



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

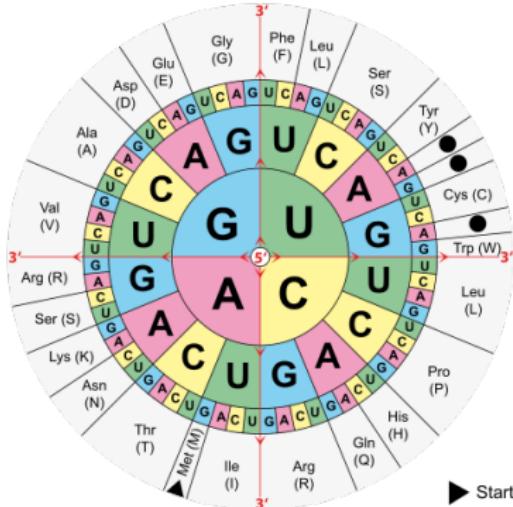
Prize share: 1/3



Photo from the Nobel Foundation archive.

Marshall W. Nirenberg

Prize share: 1/3



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

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

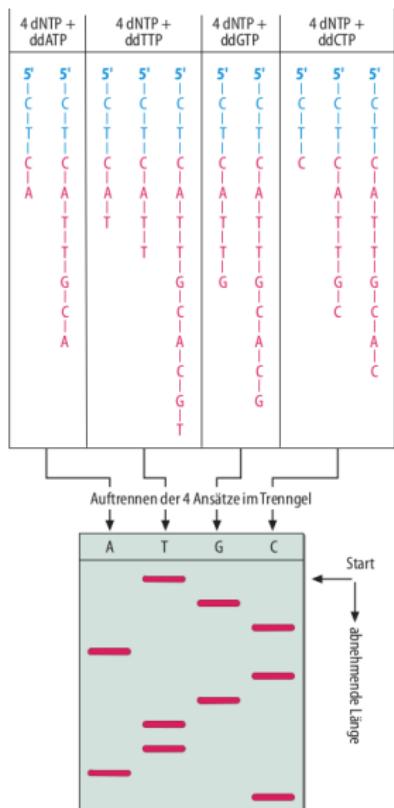
Encoded information of naturally occurring DNA unknown



- 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 ϕ 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

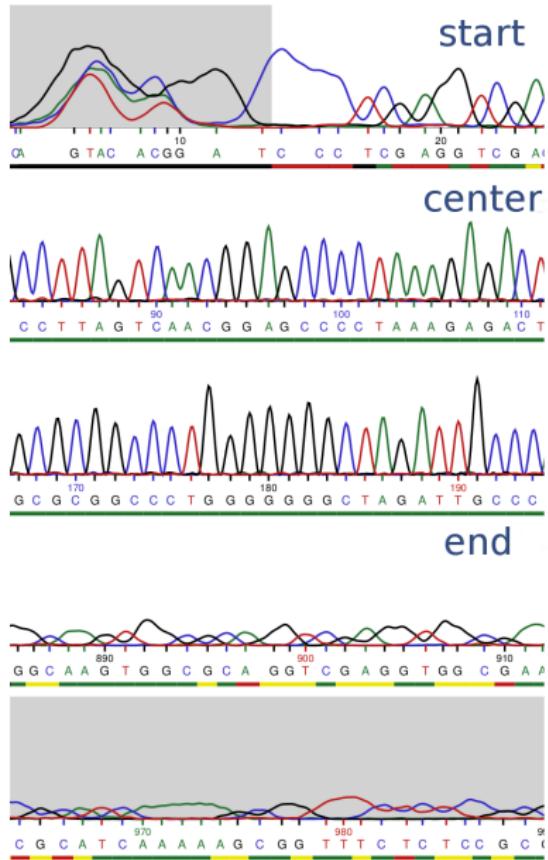
Prize share: 1/4

- 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

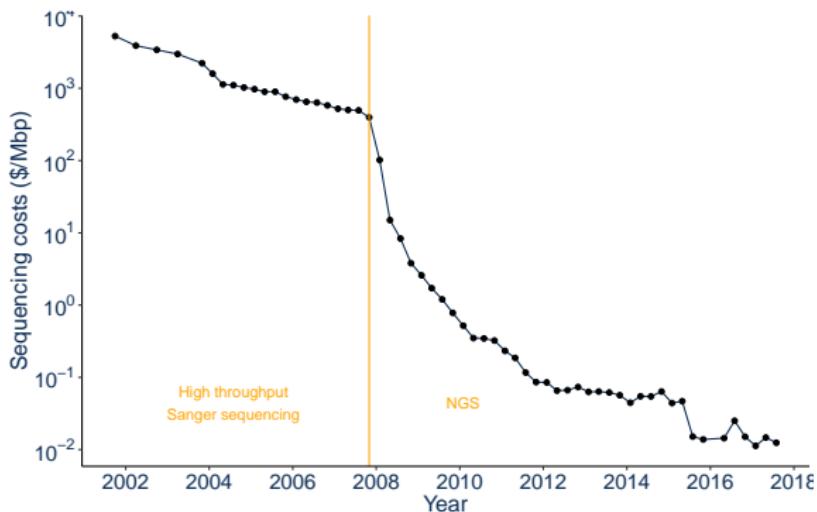


Figure 3: Sequencing costs per one million bases of raw sequence

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

2025: Sequencing of a human genome: \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

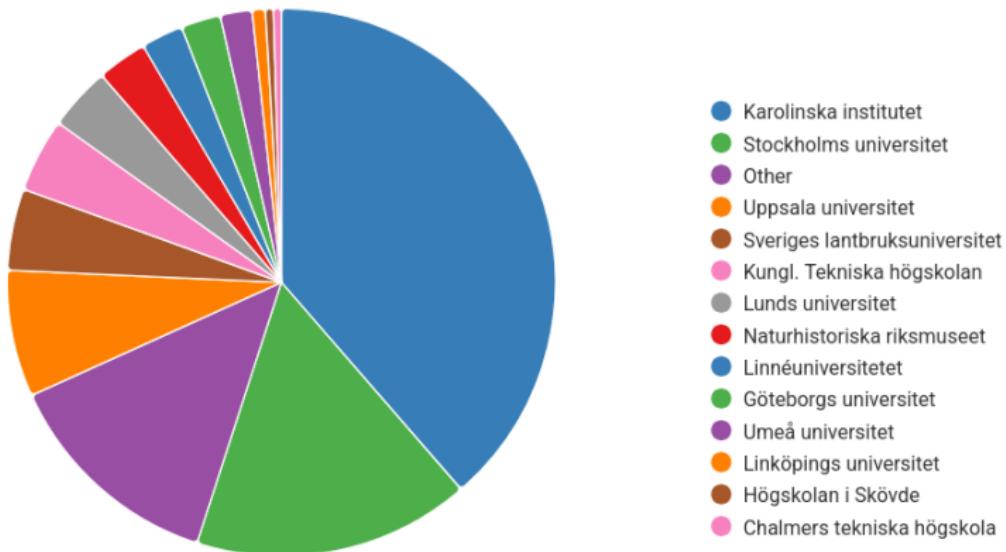
Sweden

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.

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

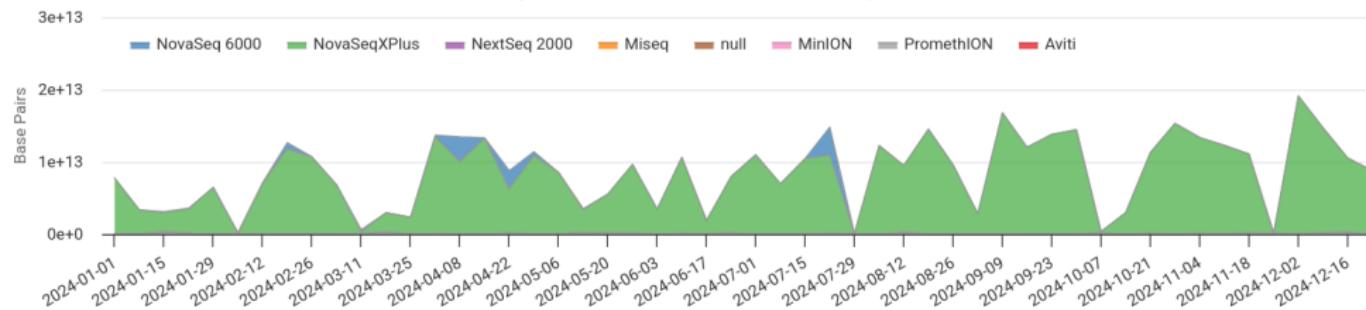
<https://ngisweden.scilifelab.se>

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

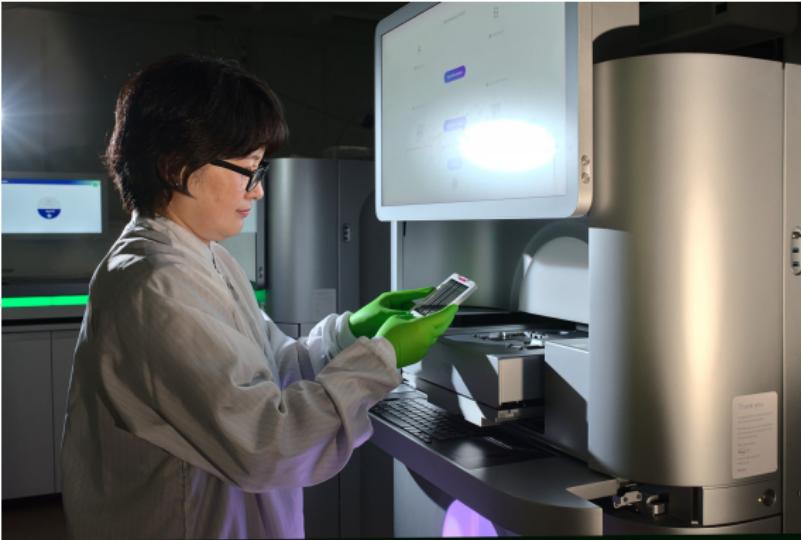
- Roche 454 sequencing (Pyrosequencing)
- Ion semiconductor sequencing
- **Illumina (Solexa) sequencing**
- **PacBio HiFi Sequencing**
- **Element Biosciences Avidite Sequencing**
- Ultima Genomics UG 100 Sequencing
- MGI DNBSEQ Technology
- Singular Genomics G4X

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)

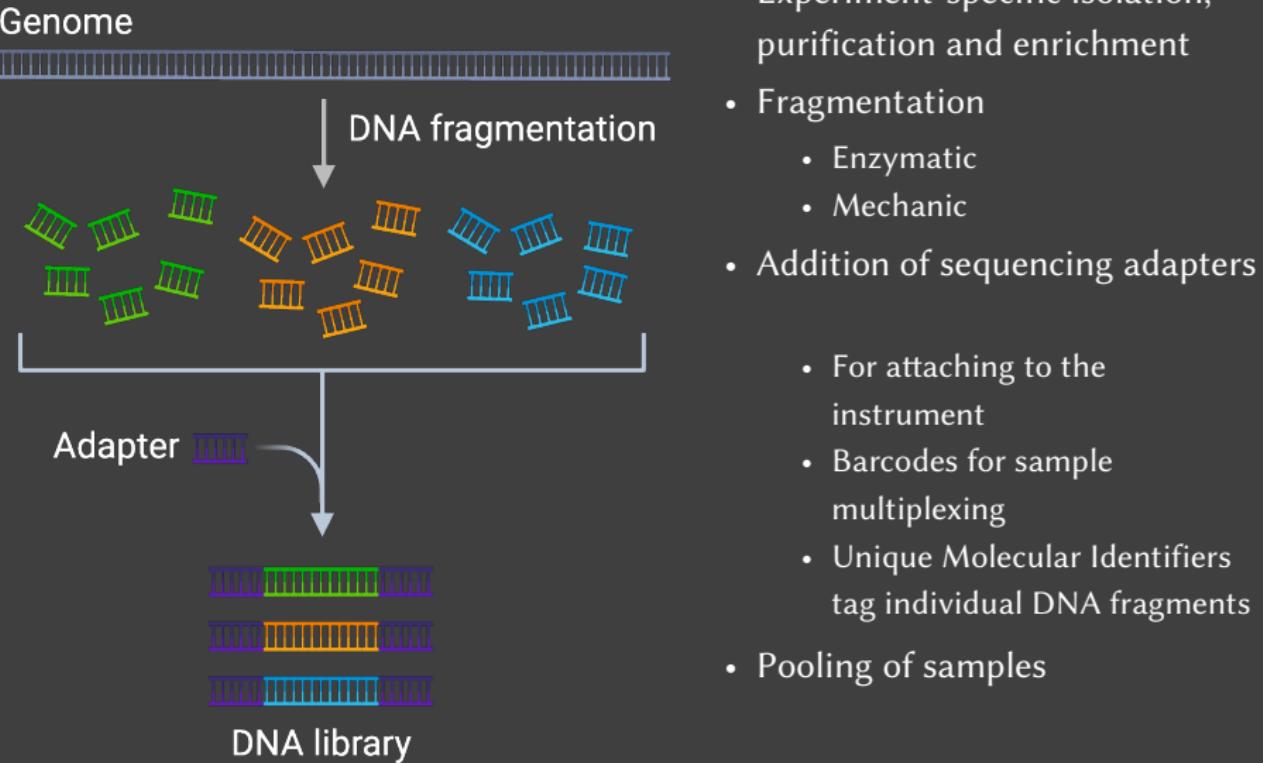


Figure by Anja Mezger

Preparation for sequencing (on the machine)

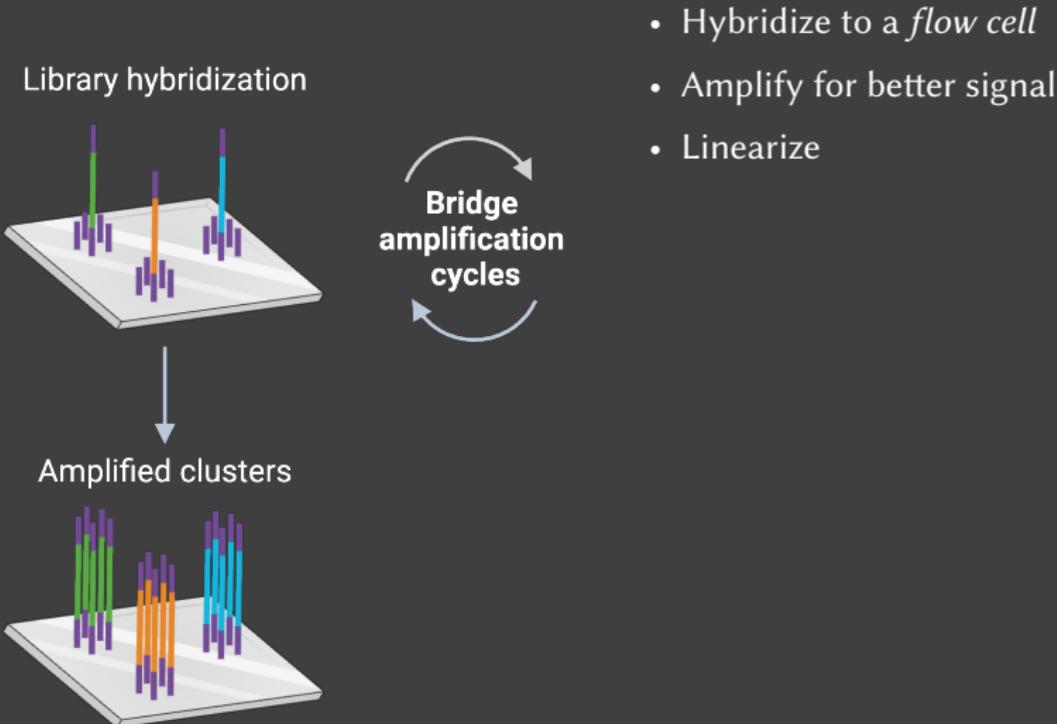
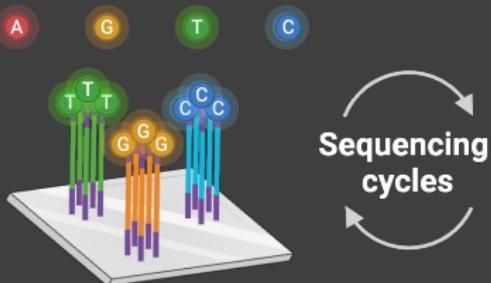


Figure by Anja Mezger

Illumina: *Sequencing by Synthesis* of DNA clusters

Fluorescently labeled nucleotides



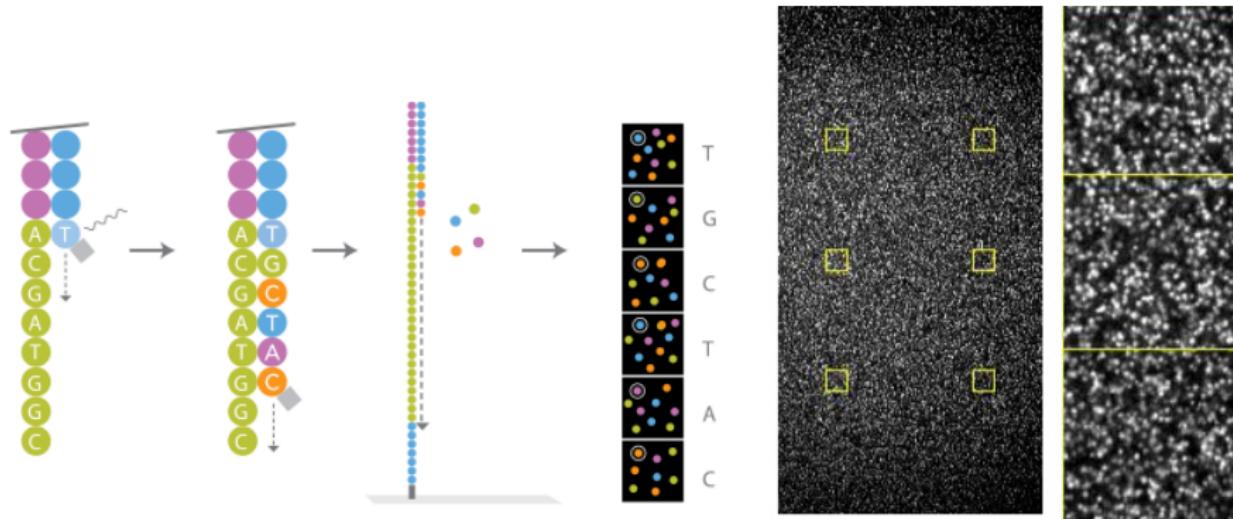
Data collection



- 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

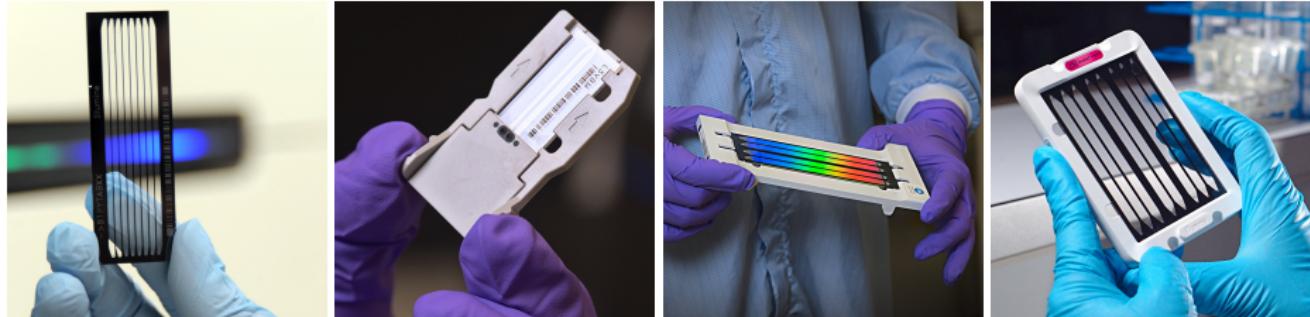
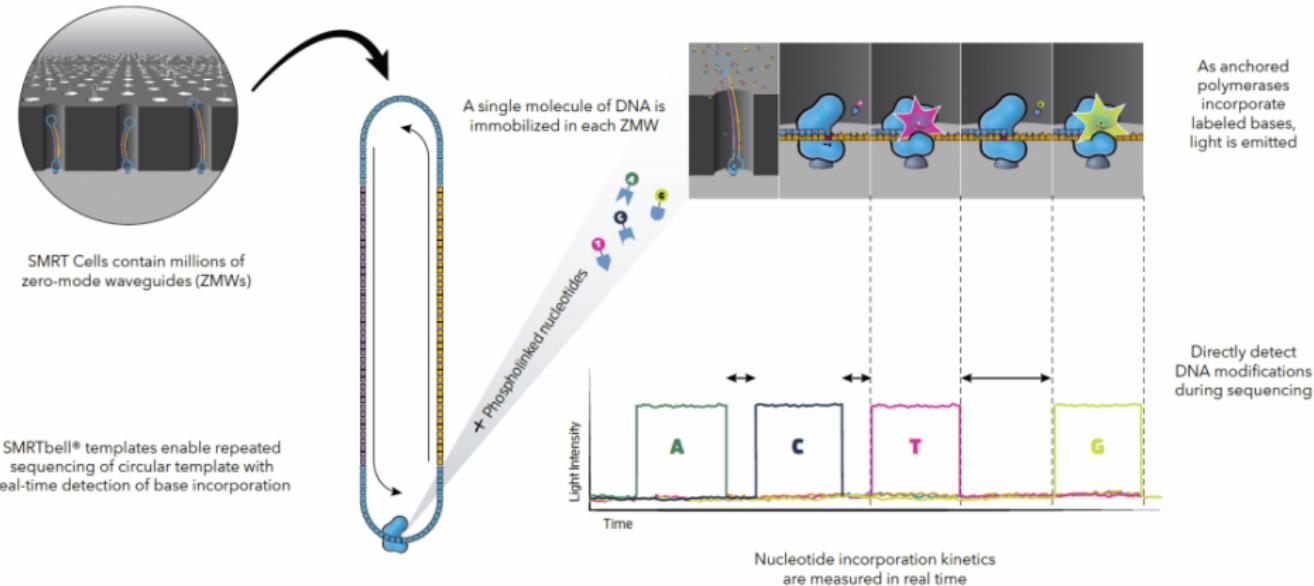


Figure 4: Various Illumina flow cells

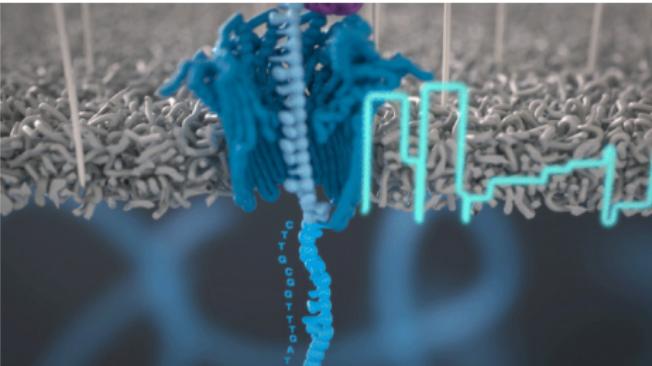
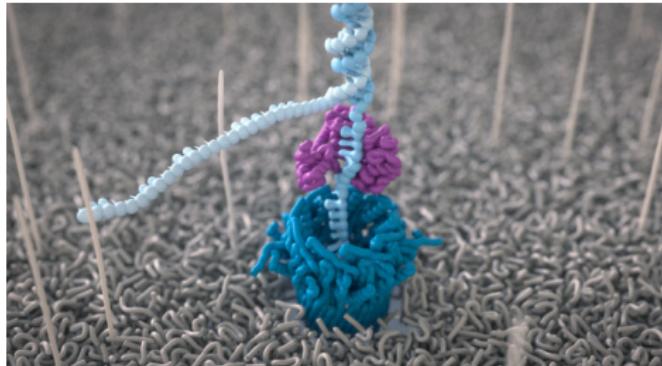
- Illumina's platform produces 2x 150bp reads from a fragment.
→ **short-read sequencing**
- Instead of 6912 fragments like with Sanger, Illumina machines can sequence Millions to Billions in parallel → **massive-parallel sequencing**

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

Milestones in DNA sequencing history

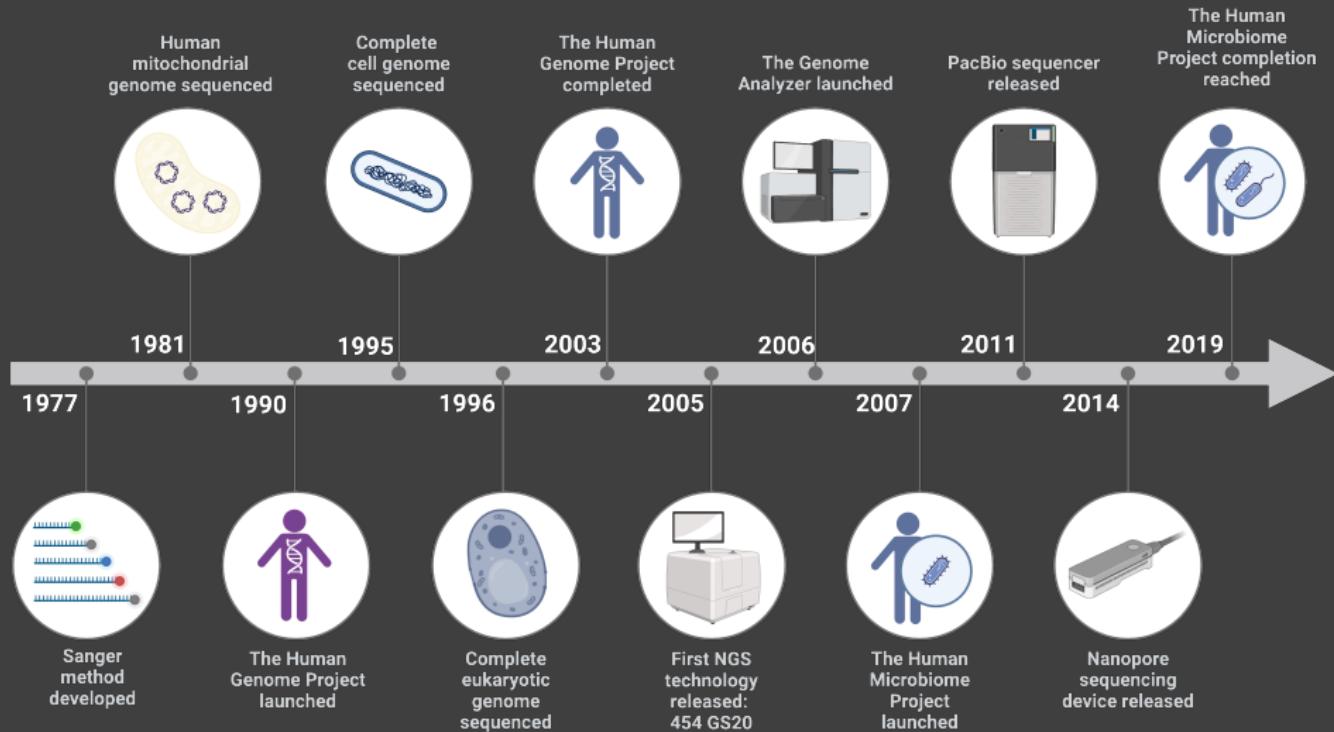


Figure by Anja Mezger

Ergebnis eines NGS Experiments

16	chr1	130313838	27	46=4X	*	0	0	GCTTCCGTAACG
0	chr11	87405681	32	4=4D46=	*	0	0	CGCGACCACTC
16	chr6	48944720	33	48=2X	*	0	0	AGGTGAGCCCCA
0	chr15	84753822	37	1X47=	*	0	0	CGGTTCCGACCC
0	chr15	84753820	33	1=2X47=	*	0	0	ACCGGTTTCGGAC
0	chr13	23667655	33	2=1I45=	*	0	0	CIGCICITTTTTC
16	chr1	38023342	33	47=2X1=	*	0	0	ACITCCAGCCCAT
0	chr19	5637750	27	4X44=	*	0	0	TCATTTCCCCCCTCIGGGGA
16	chr4	45120915	33	46=2X	*	0	0	TGIGCCITCCAAA
0	chr14	47993566	36	47=1X	*	0	0	AACCGGGCGACCC
0	chr7	148633542	33	2=2X46=	*	0	0	GCITCCITTCCTTTC
16	chr13	55122644	30	46=3X1=	*	0	0	TAACCGAAATCC
16	chr2	72817013	32	45=2X1=	*	0	0	TCCCHCCCCCCC
16	chr6	16848337	33	46=2X	*	0	0	ACCACAAAAAGTC
0	chr15	84527419	33	2=1I47=	*	0	0	CCGAGGGGAAAC
16	chr3	89810425	42	48=	*	0	0	TGAAACAAGACCC
0	chr11	72271813	33	2X46=	*	0	0	TGAACCTCCCGG
16	chr7	120998968	33	47=1I2=	*	0	0	AGGCAACCACAC
0	chr4	59048101	33	2X46=	*	0	0	CGACCCCCCATC
0	chr2	167514801	27	4X46=	*	0	0	ACCGACCCCCCCC
16	chr19	27753177	37	47=1X	*	0	0	CCCTATCCTTTC
16	chr12	81213953	42	48=	*	0	0	GAGTCCCCACAC
16	chr14	47441791	25	30=1X15=1X1=2X	*	0	0	CG
16	chr11	6344363	33	46=2X	*	0	0	CACTCCAGCCCCCCCCCCCC
0	chr2	72651687	37	1X47=	*	0	0	CTGGACATTCAT
16	chr5	101001181	36	46=1X1=	*	0	0	CAGAACCCGGTCC
16	chr5	101001181	28	46=1X1=2X	*	0	0	CAG
16	chr3	105762111	27	42=IX2=1X1=1X	*	0	0	CGG
0	chr5	147889151	25	1X1=2X31=1X14=	*	0	0	ACCC
0	chr13	89618690	42	48=	*	0	0	AAAGAACCTTAA
0	chr12	52449306	36	1=1X46=	*	0	0	CGGACCCCCGGTCC
0	chr11	74721981	37	1X47=	*	0	0	TAGCTTGTCGTC
0	chr10	74715842	30	3X47=	*	0	0	ACGGGGACCCG
0	chr10	5287032	30	3X47=	*	0	0	ACCCCCAAACCTAATAACGGT
0	chr3	96554980	37	1X47=	*	0	0	TGCTCAAACATA

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>