



# 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.

# 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

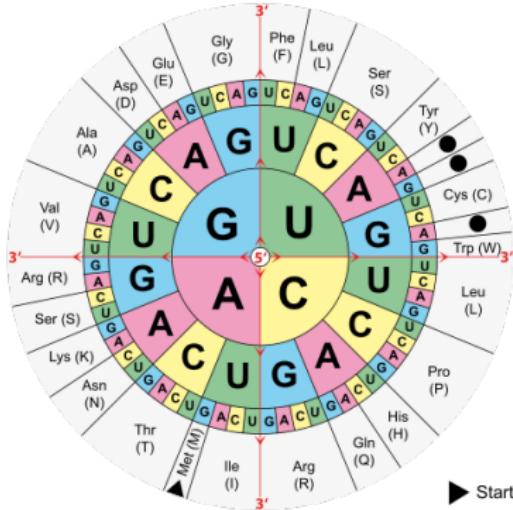
Prize share: 1/3



Photo from the Nobel Foundation archive.

Marshall W. Nirenberg

Prize share: 1/3



- The genetic code is (almost) universal<sup>[1]</sup>
  - 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

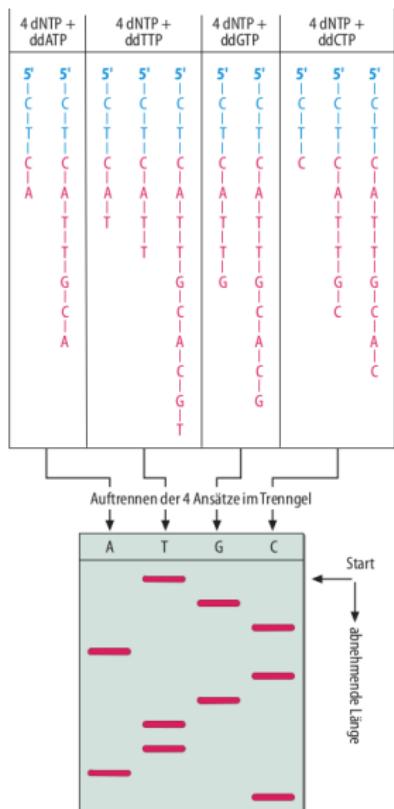
# Encoded information of naturally occurring DNA unknown



- 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.

*F. Sanger*

# 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  $\phi$ X174)

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

Medical Research Council Laboratory of Molecular Biology, Cambridge

Figure 2: [Sanger et al., 1977]

# 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**

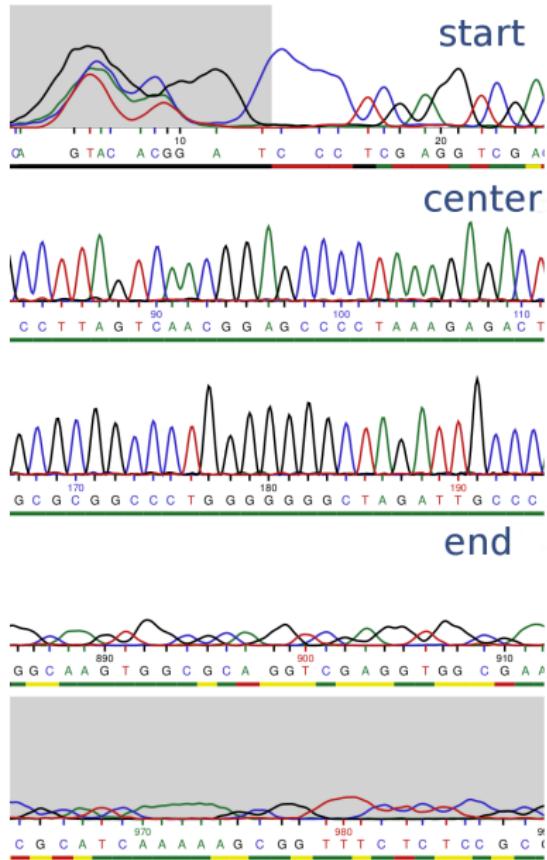
Prize share: 1/4

- Ample DNA input needed  
PCR was introduced in 1989
- Four reactions per sequence
- Read length  $\sim$  200bp



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

# Advanced Sanger sequencing for the Human Genome Project

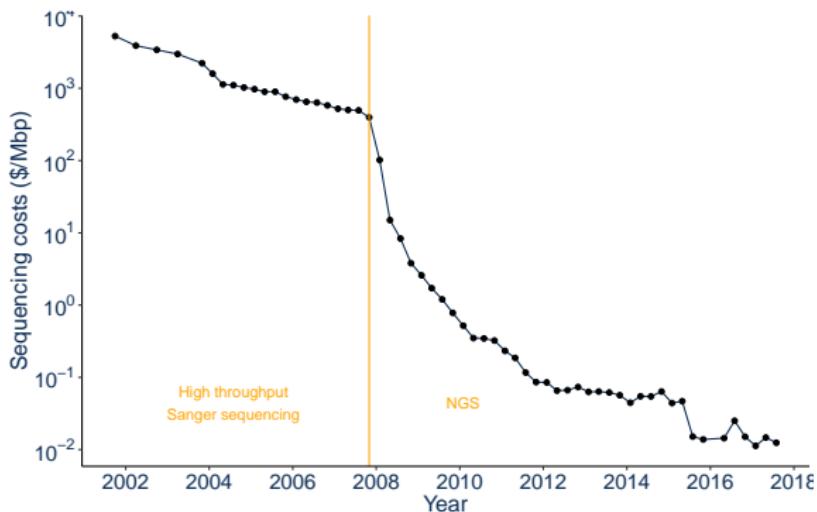


- Fluorescent chain terminators.
- Capillary electrophoresis for size separation of amplicons.
- Parallelized and automated.
- Sequencing technology of the Human Genome Project (1990-2004).

## **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: ~ 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**

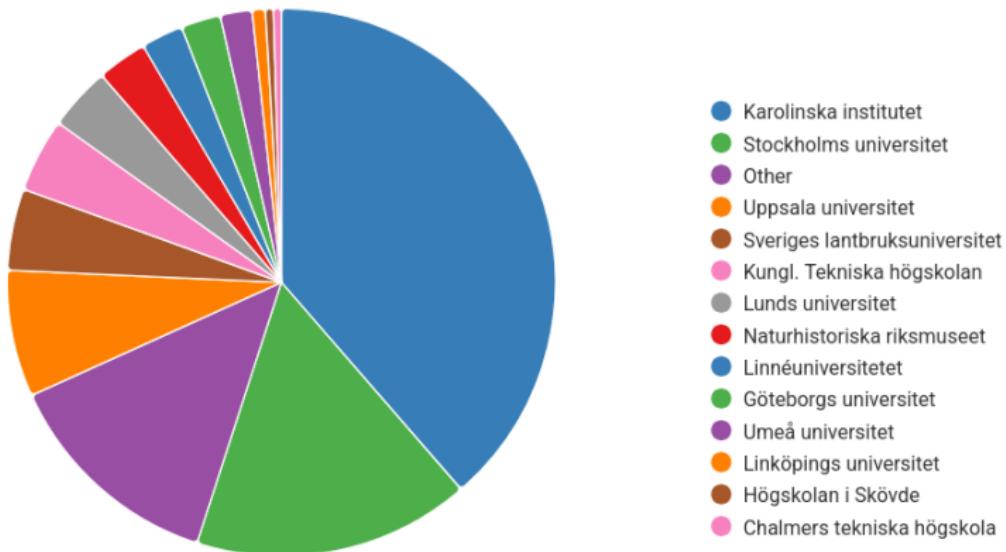
---

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.

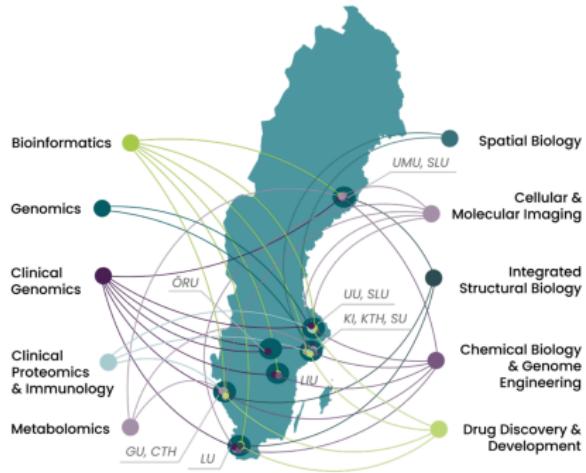
## Project Affiliations in 2024



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



- 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

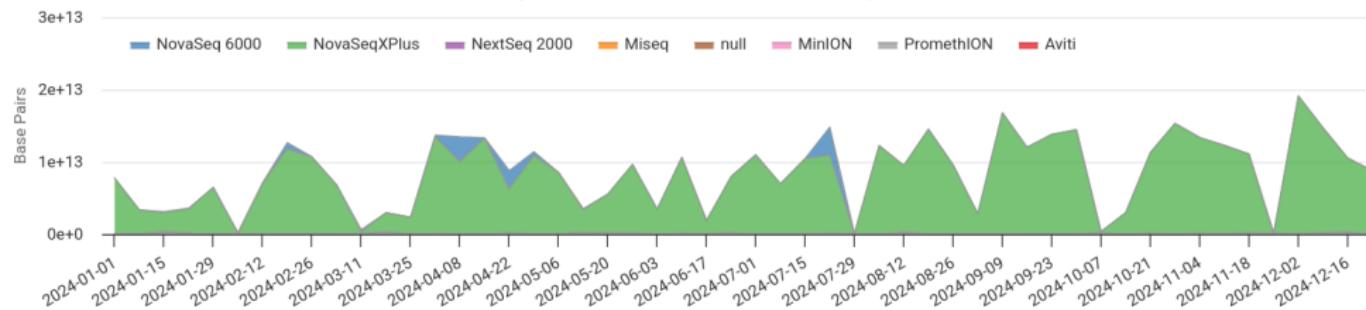


# NGI-S employs various sequencing technologies



## Sequencing Throughput

Average for 52 weeks: 1234 Gbp per day  
(1 Human genome equivalent every 3.77 minutes)



- 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

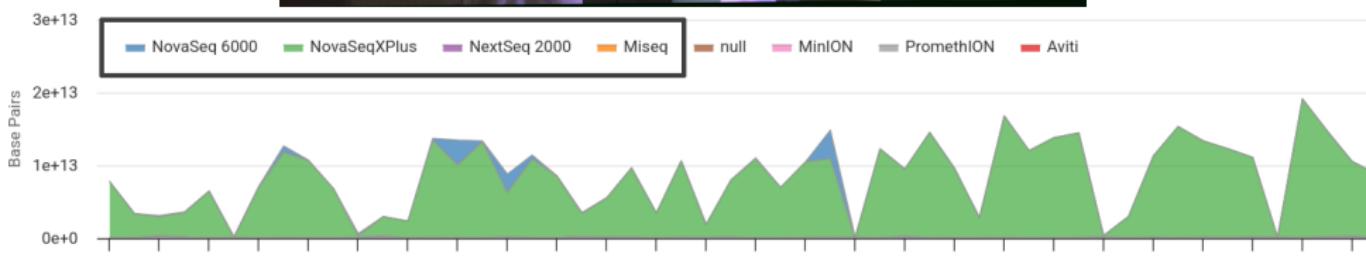
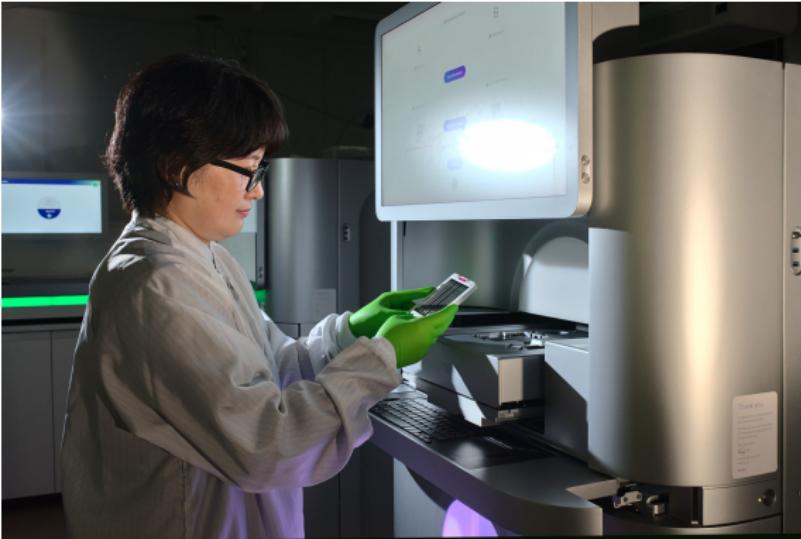
- Roche 454 sequencing (Pyrosequencing)
- Ion semiconductor sequencing
- **Illumina (Solexa) sequencing**
- **PacBio HiFi Sequencing**
- **Element Biosciences Avidite Sequencing**
- Ultima Genomics UG 100 Sequencing
- MGI DNBSEQ Technology
- Singular Genomics G4X

## 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)

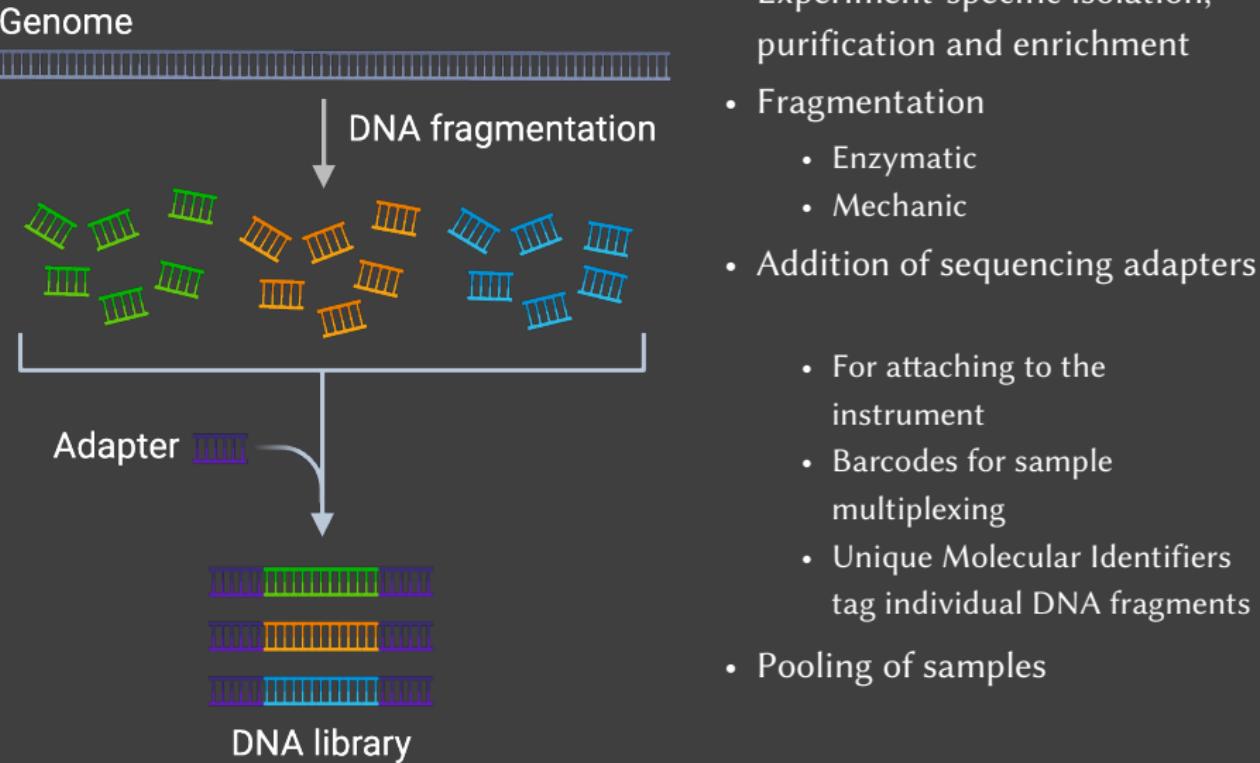


Figure by Anja Mezger

# Preparation for sequencing (on the machine)

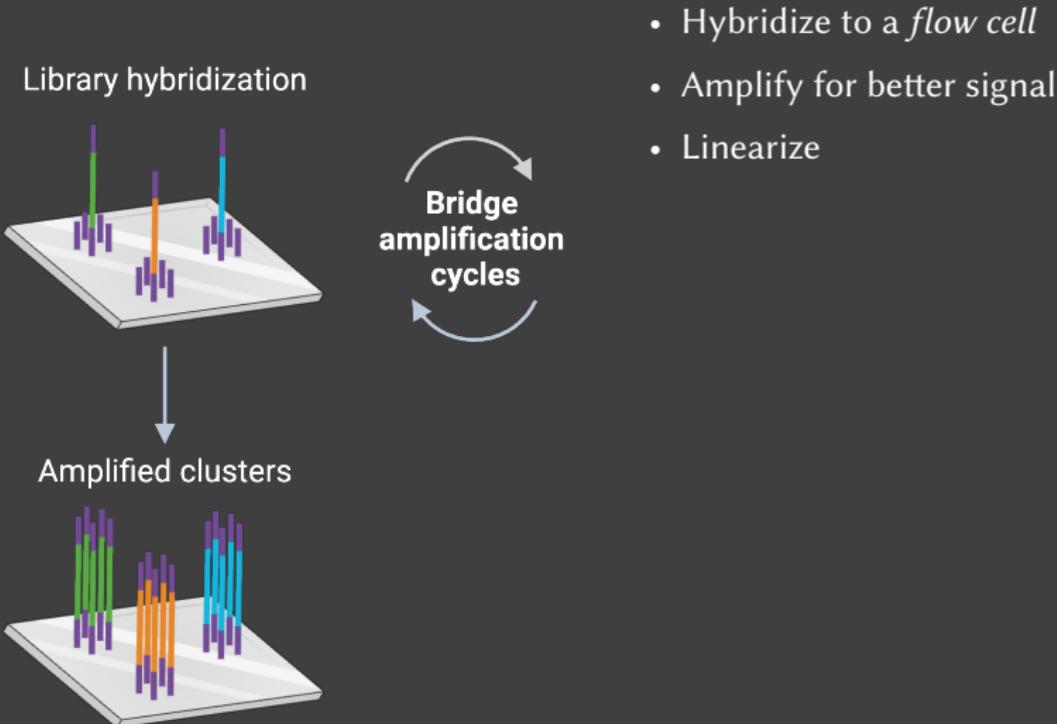
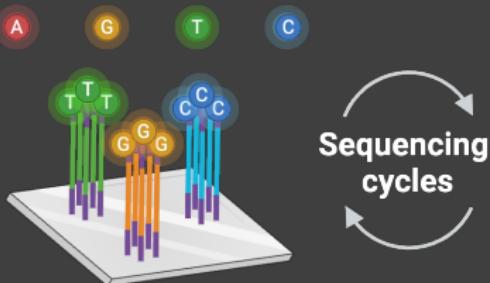


Figure by Anja Mezger

# Illumina: *Sequencing by Synthesis* of DNA clusters

Fluorescently labeled nucleotides



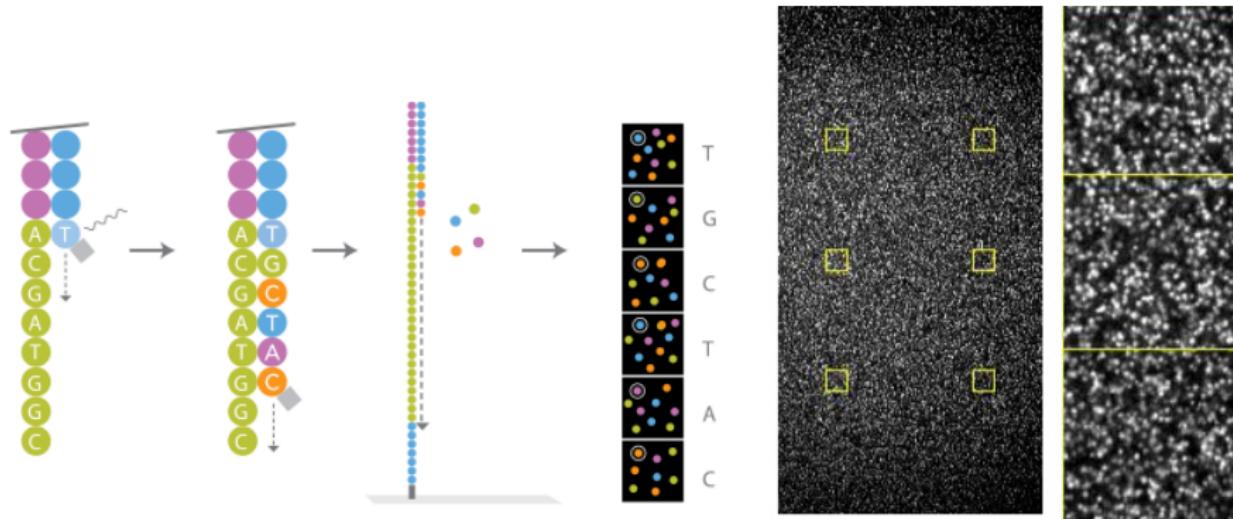
Data collection



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

Figure by Anja Mezger

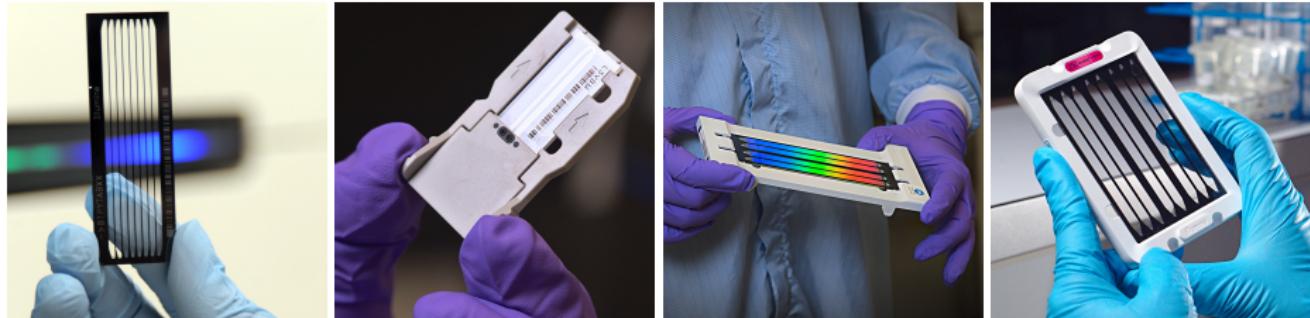
# Illumina: Sequencing by Synthesis of DNA clusters



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

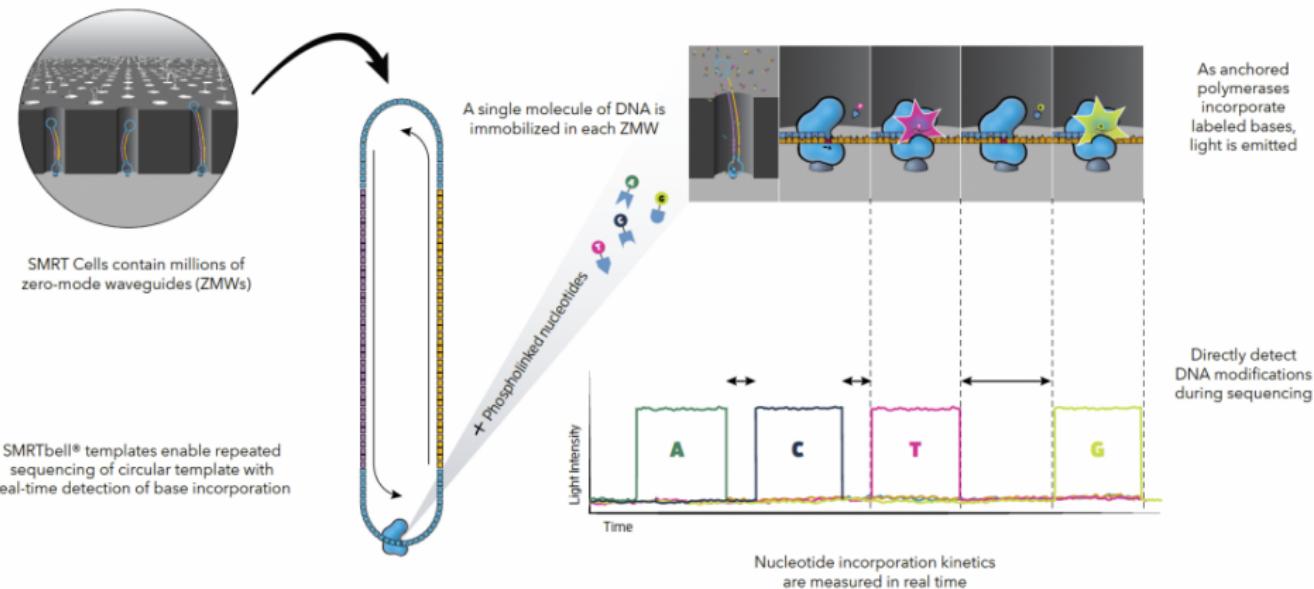
## Flow cells instead of plates: Massive parallel sequencing



**Figure 4:** Various Illumina flow cells

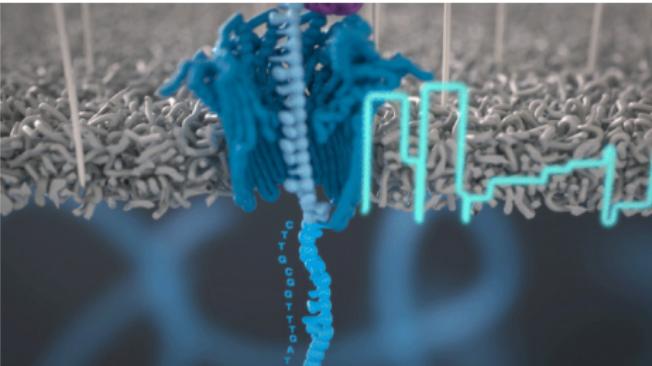
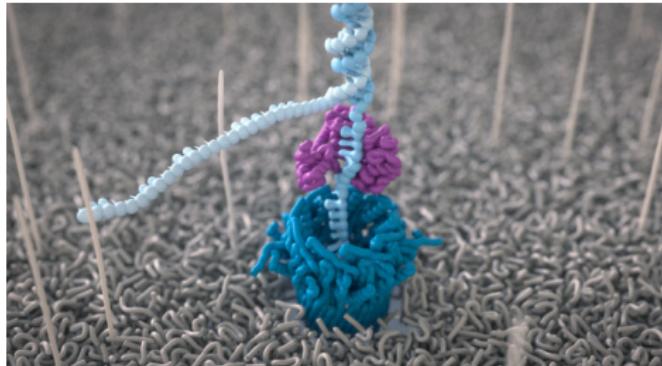
- Illumina's platform produces 2x 150bp reads from a fragment.  
→ **short-read sequencing**
- Instead of 6912 fragments like with Sanger, Illumina machines can sequence Millions to Billions in parallel → **massive-parallel sequencing**

# 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**

---

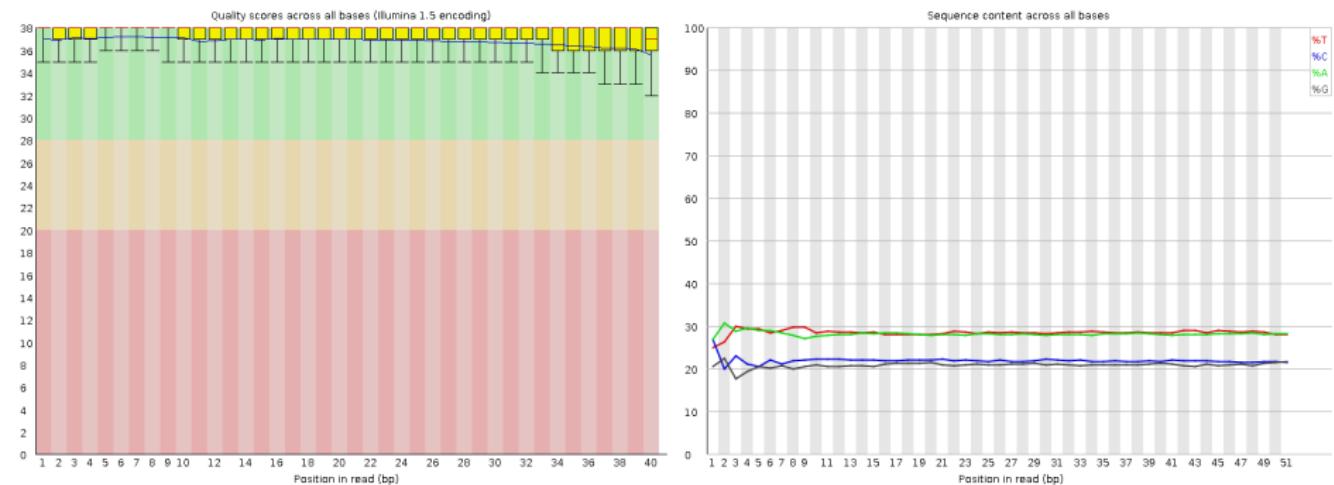
# Sequencing result: Terabytes of data in FastQ-format

## A single read:

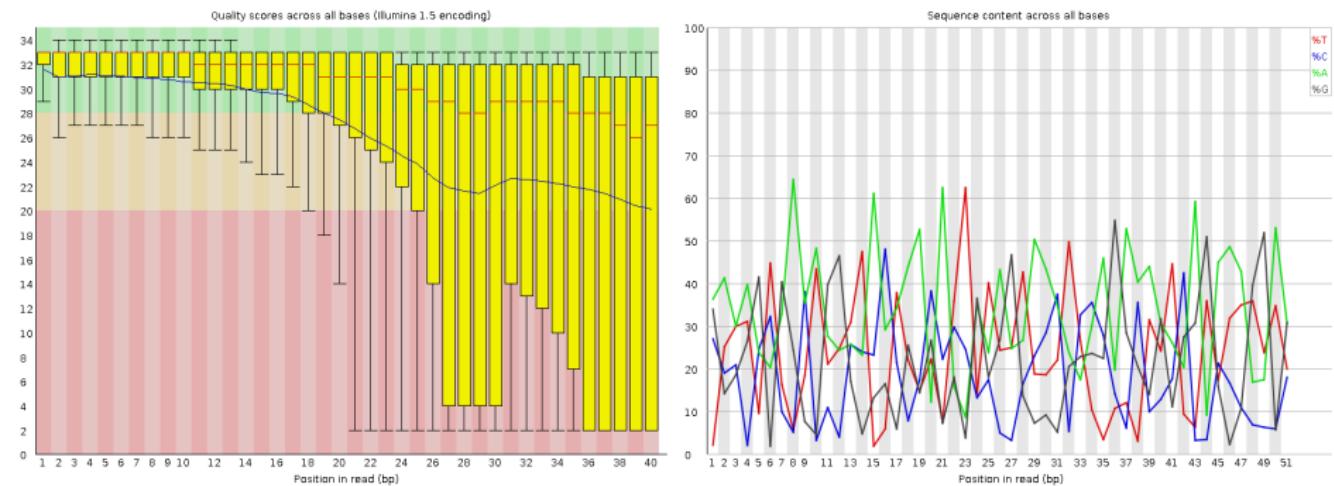
1. Read ID
  2. DNA-sequence
  3. +
  4. Error rate of the base call
- 

```
@M00463:56:000000000-AD76D:1:1101:15189:1873 1:N:0:1
ATAAACACGGTCTTTCCAGGTCAAGCCGGACGGTACCGCCCTGTGGCCATCGAA
+
-86<<F9FB7FFFGGGGGFACGFFDEFGGGEFE>FGGGDGFGFGGCFIG?FF7E@
```

# Quality control: Good data



# Quality control: Poor data



# Common bioinformatic analyses

```
>NC_001422.1 Escherichia phage phiX174
GAGTTTATCGCTTCATGACGAGAATTAACTCTGGATATTCGATGAGTCGAAAAATT/
GATAAAAGCAGGAATTAACTACTGCTTGTTACGAATTAAATCGAAGTGGACTCTGGCGAAAATG/
ATTCGACCTATCCTTGGCAGCTCGAGAAGCTTACTTGCACCTTCGCCATCAACTAACGAT
TCAAAAAACTGACCGGTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTCGTCAGGACTG
GATATGAGTCACATTGTTAGAGATTCTTGTGACATTAAAAGAGCTGGATTAA
TGAGTCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAGTTACTGAAACATCGT
TCCAGACCCTTGGCTCTATTAGCTCATCGGCTCTGCCGTTTGGATTAACCGAGATG
CGATTTCCTGACGAGTAACAAAGTTGGATTGACTCTGACCGCTCTGCTGCTGCTGGTTG
TGGCTTATGGTACGCTGGACTTGGACCTCCGCTTCCGCTCTGCTGCTGCTGGTTG
TCATGGCTTATTAGTGTACCCGTCAACCTAACAGCGCTGTCTCATGGAAAGCTGTAAT
GGAAAAACATTATAATGGCTCGAGCGTCCGGTAAAGCCGTAATTGTCGCTTACCTGGC
CGCGCAGGAAACACTGACGTTTACTGACGCGAGAAGAAAACGTCGCTCAAATTAACGTCGGA
TGATGTAATGCTCAAAGTAAAAACGTTCTGGCCTCGCCCTGGTGTCCGAGCCGTTGCGAG
AAAGCCAGCGTAAGGGCGCTCGCTTGGTATGAGGTGTAACAAATTGAGGGCTT
CCCTTACTTGAGGATAAATTATGCTAATATTCAAACACTGGCCGAGCGTATGCCGATGACCTT
TCTGGCTTCTGCTGGTCAAGATTGGCTCTTATTACCATTCACACTCGGTTACGTTGCTG
TCCTTGAGAGATTGACGCCGTTGGCGCTCTCGCTTCTCATGGCTGCGCTTGTATGAG
CTGAGACATTAACTTTTATGCTCTCATGCTCACGTTATGGTAACAGTGGATTAAAGTTG
GGATGGTGTAAATGCCACTCTCCCGACTGTTAACACTACTGGTATATTGACCATGCCGTT
GGCACGATTAAACCGTGTACCAATAAAATCCCTAACGATTTGGTTCAGGGTTATTGAAATCTAT
ACTATTAAAGGCCGCTGGATGCTGACCGTACCGAGGCTAACCTAACTGAGCTTAATCAAGATG
TCGTTATGGTTCCGTTGCTGCCATCTCAAAACATTGGACTGCTCCGCTTCTCTGAGACTG
[...]
```

- **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

Theeeeeeee answ\* end

Indel

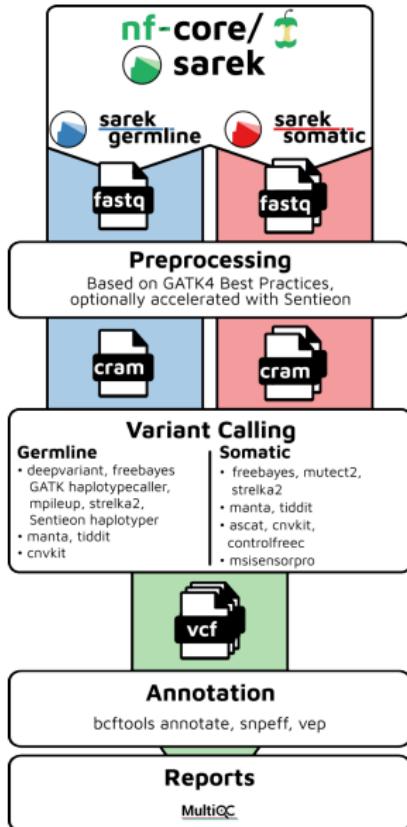
answer

←→

answer

Multimapper

# Analyses as if we are back today



- Data pipelines combine sequential steps.
- Workflow managers execute pipelines and scale analyses to many samples.
- Collaborative communities for workflow development help you to get started:
  - <https://nf-co.re>
  - <https://anvio.org>

## Weblinks

- Own lecture on NGS data analysis  
<https://github.com/MatthiasZepper/Lecture-OmicsDataAnlysis>
- 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/>

## Sequencing applications

---

# We are surrounded by genetic information



Library of the Human Genome, Wellcome Collection. Photo by Ben Gilbert and Thomas Farnetti

We are surrounded by genetic information: Many applications

**Science**

**RESEARCH ARTICLES**

**Three-dimensional intact-tissue sequencing of single-cell transcriptional states**

Xiao Wang<sup>1\*</sup>, William E. Alles<sup>1,2\*</sup>, Matthew A. Wright<sup>1,3</sup>, Emily L. Sylwestrak<sup>1</sup>, Nikolay Samusik<sup>1</sup>, Sam Vesuna<sup>1</sup>, Karun Chari Ramakrishnan<sup>1</sup>, Jia Liu<sup>1</sup>, Garry P. Nolan<sup>1,2</sup>, Felice-Alessio Rava<sup>1</sup>, Karl Deisseroth<sup>1,3,4,5</sup>

<sup>1</sup>Department of Bioengineering, Stanford University, Stanford, CA 94301, USA; <sup>2</sup>Stanford Bio-X, Stanford, CA 94301, USA; <sup>3</sup>Stanford University School of Medicine, Stanford, CA 94301, USA; <sup>4</sup>Stanford Neurosciences Institute, Stanford, CA 94301, USA; <sup>5</sup>Stanford Department of Psychology, Stanford, CA 94301, USA

**LETTER**

doi:10.1038/nature10554

**Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination**

Eduardo<sup>1</sup>, Zakeya Al Rasbi<sup>1</sup>, Marina Boziki<sup>1</sup>, Caroline Johner<sup>2</sup>, Hartmut Wekerle<sup>1</sup>

**Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing**

Yong-hui Jiang,<sup>1,18</sup> Ryan K.C. Yuen,<sup>2,18</sup> Xin Jin,<sup>3,4,5,18</sup> Mingbang Wang,<sup>3,18</sup> Nong Chen,<sup>3</sup> Xueli Wu,<sup>3</sup> ...  
Jing Wang,<sup>1,6,7,8,18</sup> ...  
Liu Li,<sup>1,9,10,11,12,13,14,15,16,17,18</sup> ...  
ARTICLE

doi:10.1038/nature10555

**A map of human genome variation from population-scale sequencing**

The 1000 Genomes Project Consortium\*

**A Draft Sequence of the Neandertal Genome**

Richard E. Green,<sup>1,2,†‡</sup> Johannes Krause,<sup>1,3,§</sup> Adrian W. Briggs,<sup>1,4,§</sup> Tomislav Maricic,<sup>1,4,§</sup> ...  
Emre S. Maenpaa,<sup>1</sup> Shahrooz Taheri<sup>2</sup>, Edward K. Gibbons<sup>3</sup>, Yanhong Li<sup>2</sup>, Peter D. Pääbo<sup>1</sup>, ...  
Anastasios Tsiplakos<sup>1</sup>, ...  
**Ancient human genome sequence of an extinct Palaeo-Eskimo**

Richness of human gut microbiome correlates with metabolic markers

Emmanuelle Le Chatelier<sup>1,\*</sup>, Trine Nielsen<sup>2,\*</sup>, Junji Qin<sup>3,\*</sup>, Edi Prifti<sup>1\*</sup>, Falk Hildebrand<sup>4,5</sup>, Gwen Falony<sup>4,5</sup>, Mathieu Mavingagu<sup>4,5</sup>, Anne-Sophie Artigues<sup>1,5</sup>, ...  
Jens Roat Kultima<sup>1</sup>, Daniel Kristoffer Forslund<sup>1</sup>, Shinichi Sunagawa<sup>1</sup>, ...  
Manimozhiyan Arumugam,<sup>1,2,3</sup> Athanasiios Tsiplakos<sup>1</sup>, ...  
Country-specific antibiotic use practices in the human gut resistome

Not a finite list, but let's tidy up

### **Applications that characterize genetic (mal)function**

- Gene expression / Transcriptomics (RNA-seq, CAGE-seq)
- Gene regulation / Epigenetics (DNA-Methylation, Histone modifications)
- Gene alterations (Hereditary diseases, cancer biology)

### **Applications that explore what is around us**

- Genome assemblies of other species (biodiversity)
- Metagenomics (environmental DNA)
- Pathogen surveillance (antibiotic resistance, epidemics)

### **Applications to elucidate evolutionary processes**

- Ancient genomes
- Population genomics

# Milestones in DNA sequencing history

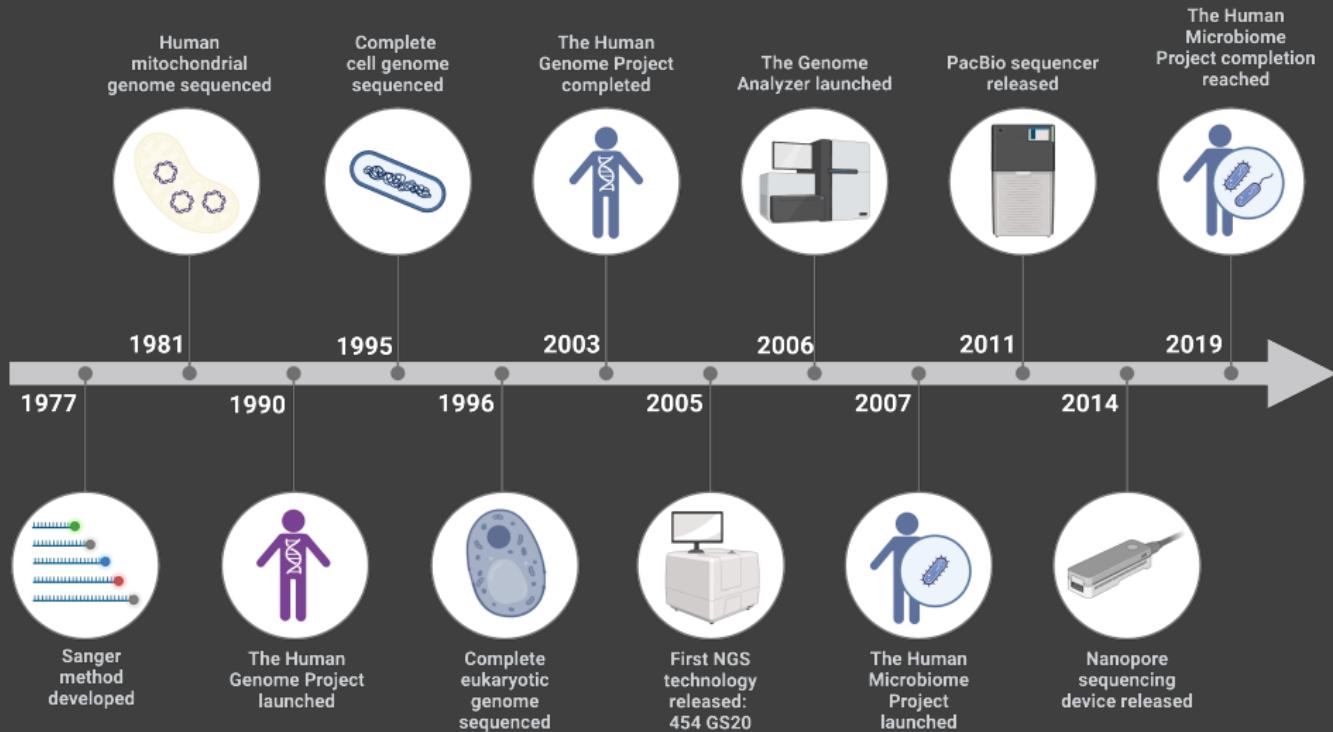


Figure by Anja Mezger

## 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>