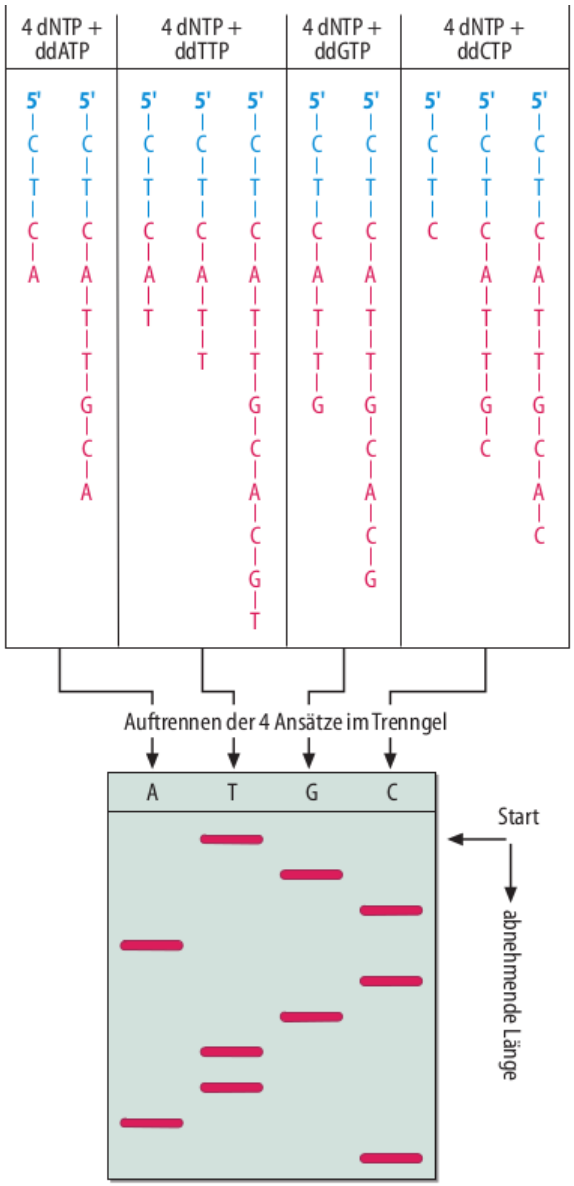
Encoded information of naturally occuring DNA unknown



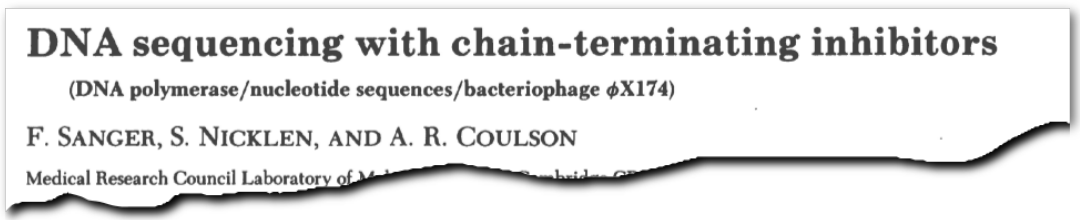


- Peptides could be sequenced since the 1950s (Sanger method, Edman degradation).
- Sequencing of DNA was one of the most urgent, unresolved problems in the early 1970s.
- Frederick Sanger (Nobel laureate for sequencing Insulin 1958) started working with DNA.

1977: Chain-termination sequencing by Frederick Sanger



- DNA fragments could be separated by size.
- Sanger's method creates sequence-derived length patterns.
- It relies on radioactive labeling and in-vitro amplification of DNA.



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 M[illegible]

Figure 2: [Sanger et al., 1977]

← Fig.5.40, Löffler: Biochemie und Pathobiochemie 7th Ed.

1980: Nobel prize for DNA sequencing

Nobel Prize in Chemistry 1980

Photo from the Nobel Foundation archive.

Paul Berg

Prize share: 1/2

Photo from the Nobel Foundation archive.

Walter Gilbert

Prize share: 1/4

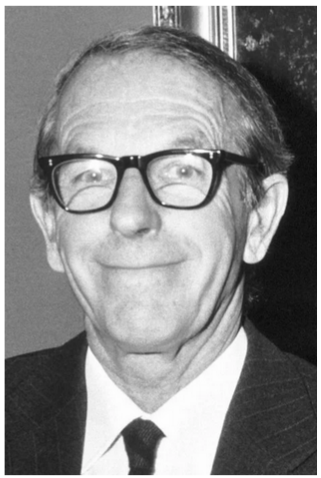


Photo from the Nobel Foundation archive.

Frederick Sanger

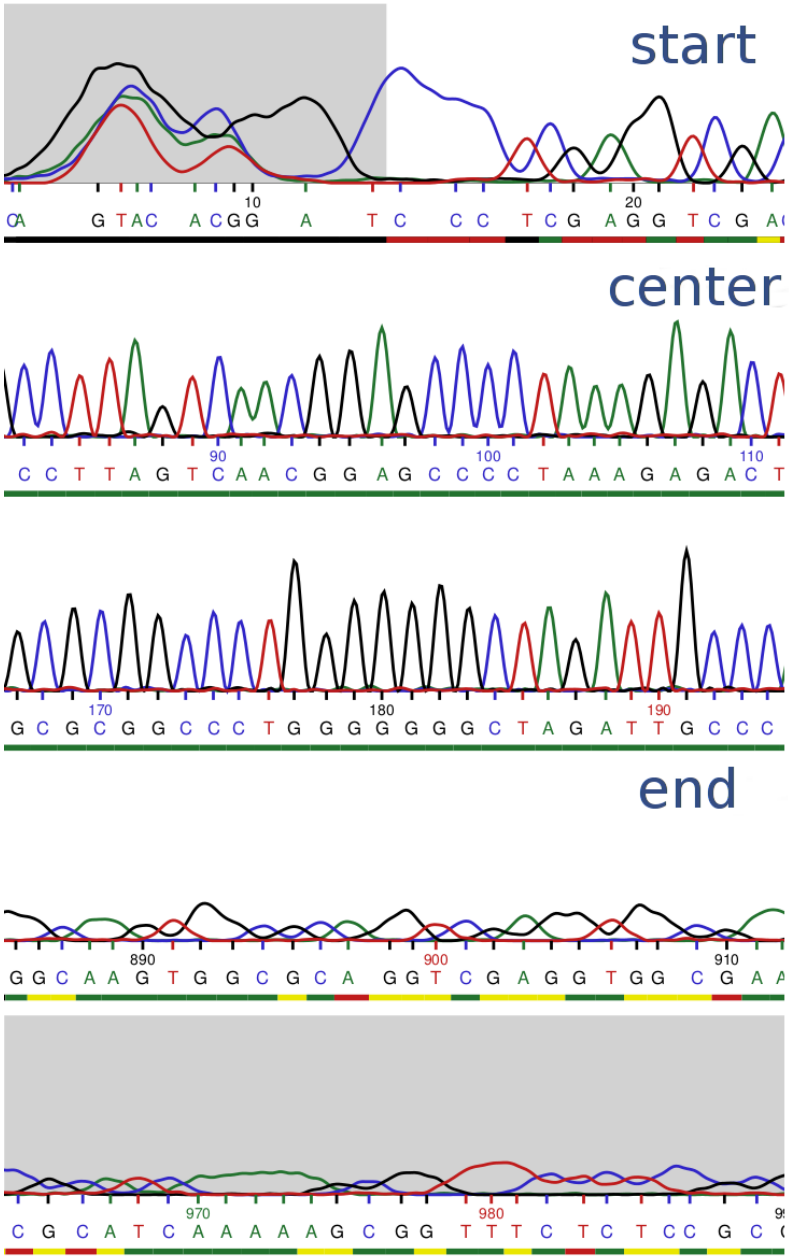
Prize share: 1/4

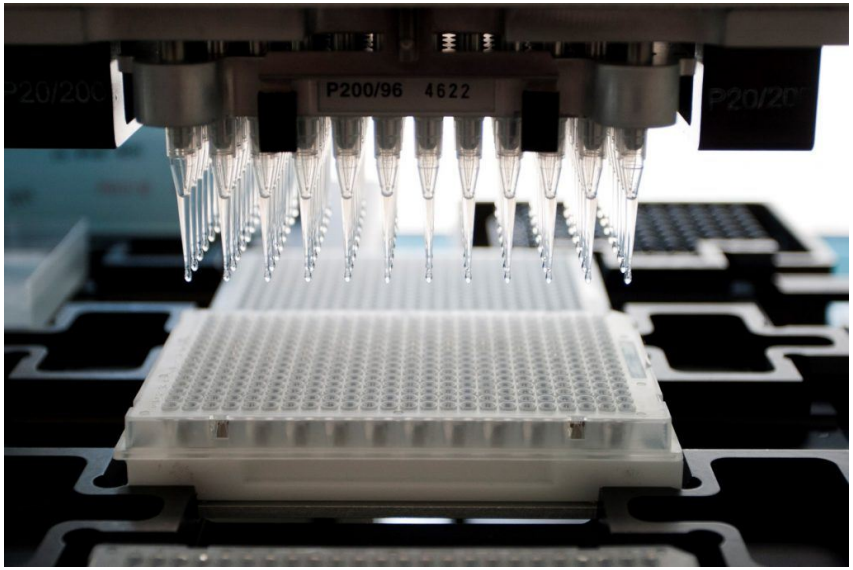
- Ample DNA input needed

  PCR was introduced in 1989
- Four reactions per sequence
- Read length ~ 200bp

https://www.nobelprize.org/prizes/chemistry/1980/summary/

Advanced Sanger sequencing for the Human Genome Project





- Fluorescent chain terminators.
- Capillary electrophoresis for size separation of amplicons.
- Parallelized and automated.
- Sequencing technology of the Human Genome Project (1990-2004).

Next-generation sequencing

New high-throughput methods were developed

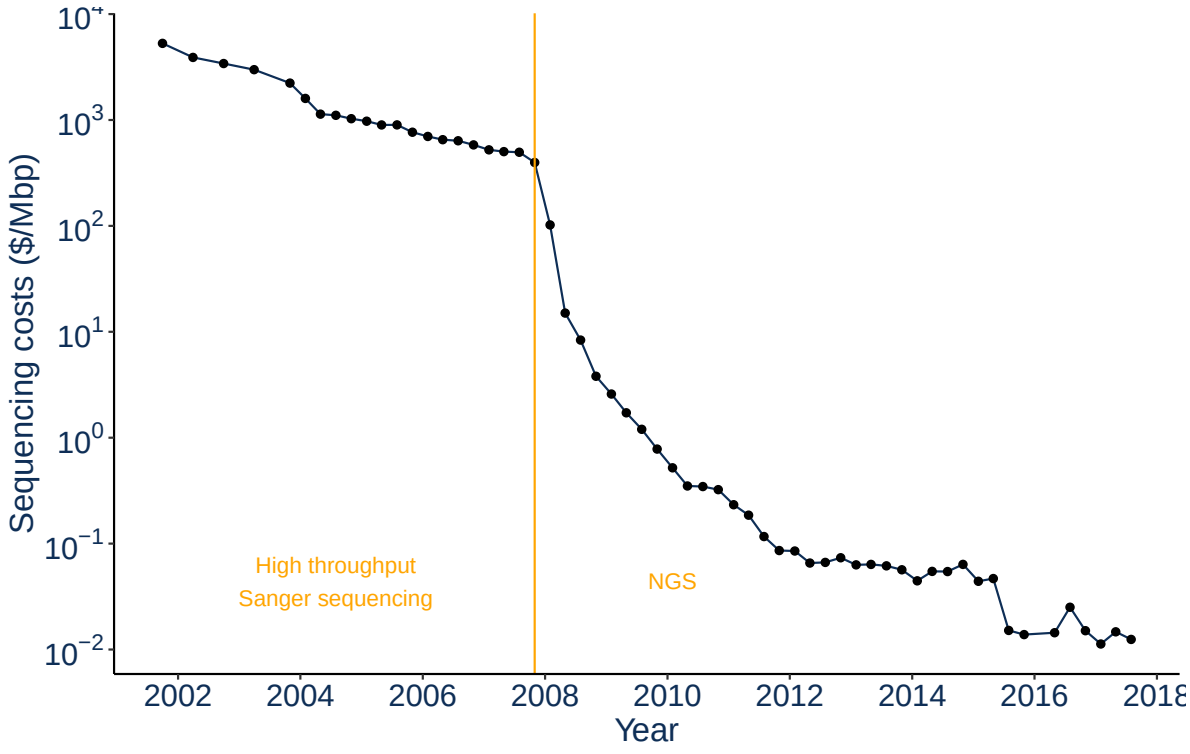


Figure 3: Sequencing costs per one million bases of raw sequence

1990-2004: Human Genome Project sequencing: US $500 million

2025: Sequencing of a human genome: $\sim$ US $100-1000

National Human Genome Research Institute (NHGRI)
https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Around 2010: Sanger sequencing was outcompeted by NGS

ABI 3730xl DNA Sequencer
(Sanger Multiplex, 2013)

- ~6912 reads of 400bp
- ~2,76 Mbp / day

Illumina HiSeq 2500
(NGS / MPS, 2013)

- ~600 Million reads of 100bp
- ~60.000 Mbp / day

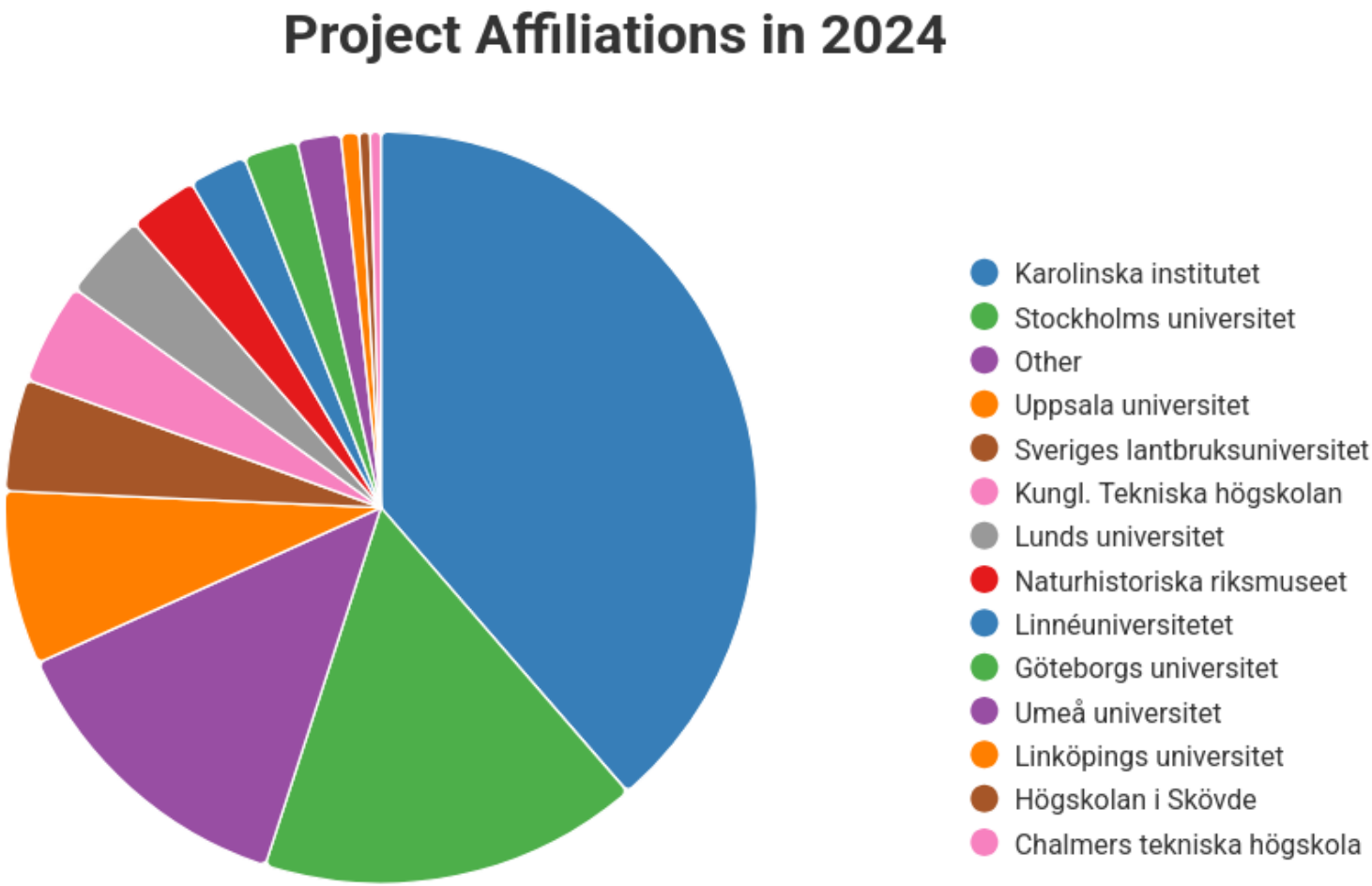
(depending on settings and sequencing chemistry used)

National Genomics Infrastructure Sweden

DNA sequencing facilities provide sequencing capacity

- DNA sequencing of paramount importance for life science.
- 2013: National Genomics Infrastructure Sweden is founded.
- Our mission is to offer a state-of-the-art infrastructure available to researchers all over Sweden.

Users of the National Genomics Infrastructure Sweden



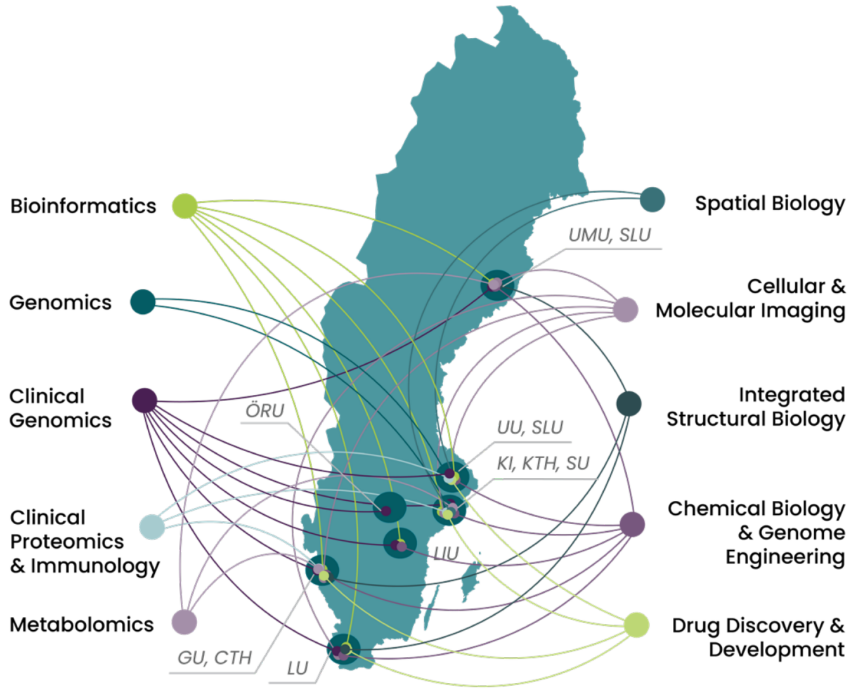
https://ngisweden.scilifelab.se/resources/ngi-stockholm-status/

National Genomics Infrastructure Sweden

NATIONAL GENOMICS INFRASTRUCTURE

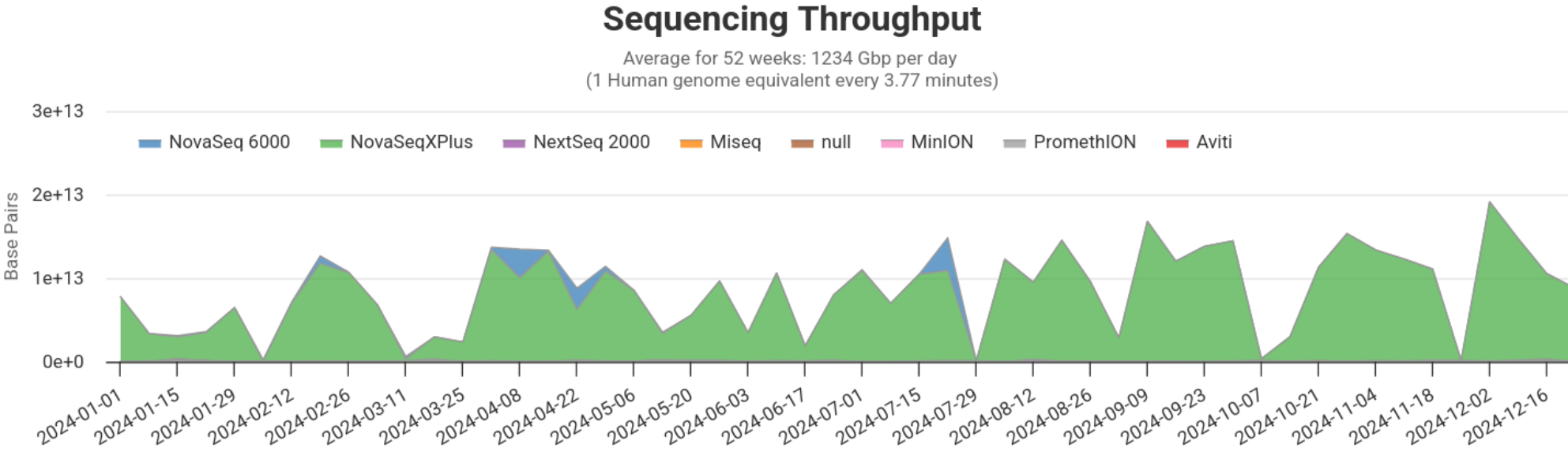
SciLifeLab

- NGI is a sequencing facility for *research projects*
- Part of the Genomics Platform at SciLifeLab
- Distributed in 3 nodes:
  - SNP&SEQ Technology Platform, Uppsala
  - Uppsala Genome Center
  - NGI Stockholm + Eukaryotic Single Cell Genomics (ESCG), Solna



https://ngisweden.scilifelab.se

NGI-S employs various sequencing technologies



- In 2024, NGI Stockholm sequenced on average 1200 Gbp/day

https://ngisweden.scilifelab.se/resources/ngi-stockholm-status/

Sequencing platforms

Sequencing platforms / technologies since Sanger

Next generation sequencing

- Roche 454 sequencing (Pyrosequencing)
- Ion semiconductor sequencing
- **Illumina (Solexa) sequencing**
- **PacBio HiFi Sequencing**

Third generation sequencing

- **Oxford Nanopore sequencing**
- **Element Biosciences Avidite Sequencing**
- Ultima Genomics UG 100 Sequencing
- MGI DNBSEQ Technology
- Singular Genomics G4X

Platforms in **bold** are in use at the National Genomics Infrastructure

Sequencing platforms / technologies since Sanger

Sequencing by synthesis

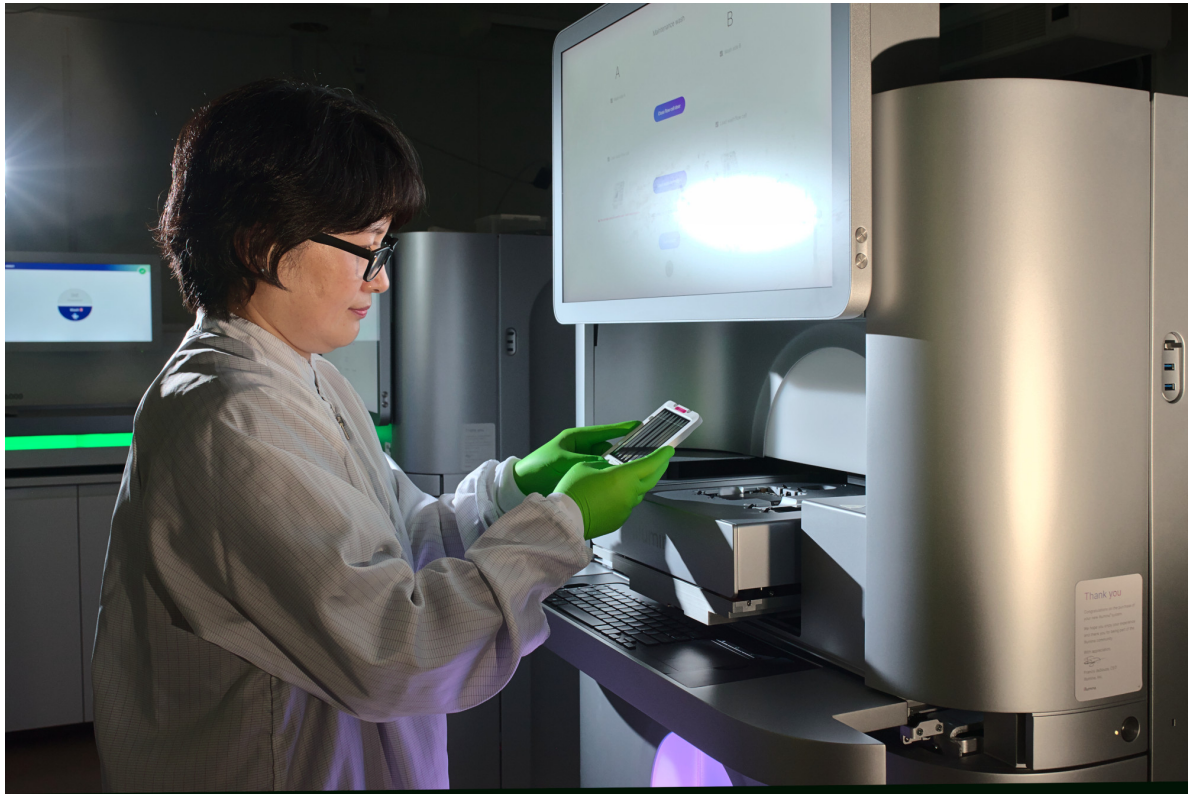
- Roche 454 sequencing (Pyrosequencing)
- Ion semiconductor sequencing
- **Illumina (Solexa) sequencing**
- **PacBio HiFi Sequencing**
- **Element Biosciences Avidite Sequencing**
- Ultima Genomics UG 100 Sequencing
- MGI DNBSEQ Technology
- Singular Genomics G4X

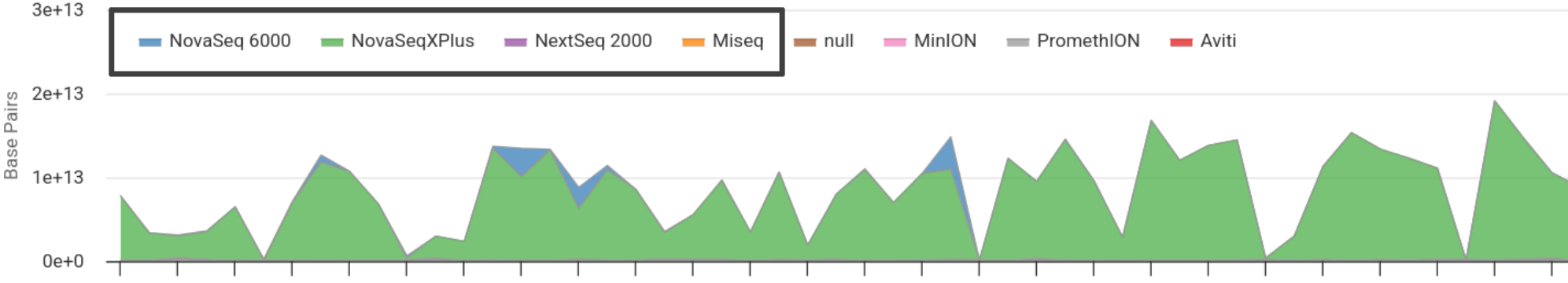
Direct DNA/RNA sequencing

- **Oxford Nanopore sequencing**

Platforms in **bold** are in use at the National Genomics Infrastructure

Illumina sequencing is *the* NGS sequencing platform





Illumina's sequencing by synthesis technology is NGI's bread-and-butter platform

Preparation for sequencing (in the lab)

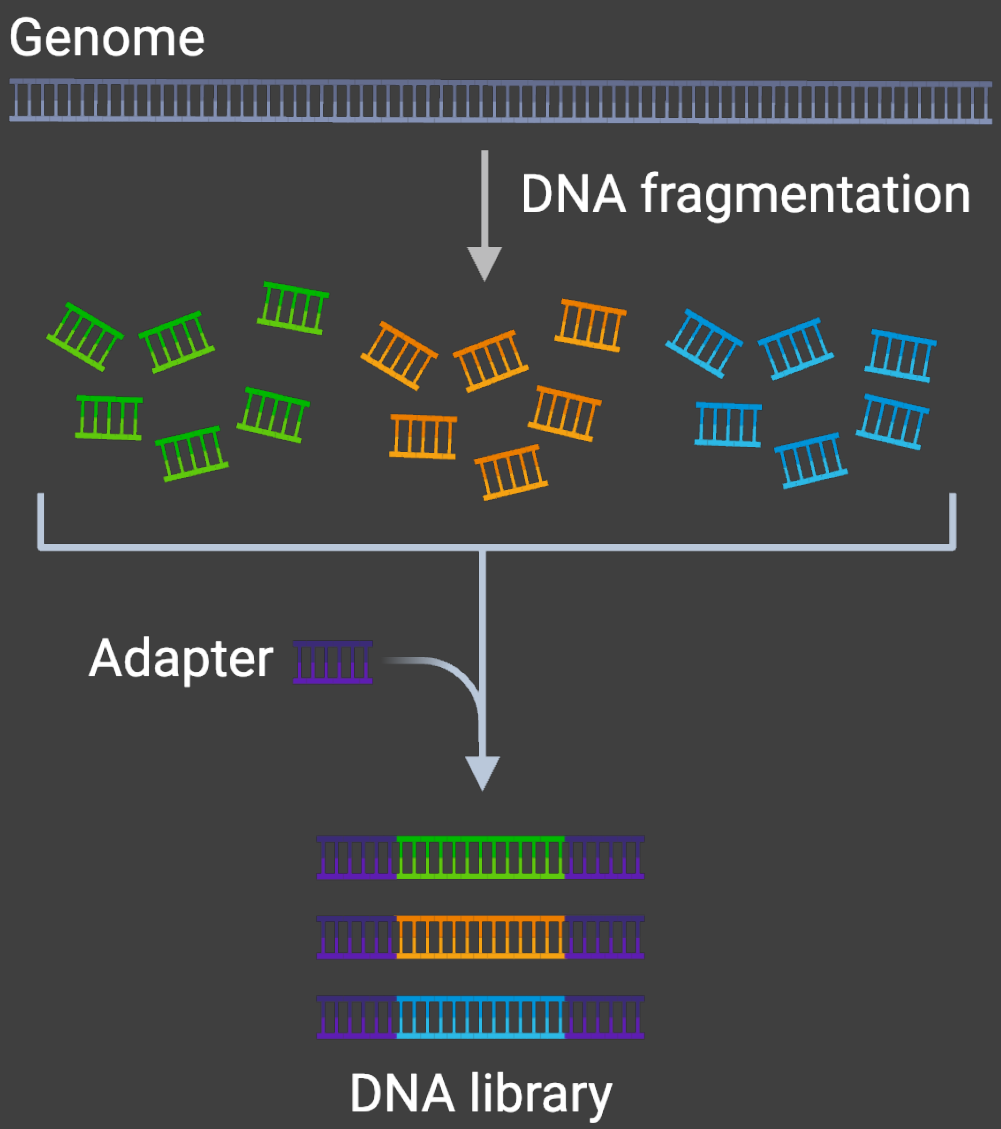
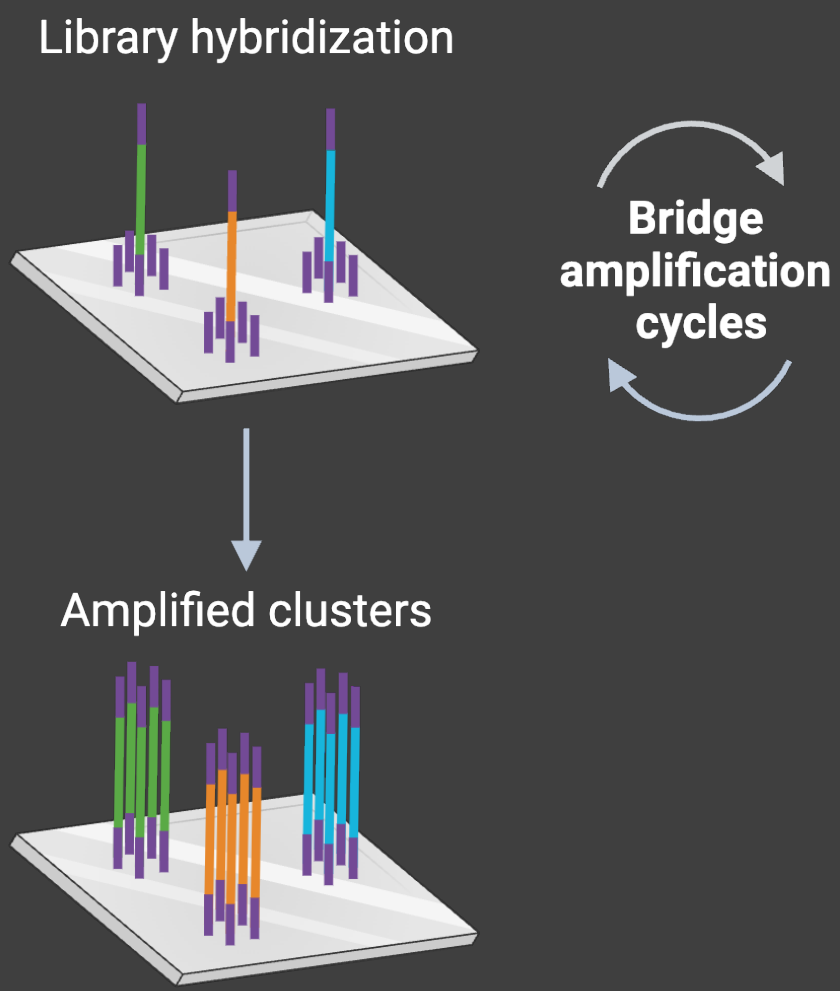


Figure by Anja Mezger

- Experiment-specific isolation, purification and enrichment
- Fragmentation
  - Enzymatic
  - Mechanic
- Addition of sequencing adapters
  - For attaching to the instrument
  - Barcodes for sample multiplexing
  - Unique Molecular Identifiers tag individual DNA fragments
- Pooling of samples

Preparation for sequencing (on the machine)



- Hybridize to a *flow cell*
- Amplify for better signal
- Linearize

Figure by Anja Mezger

Illumina: *Sequencing by Synthesis* of DNA clusters

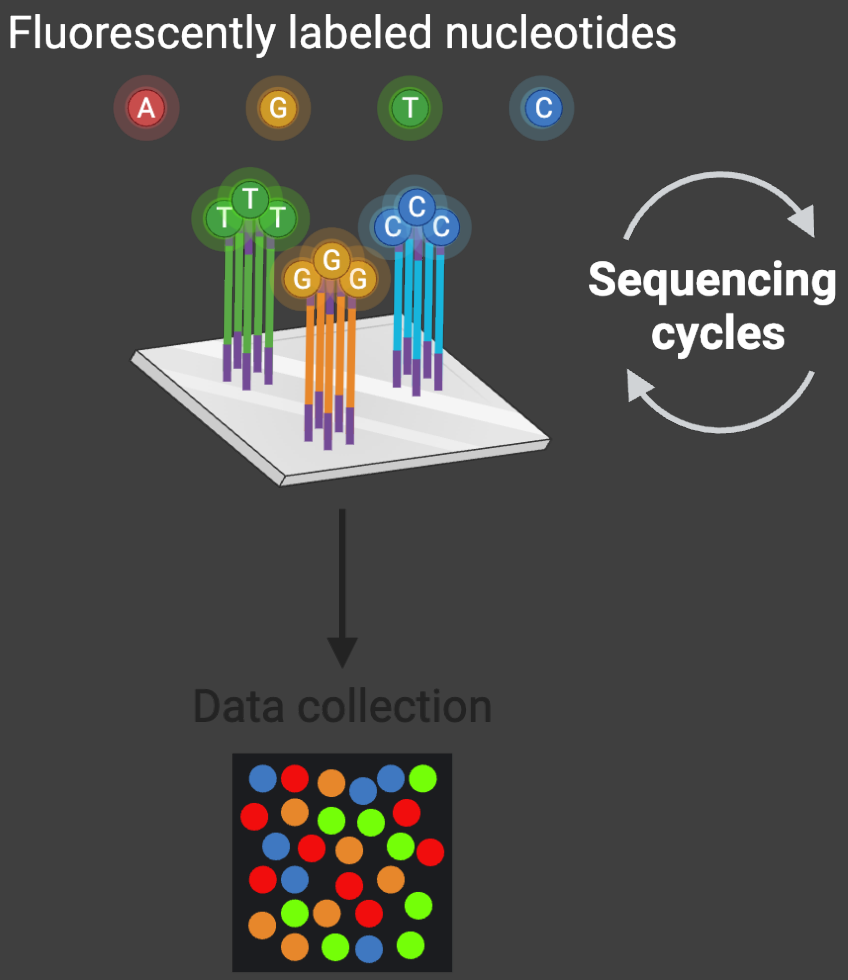
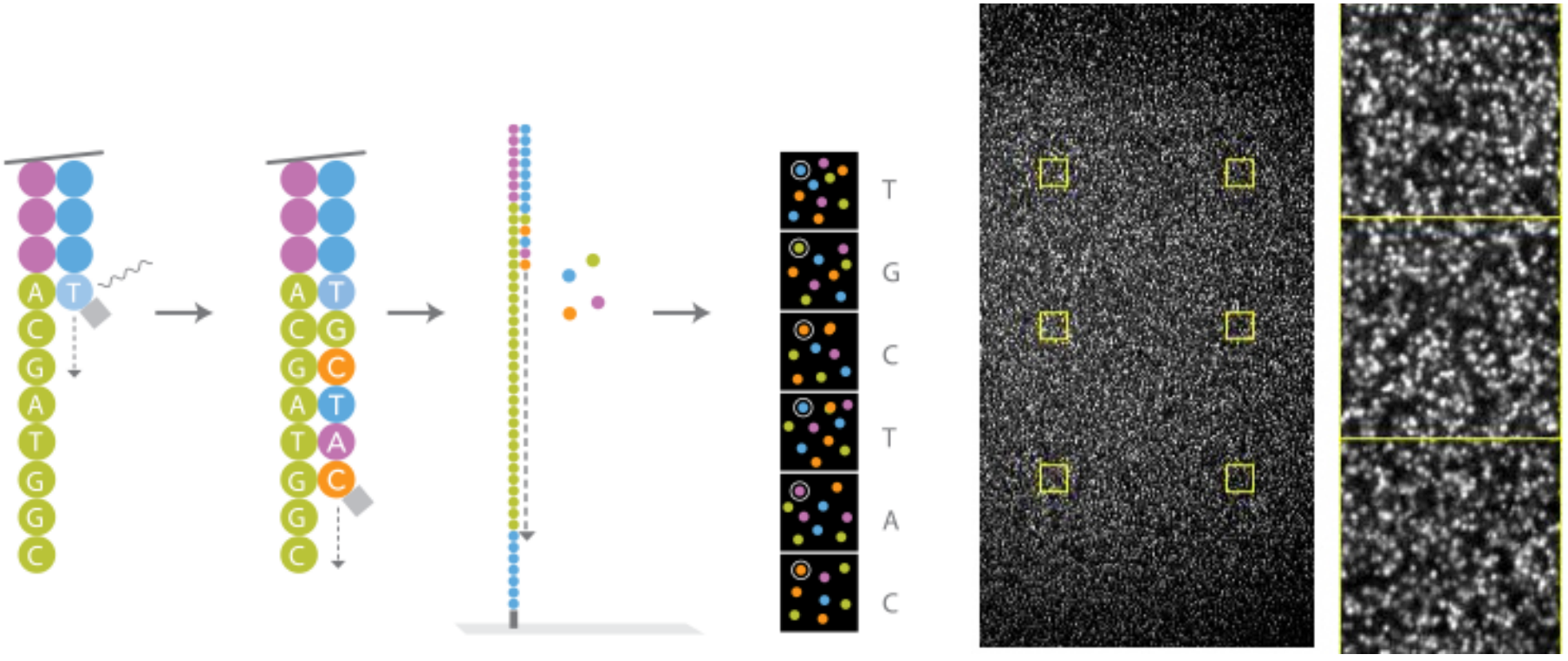


Figure by Anja Mezger

- DNA is amplified (again)
- Base integration yields a light signal (details vary among Illumina machines)
- Sequence is derived from a time-series of images

Illumina: *Sequencing by Synthesis* of DNA clusters



1. Integration of base is monitored directly
2. Image sequence is recorded
3. For each cluster, the light/dark pattern is converted into a DNA sequence

→ Highly parallelized, direct monitoring as synthesis proceeds

Flow cells instead of plates: Massive parallel sequencing

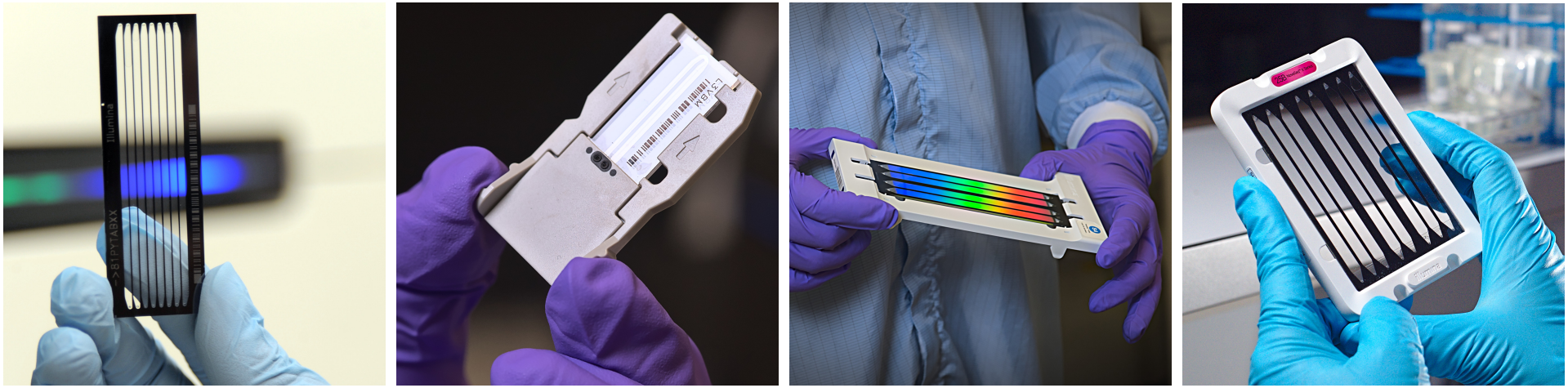
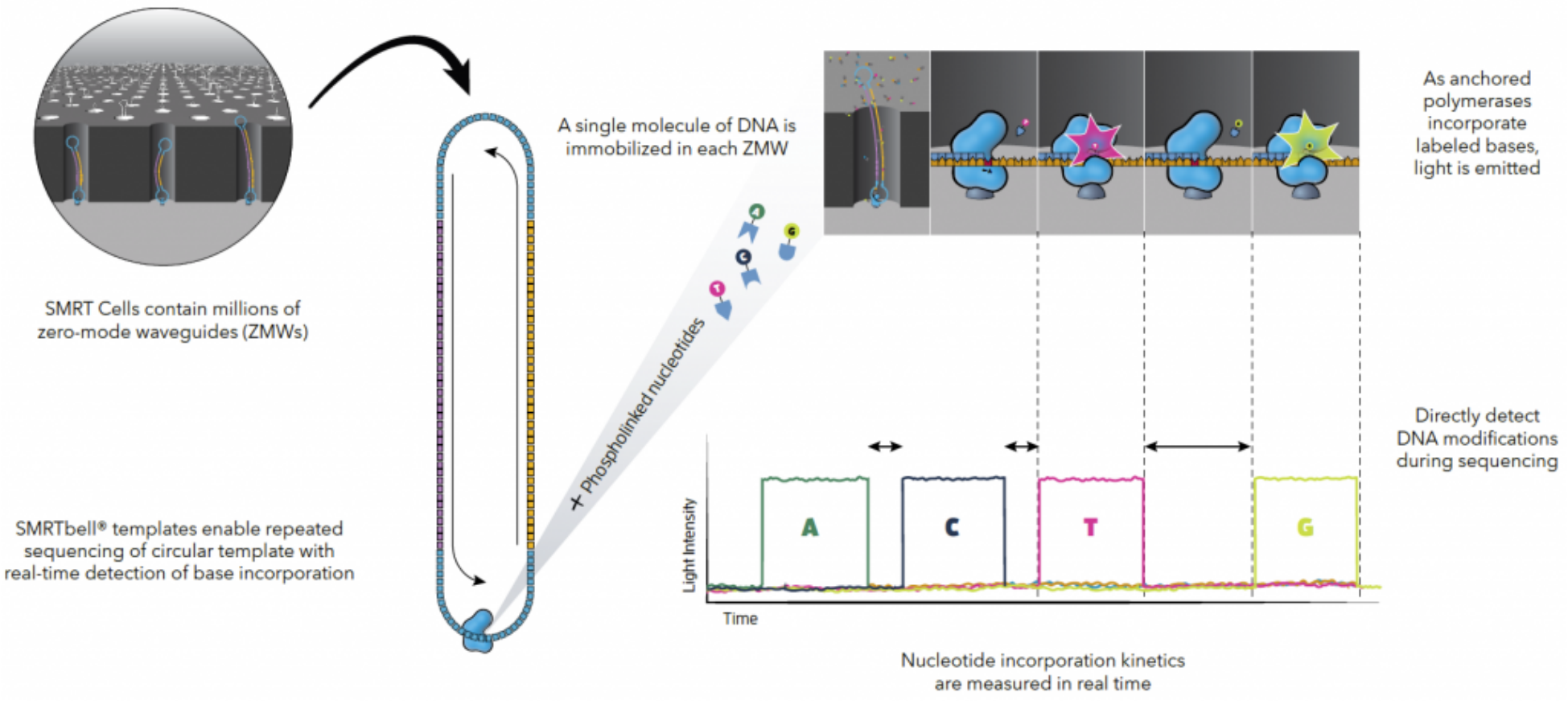


Figure 4: Various Illumina flow cells

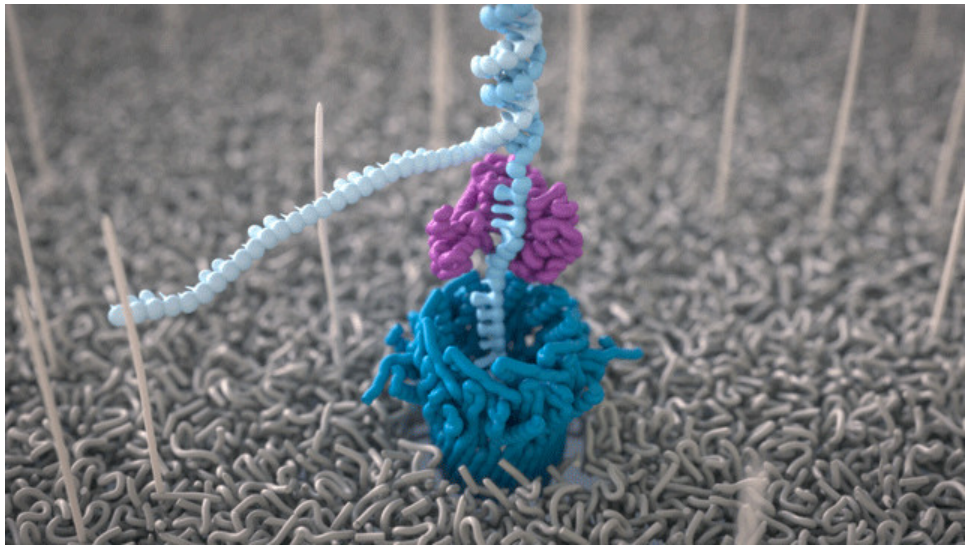
- Illumina's platform produces 2x 150bp reads from a fragment.
  → **short-read sequencing**
- Instead of 6912 fragments like with Sanger, Illumina machines can sequence Millions to Billions in parallel → **massive-parallel sequencing**

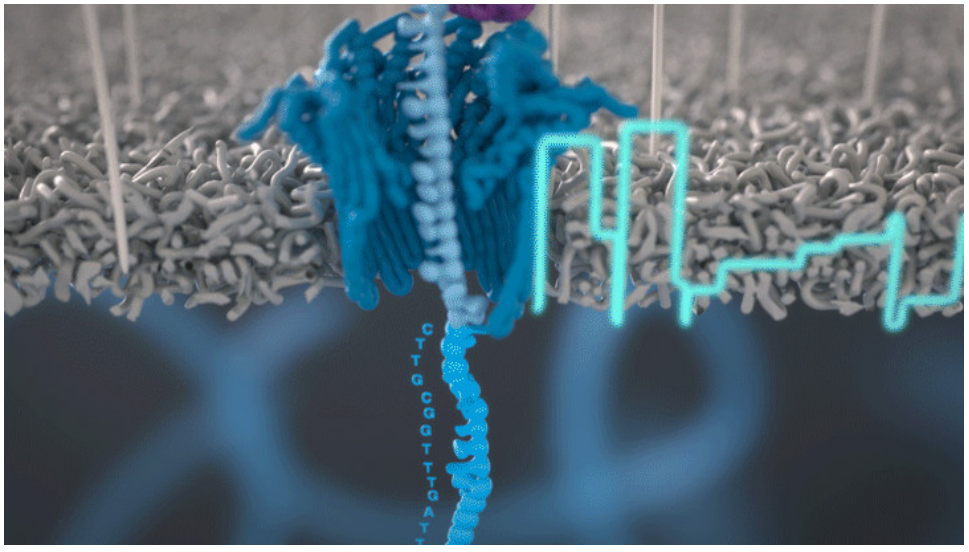
PacBio: Single-molecule sequencing by synthesis



1. PacBio can generate longer reads than Illumina.
2. Circular libraries, fragment is sequenced repeatedly.

Oxford Nanopore: Sequencing by electric conductivity





1. DNA is sequenced without amplification
2. A motor protein pulls a DNA strand through a pore (protein channel or solid state)
3. Bases cause specific conductivity changes
4. Direct reading of RNA and detection of methylated bases.

NGI provides sequencing platforms for every need

Standard

- Illumina sequencing

Longer reads, less base-call errors

- PacBio HiFi Sequencing

Much longer reads, many more base-call errors

- Oxford Nanopore sequencing

Short-reads, fewest base-call errors:

- Element Biosciences Avidite Sequencing

Sequencing data handling

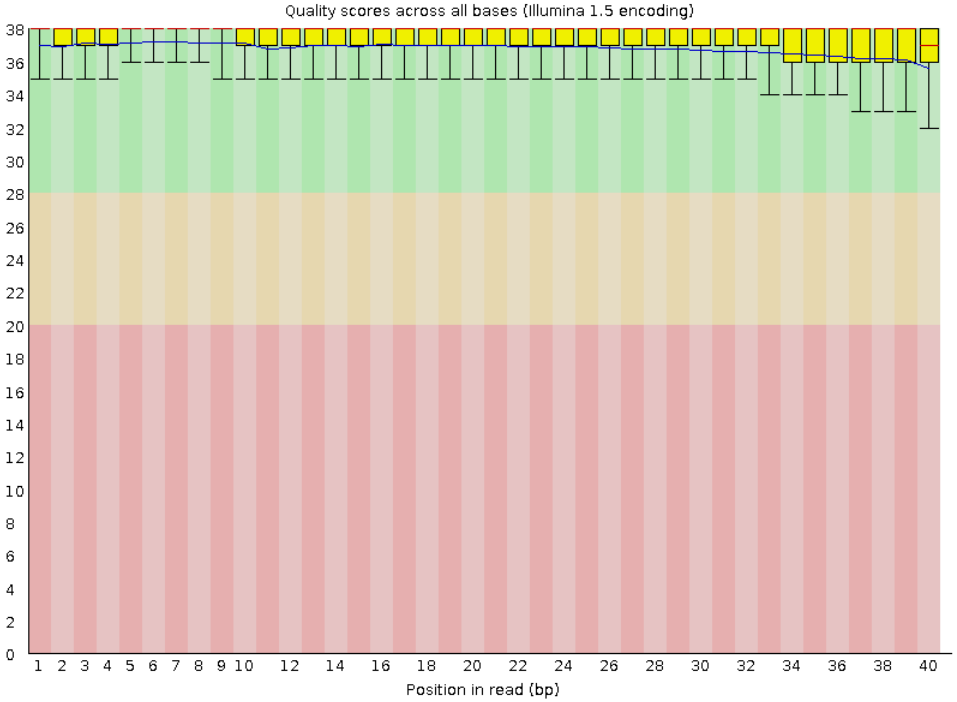
Sequencing result: Terabytes of data in FastQ-format

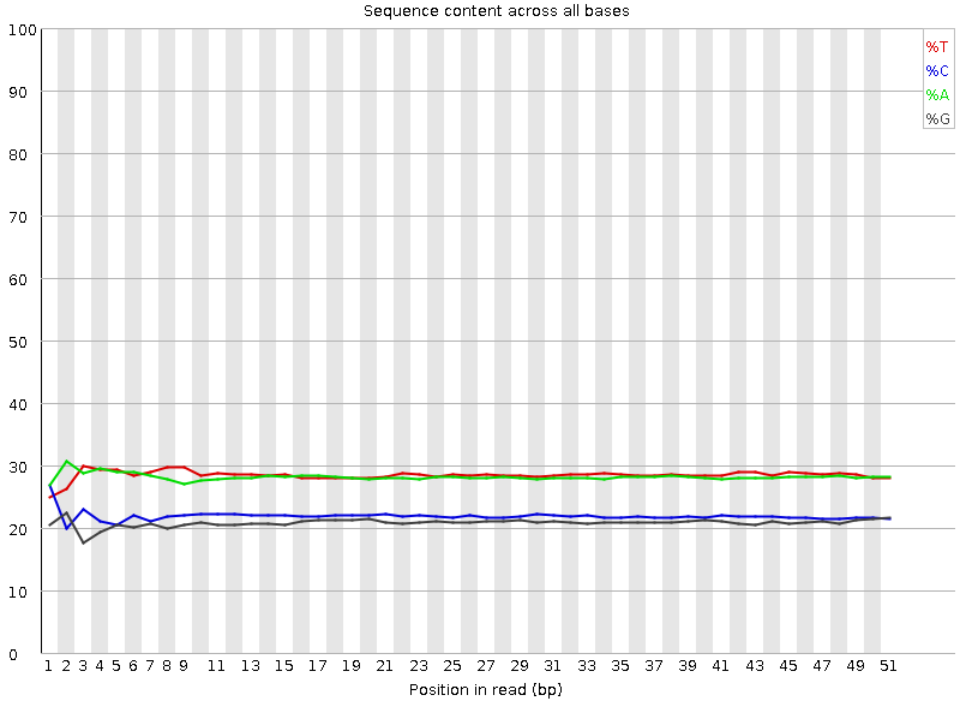
A single read:

1. Read ID
2. DNA-sequence
3. +
4. Error rate of the base call

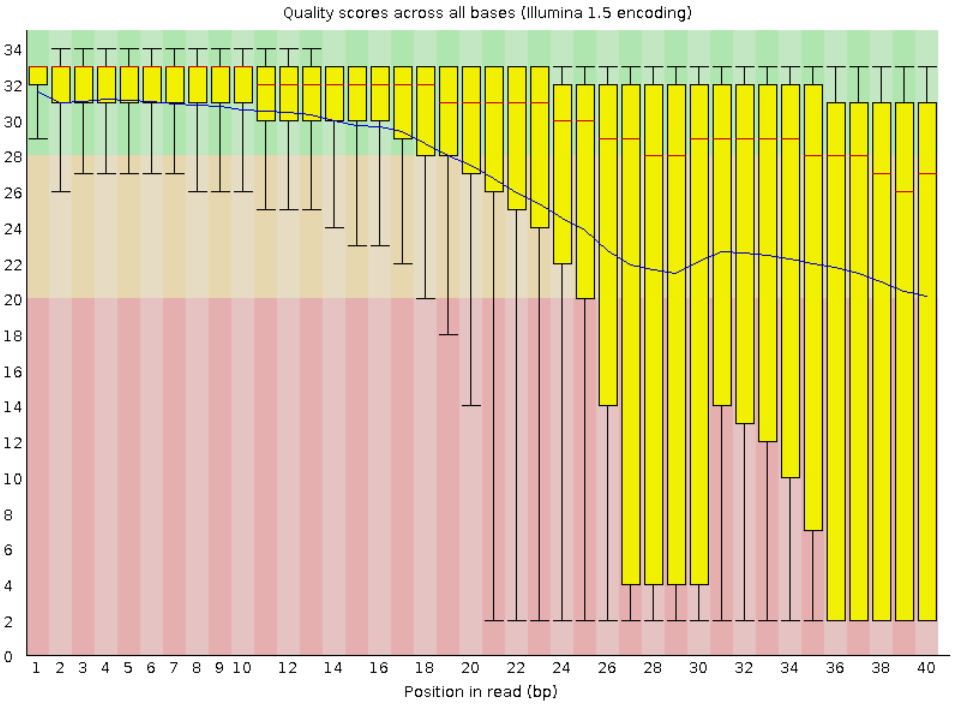
```
@M00463:56:000000000-AD76D:1:1101:15189:1873 1:N:0:1
ATAAACACGGTCTTTTTCCAGGTCAAGCCGGACGGTACCGCCCTGTGGCCATCGAA
+
-86<<F9FB7FFFGGGGGGFACGFFDEFGGGEFE>FGGGDGFGGFGGCFG?FF7E@
```

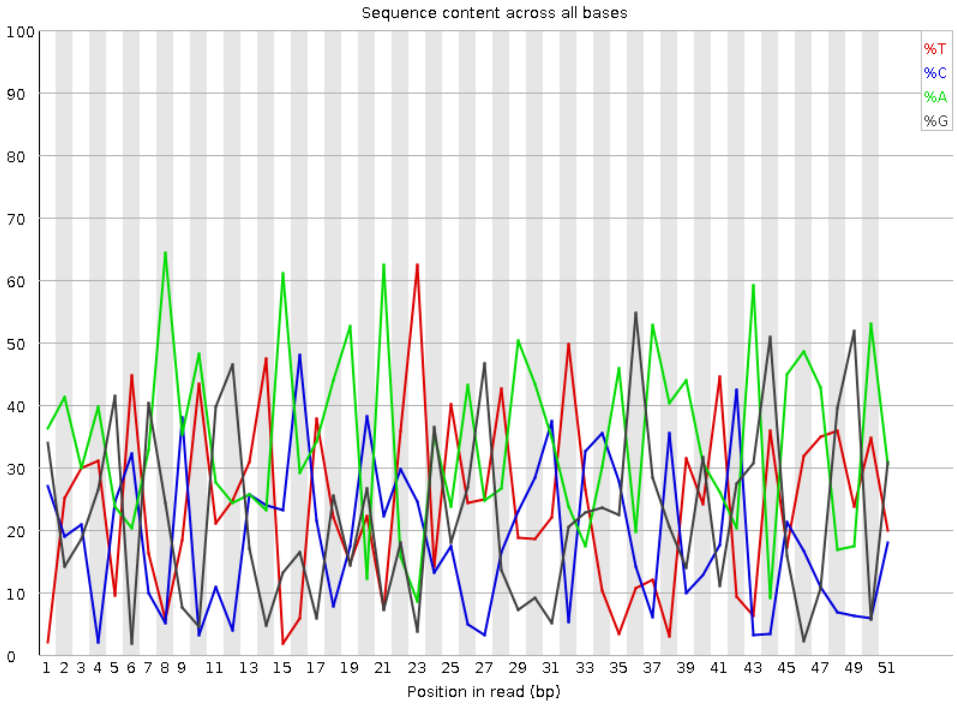
Quality control: Good data





Quality control: Poor data





Common bioinformatic analyses

```
>NC_001422.1 Escherichia phage phiX174
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTA
GATAAAGCAGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGGCGGAAAATGA
ATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGAT
TCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGC
GATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATTTTAAAAGAGCGTGGATTAC
TGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCGTA
TCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATC
CGATTTTCTGACGAGTAACAAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGTTGA
TGCGTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCCTGCTCCTGTTGAGTTTATTGCT
TCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAAT
GGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCG
CGCGCAGGAAACACTGACGTTCTTACTGACGCAGAAGAAAACGTGCGTCAAAAATTACGTGCGGAA
TGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGG
AAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTT
CCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTT
TCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTCCGGTTATCGCTGC
TCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGAC
CTGTAGACATTTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCA
GGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTT
GGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTTTCAGGGTTATTTGAATATCTAT
ACTATTTTAAAGCGCCGTGGATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATC
TCGTTATGGTTTCCGTTGCTGCCATCTCAAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGA
[...]
```

- **pairwise) Alignment:**
  Find the exact origin of a short fragment in a long reference.
- **Quasi-mapping:**
  Which reference is the most-likely origin?
- **De-novo assembly:**
  Create a long reference from short fragments.

Analyses as if we were back in the sixties

De novo assembly of *contigs* (fault-tolerant)

nswer, my friend

owin' in the wind. The ans

e amber my fr

Technical read error / mutation

y friend is blowin'

The answer is blowin' in

my friend, is blow

e wind. The answe

e answer, my fr

in the wind. The answer is blowin'

The answer, my friend, is blowin' in the wind. The answer is blowin' in the wind.

Alignment (fault-tolerant)

my vriend

Technical read error/ mutation

Theeeeeeeee answ*end

Indel

answer ⟷ answer

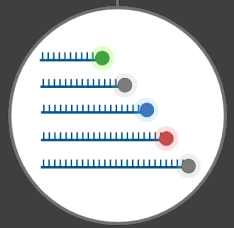
Multimapper

Weblinks

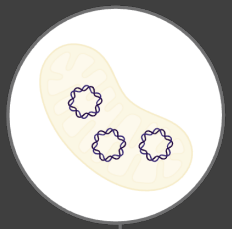
- Course Materials on sequencing data science
  http://data-science-sequencing.github.io
- DNA Sequencing Coursera class slides
  https://github.com/BenLangmead/ads1-slides
- Genome Browser (Easy access to selected genomes)
  http://genome-euro.ucsc.edu
- European Nucleotide Archive (Complete genomes and contigs)
  https://www.ebi.ac.uk/ena
- Current human reference genome (version 38)
  http://ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

Milestones in DNA sequencing history

1977 — Sanger method developed

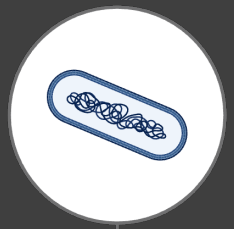


1981 — Human mitochondrial genome sequenced

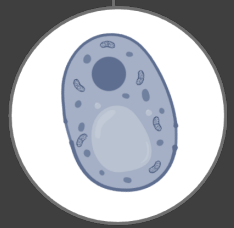


1990 — The Human Genome Project launched

1995 — Complete cell genome sequenced

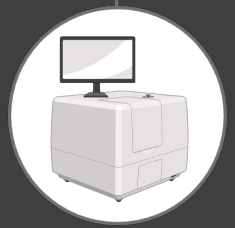


1996 — Complete eukaryotic genome sequenced



2003 — The Human Genome Project completed

2005 — First NGS technology released: 454 GS20

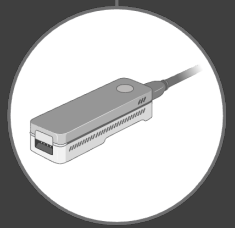


2006 — The Genome Analyzer launched

2007 — The Human Microbiome Project launched

2011 — PacBio sequencer released

2014 — Nanopore sequencing device released



2019 — The Human Microbiome Project completion reached

Figure by Anja Mezger

Sequencing applications

References

References i

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., ... Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome.. *Nature*, *409*, 860–921. https://doi.org/10.1038/35057062

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors.. *Proceedings of the National Academy of Sciences of the United States of America*, *74*, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., ... Zhu, X. (2001). The sequence of the human genome.. *Science (New York, N.Y.)*, *291*, 1304–1351. https://doi.org/10.1126/science.1058040