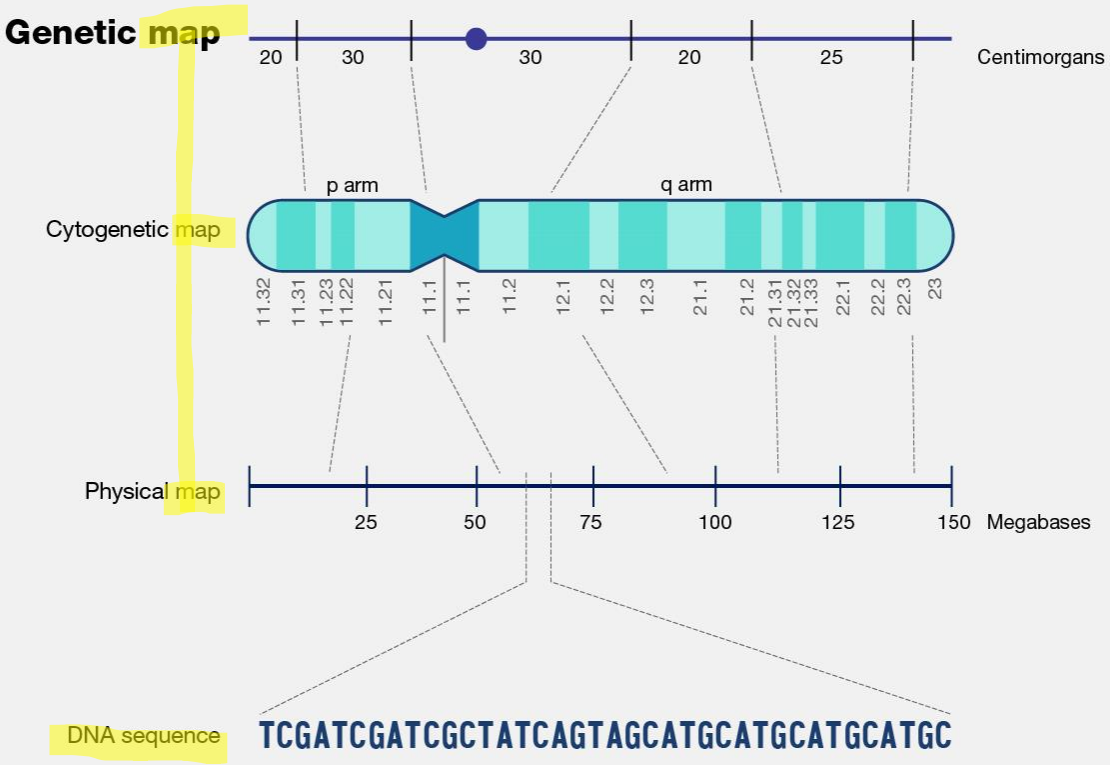
CENTER OF
PLANT SCIENCES
Sant'Anna
Scuola Universitaria Superiore Pisa

Advanced Genomics

The dawn of sequencing: Sanger and the HGP

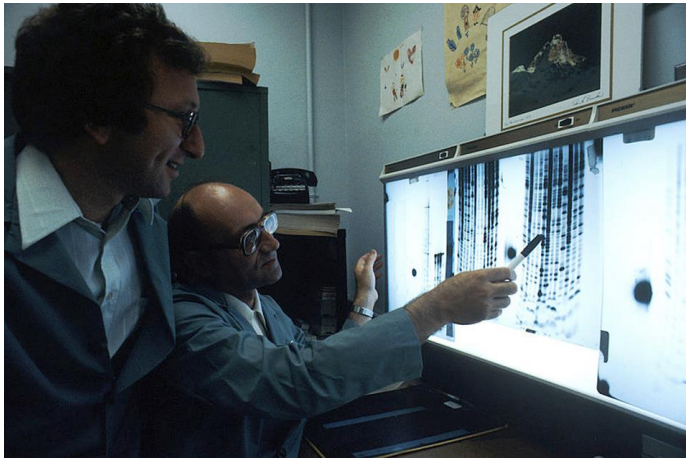


Two pillars of genome reconstruction:

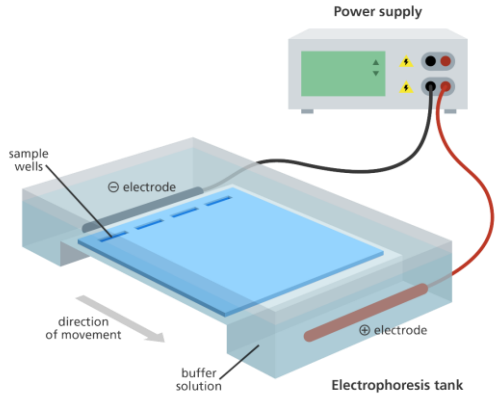
1. Map
2. Sequence

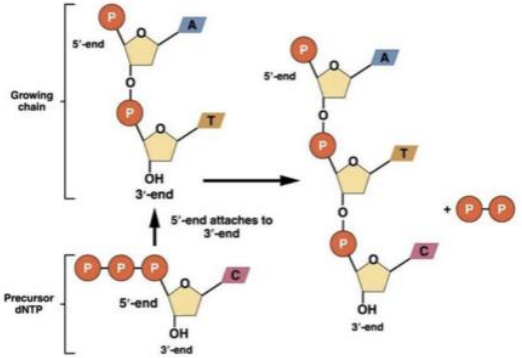
DNA sequencing

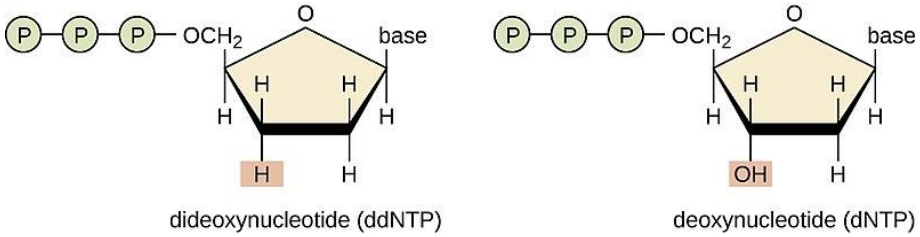
- First version in 1968
- A combination of:
  - Electrophoresis
  - *in vitro* DNA replication
  - random incorporation of chain-terminating dideoxynucleotides
- The leading technology for 40 years
- Still used today



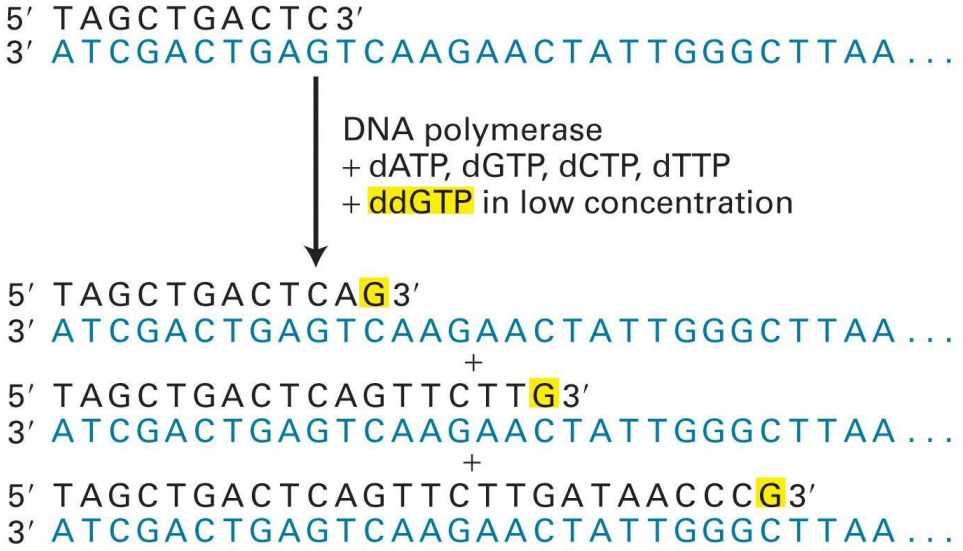
Sir Frederik Sanger (1918-2013)
TWO Nobels in chemistry (1958, 1980)







- Chain termination method: the incorporation of ddNTPs will stop the reaction
- The reaction is run in parallel for different ddNTPs



The reaction is stopped at all possible sites; products are separated by lenght using electrophoresis and the template sequence is reconstructed

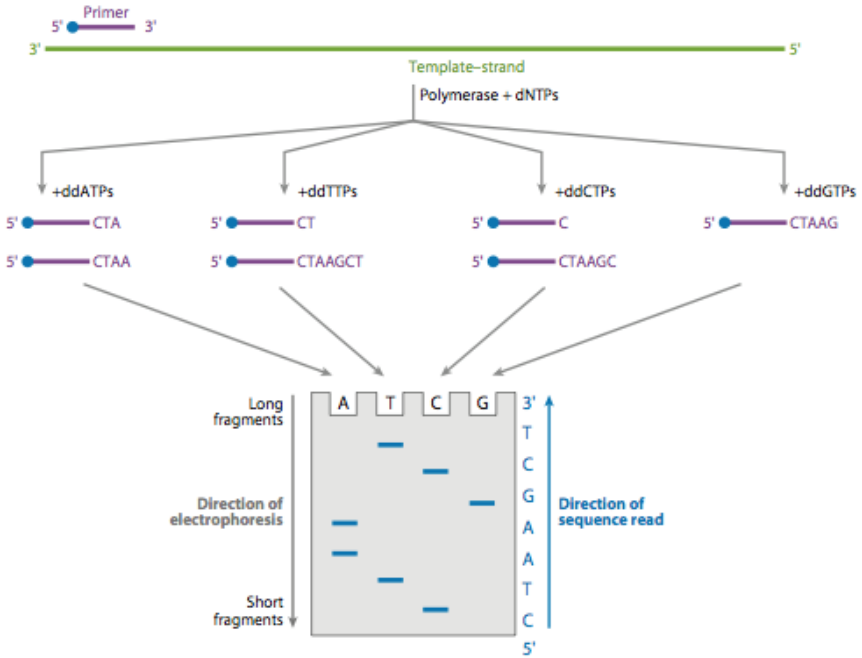
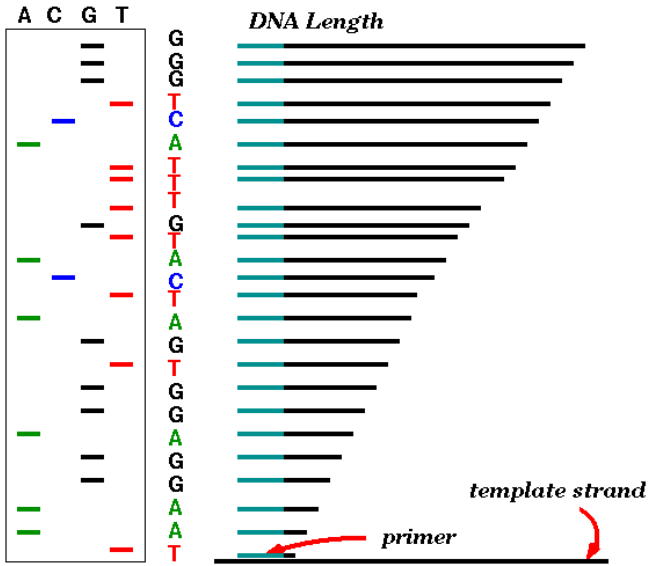
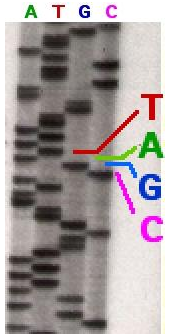
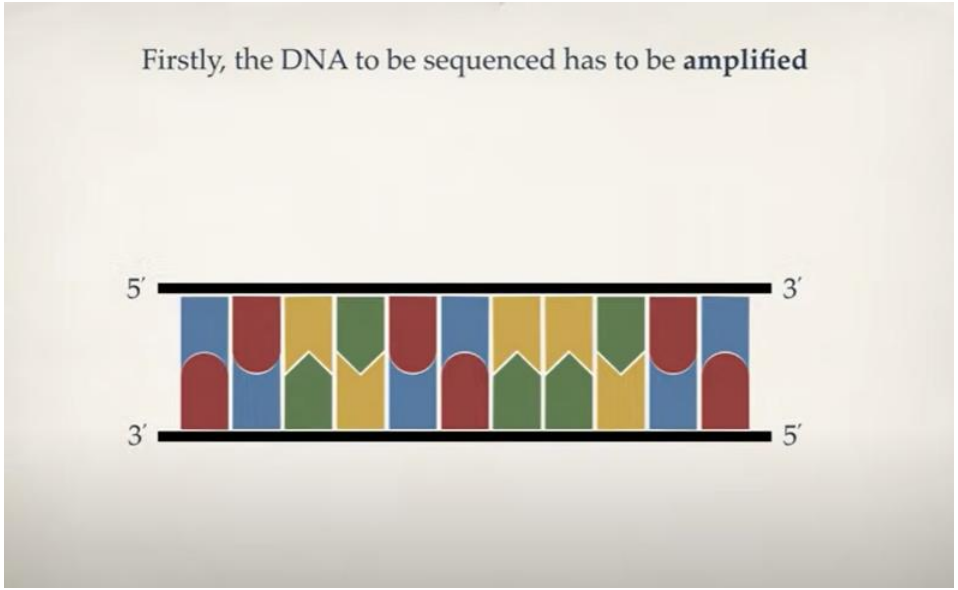


Figure 1

Sanger sequencing.







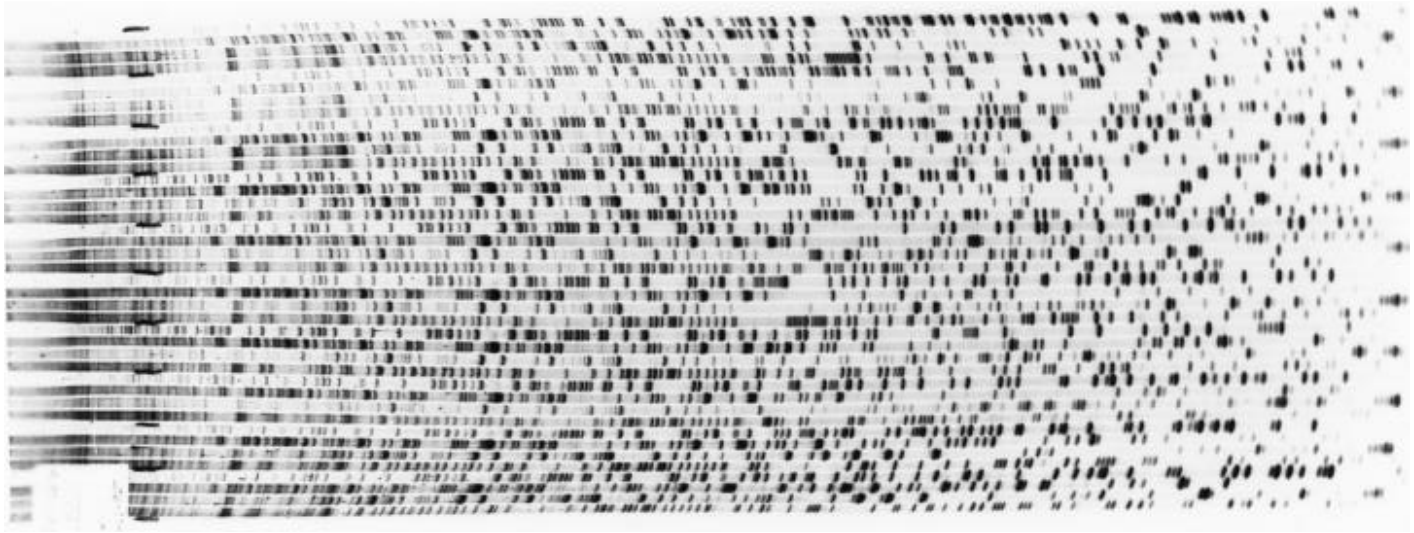
https://www.youtube.com/watch?v=FvHRio1yyhQ&ab_channel=My2Sense

Pros of the Sanger method

- Simple chemistry, cheap and easy to implement
- Very high accuracy
- Sequence lenght of 500+ bp

Cons of the Sanger method

- Low throughput



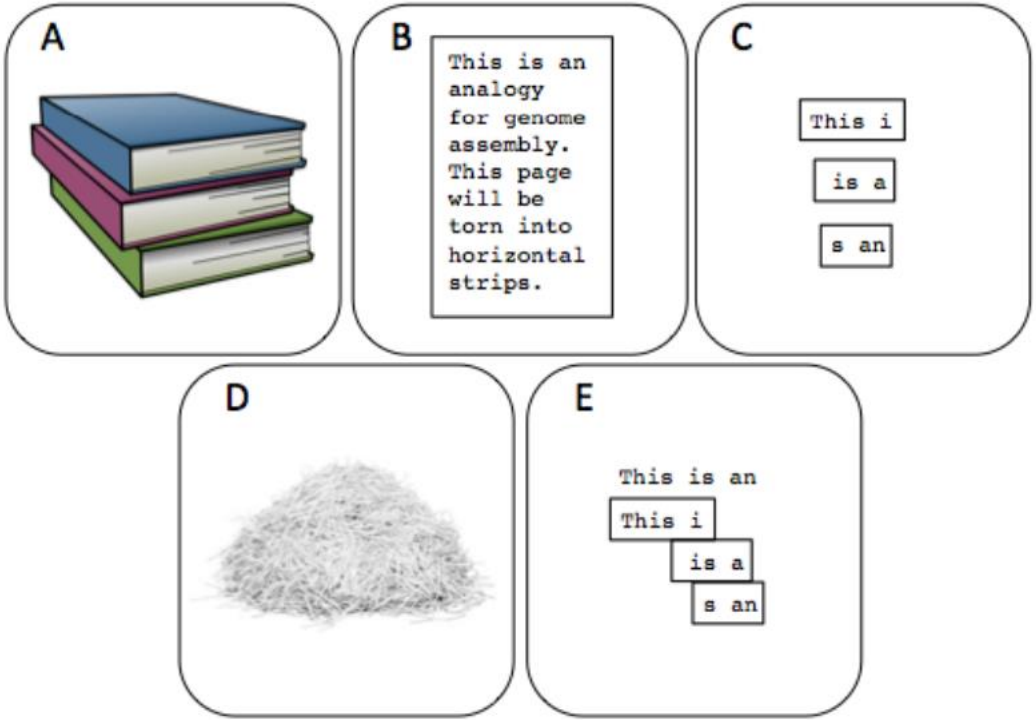
Acknowledging complexity

We saw that most genomes are BIG; very long molecules (chromosomes) which contain the nucleotide sequence we want to be able to read

- Genomics is limited by the length of the DNA sequence that we can read at once. The limit comes to the capacity to sustain DNA polymerization in vitro (similarly to what happens with PCR)
- Only recently we managed to read continuous sequences in the range of Mbs (we will see about that when talking about third gen sequencing)

As we stand, all genomic approaches are based to some degree of fragmentation of genomes bigger than a few Mbs

To reconstruct a genome from small fragments is like shredding a book and trying to reconstruct its original content



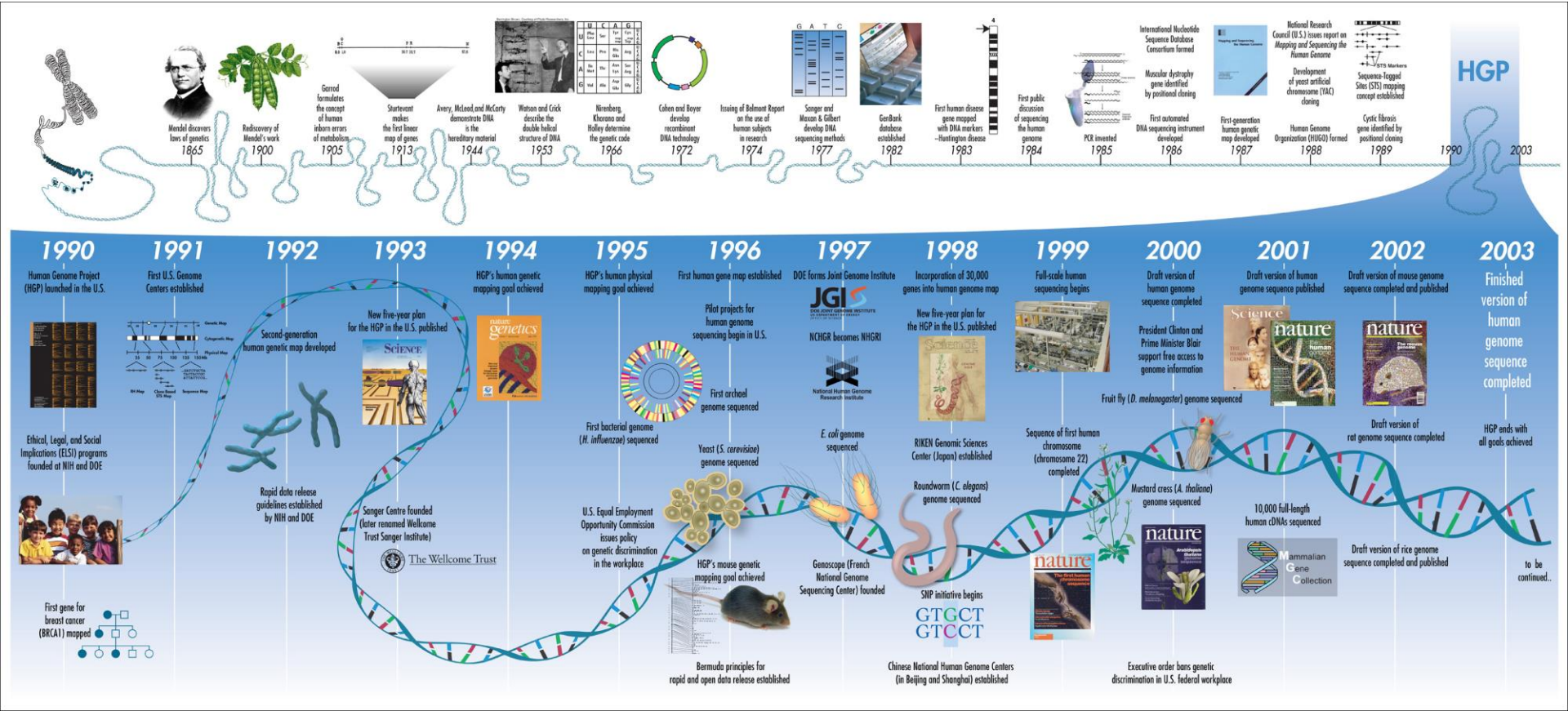
Molecular markers are like key moments in the story; if you are reading The Shining, you know that Jack Torrance turns crazy *after* the winter starts at the Overlook Hotel

Still, this is a momentous challenge that requires reducing the complexity of the task. Working chapter by chapter, perhaps

The dawn of genomics: the human genome project (HGP)

- Genomics is a legacy of the HGP
- Aim: to map the complete set of nucleotides contained in a human haploid reference genome
- Started in 1990 and completed in 2003
- $3 billion project founded by the US Department of Energy and the National Institutes of Health (NIH) to the International Human Genome Sequencing Consortium (UK, USA, FRA, DEU, JAP)
- Still the largest collaborative effort in biological sciences
- Gapless assembly only published in 2022!

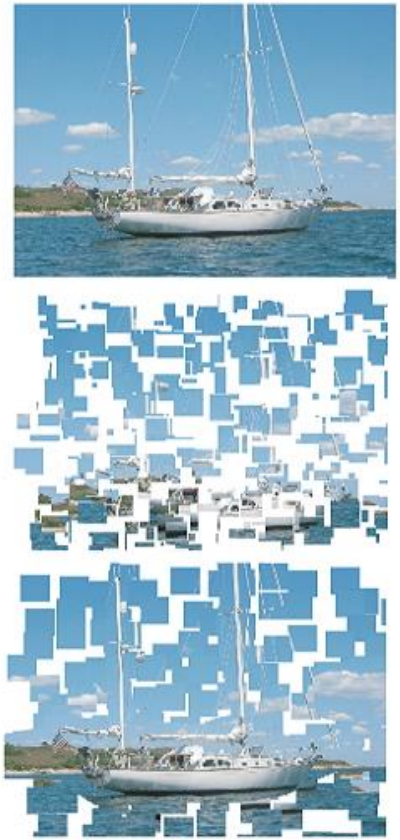
A concerted, coordinated effort



Not only human genome sequence

- Several other species sequenced as complement to humans
- Strong impulse to the developement of bioinformatics
- Push for optimization in automation and sequencing related to DNA sequencing
- Establishment of the Bermuda principle (1996):
  - Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours)
  - Immediate publication of finished annotated sequences
  - Aim to make the entire sequence freely available in the public domain for both research and development in order to maximise benefits to society

How to get to a genome?



1

Sampling and DNA/RNA extraction

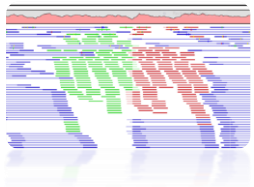
2

Molecule disruption

3

sequencing

4

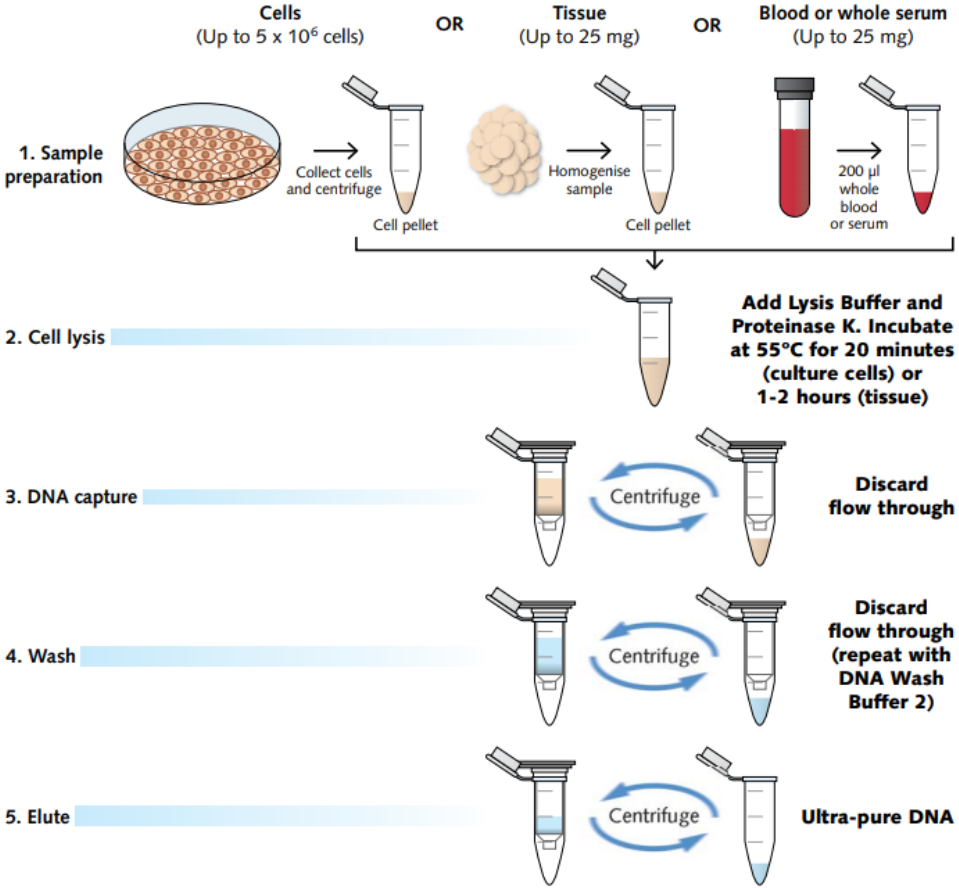


analysis

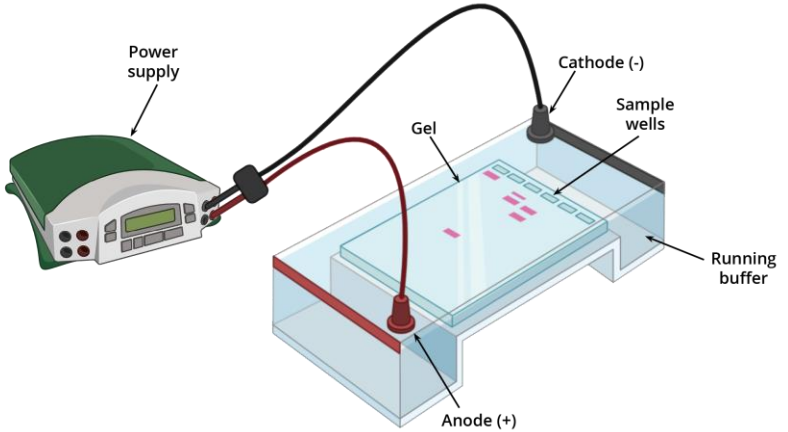
DNA extraction

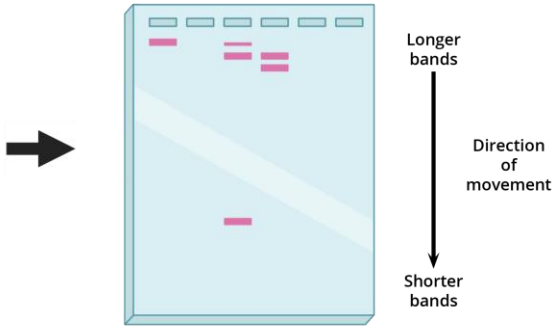
- Steps vary and depend on extraction protocol and target organism
- First step is to break down tissues
- Procedures may be performed in liquid nitrogen to prevent DNAse activities and to make tissues brittle

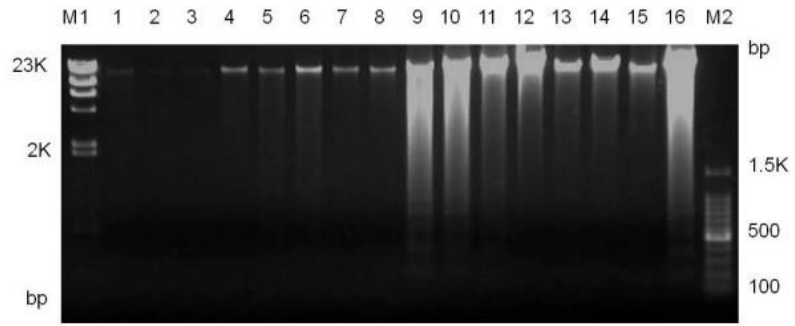

DNA
invitrogen
by Thermo Fisher Scientific

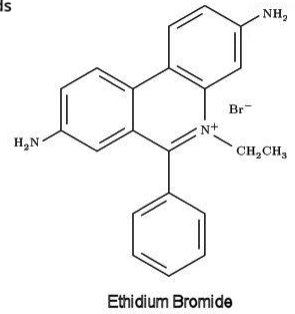

Cells
(Up to 5 x 10⁶ cells)
OR
Tissue
(Up to 25 mg)
OR
Blood or whole serum
(Up to 25 mg)
1. Sample preparation
Collect cells and centrifuge
Cell pellet
Homogenise sample
Cell pellet
200 µl whole blood or serum
2. Cell lysis
Add Lysis Buffer and Proteinase K. Incubate at 55°C for 20 minutes (culture cells) or 1-2 hours (tissue)
3. DNA capture
Centrifuge
Discard flow through
4. Wash
Centrifuge
Discard flow through (repeat with DNA Wash Buffer 2)
5. Elute
Centrifuge
Ultra-pure DNA

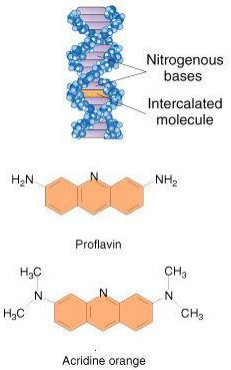
- DNA quality and quantity can be checked with gel electrophoresis or spectrophotometry

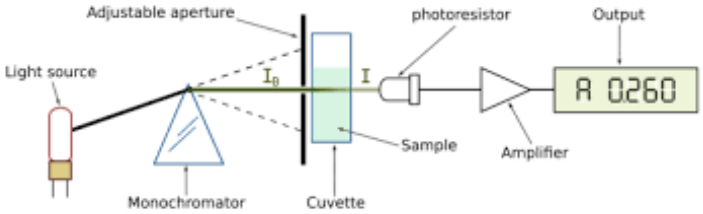


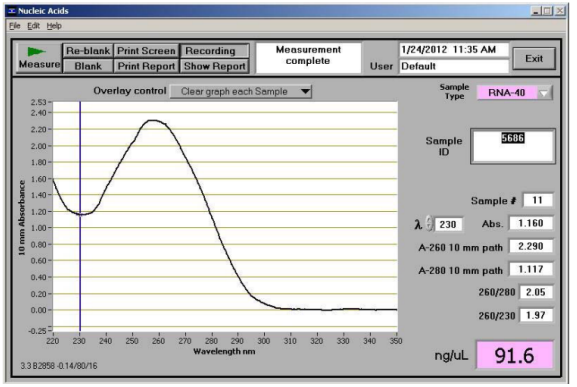


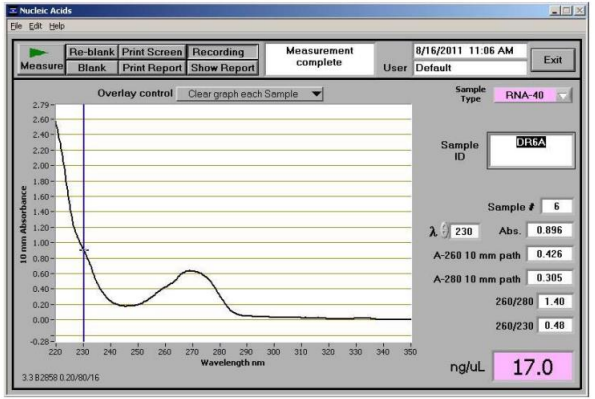






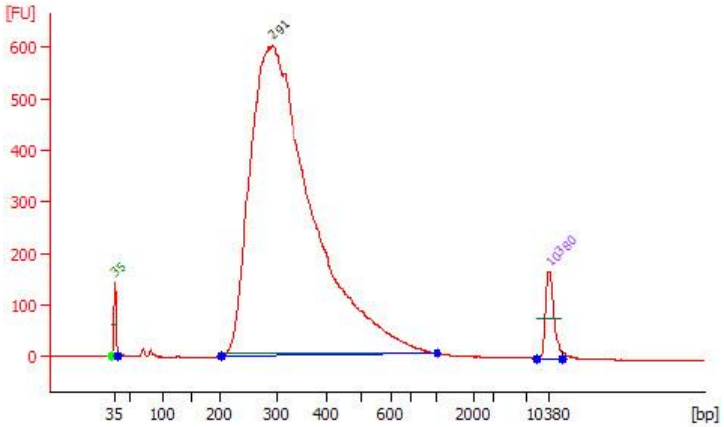


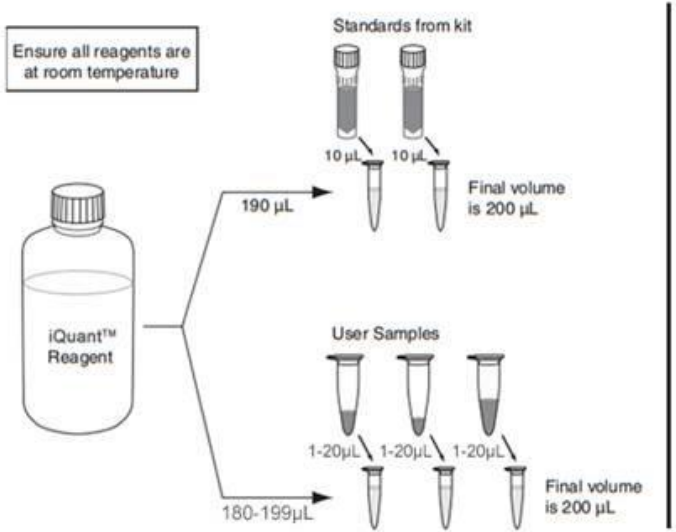




- Nucleic acids absorb UV light at 260 nm due to the aromatic base within their structure.
- Proteins and phenolic compounds have a strong absorbance near 280 nm
- Many organic compounds have strong absorbances at around 225 nm. In addition to phenol, TRIzol, and chaotropic salts, the peptide bonds in proteins absorb light between 200 and 230 nm.

- More precise and reliable marker-based methods: bioanalyzer and qbit





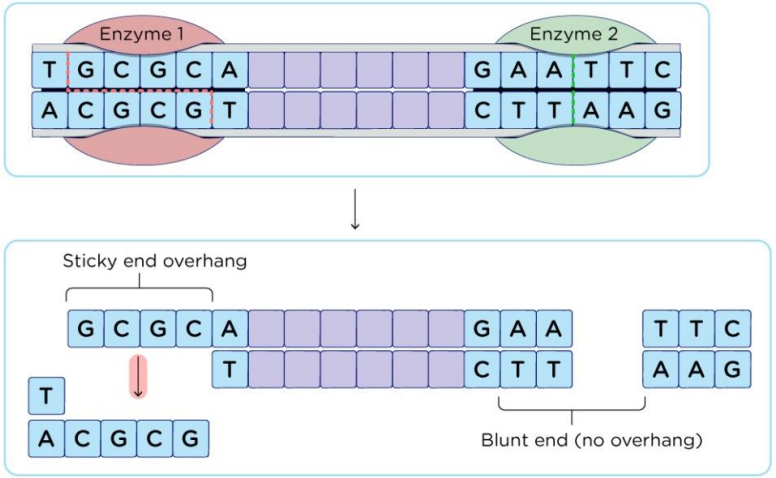
Vortex all assay tubes for 2-3 seconds

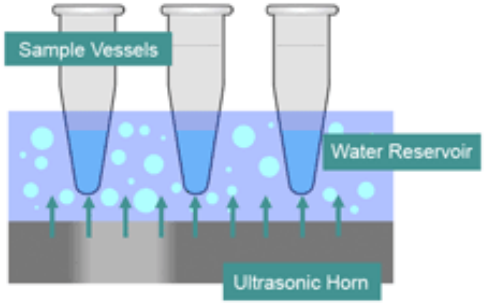
Incubate at room temperature for 2 minutes

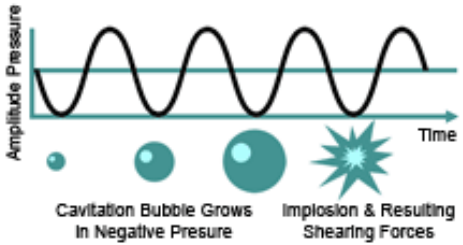
Read tubes in Qubit® Fluorometer

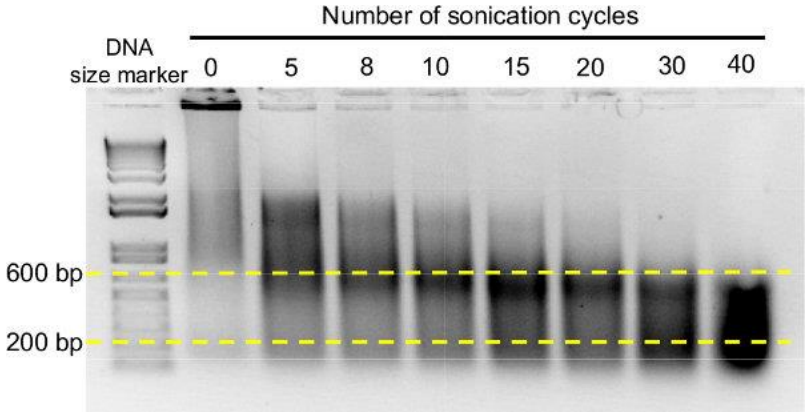
DNA fragmentation

- Need to “simplify” input DNA
- Optional
- Either by sonication or by restriction enzymes

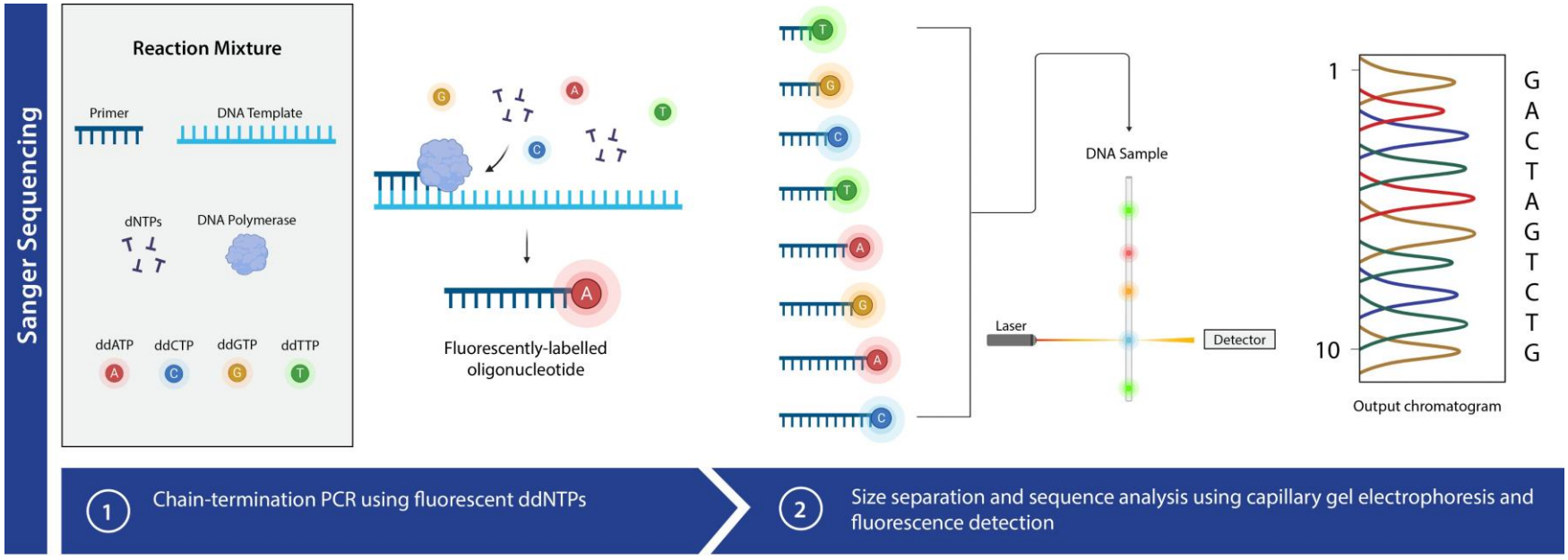


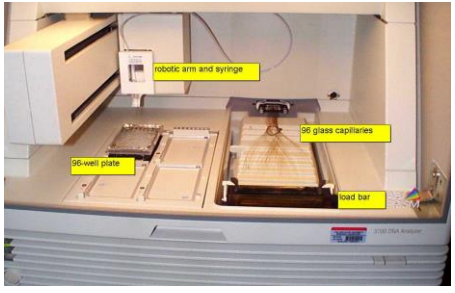




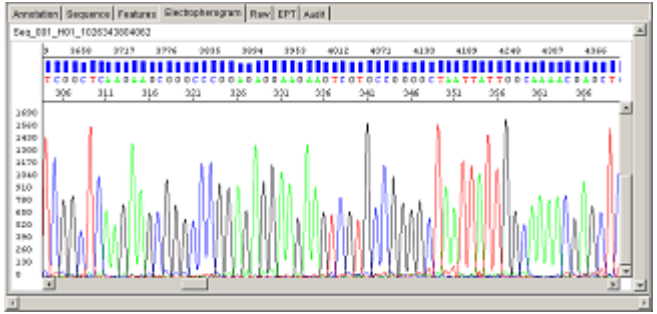


Sequencing

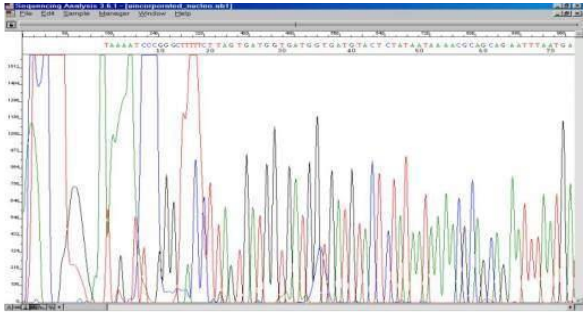




- Most modern automated sequencers used very thin capillary tubes
- Up to 96 capillaries running at the same time



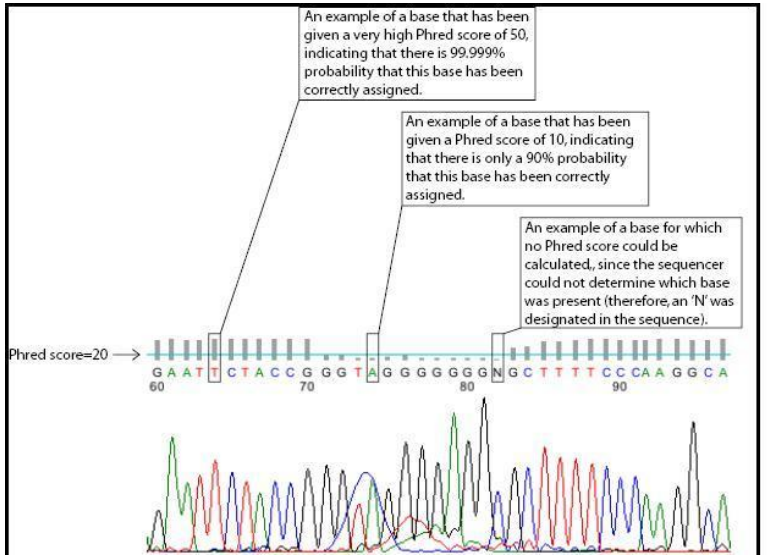
Ideal chromatogram



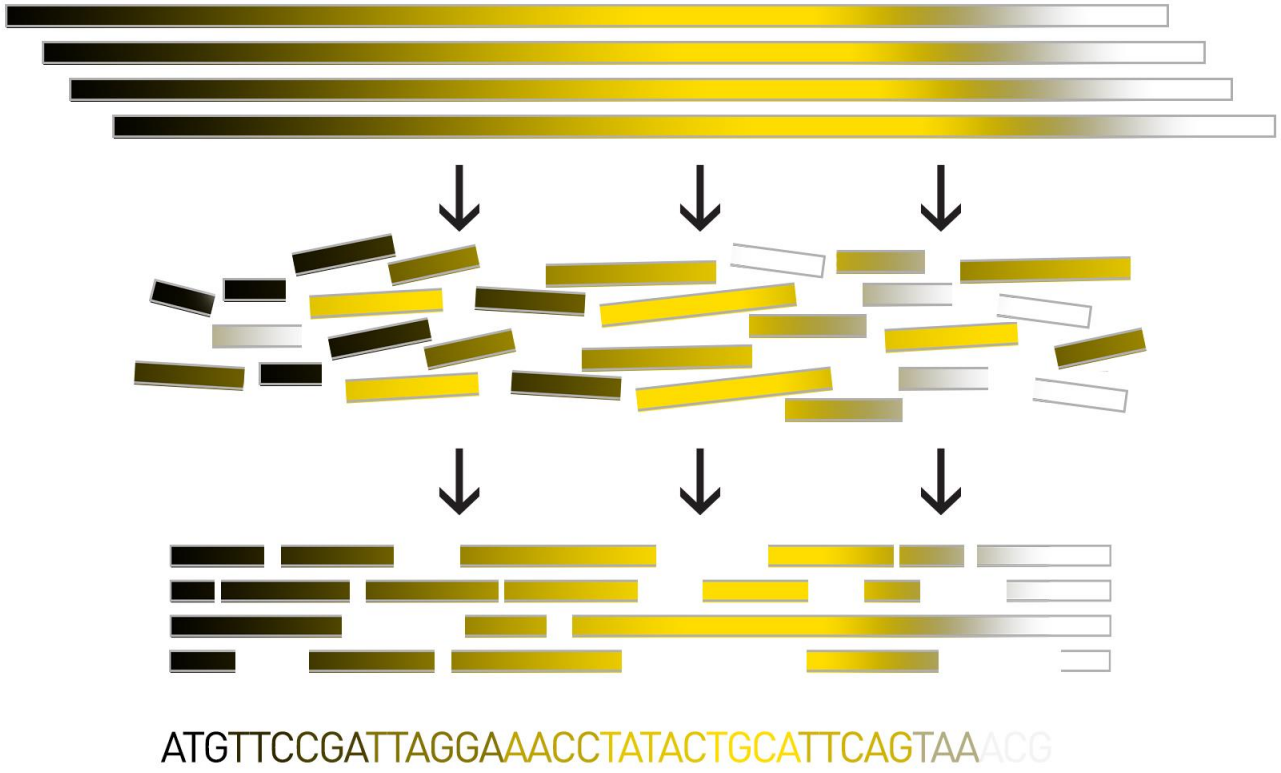
Real chromatogram

- Need for automated detection methods, filtering, correction for compression, and calling of correct nucleotides
- PHRED program developed to calculate accurate quality scores logarithmically linked to error probabilities

$$Q = -10 \log_{10} P$$



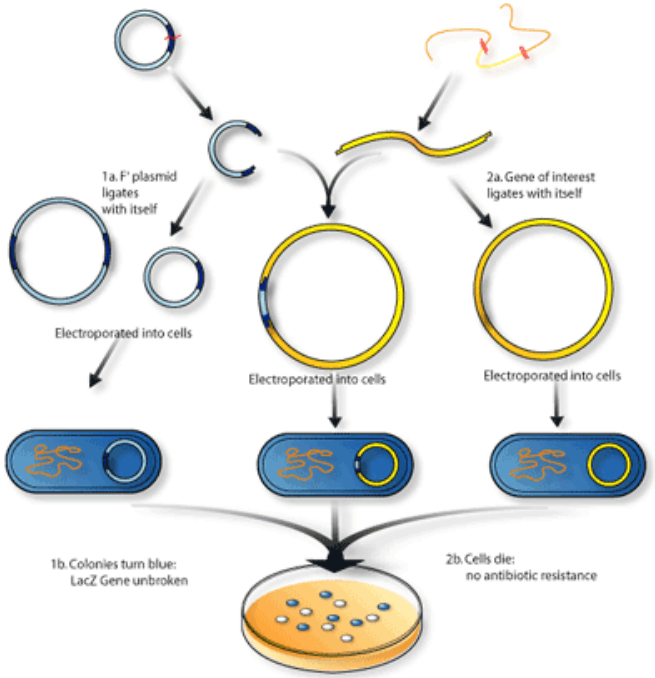
Analyze reads, assemble genomes



The strategy to tackle complexity in the HGP

Hierarchical whole genome sequencing based on artificial chromosomes

- The HGP developed two host systems: *Saccaromyces cerevisiae* and *Escherichia coli*
- The idea is to use their DNA to host longer fragments of human DNA, and sequence them as a individual entity (chapter by chapter, if you want to follow up with the book analogy)
  - Yeast artificial chromosomes (YAKs) can contain more than 1Mb of sequence and be introduced and propagated in yeast
  - Bacterial artificial chromosomes (BACs) can contain shorter fragments (a few hundred Kbs) in *E. coli*, with a fast and easy propagation



Hierarchical shotgun sequencing

Genomic DNA

BAC library

Organized mapped large clone contigs

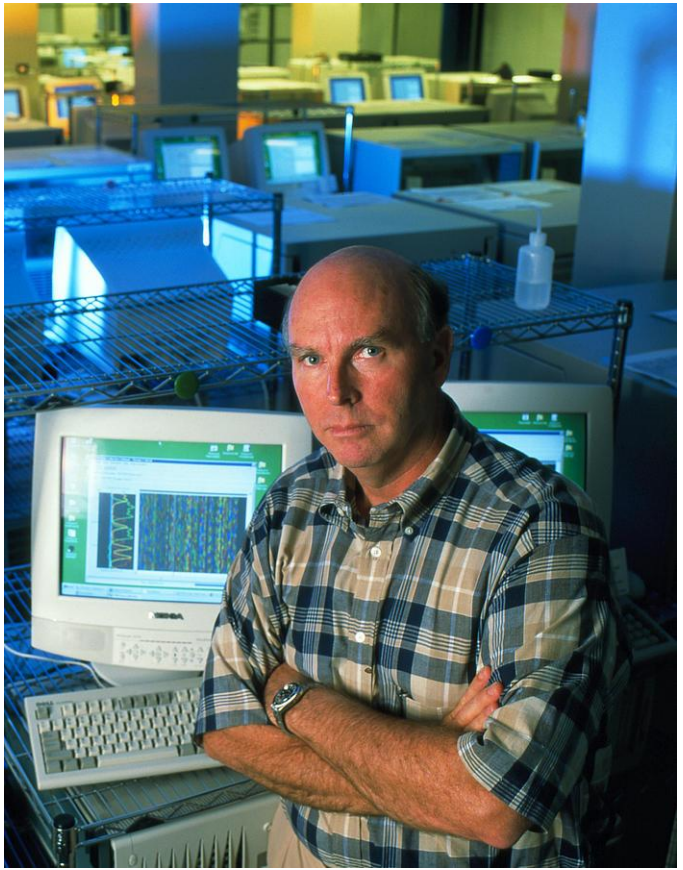
BAC to be sequenced

Shotgun clones

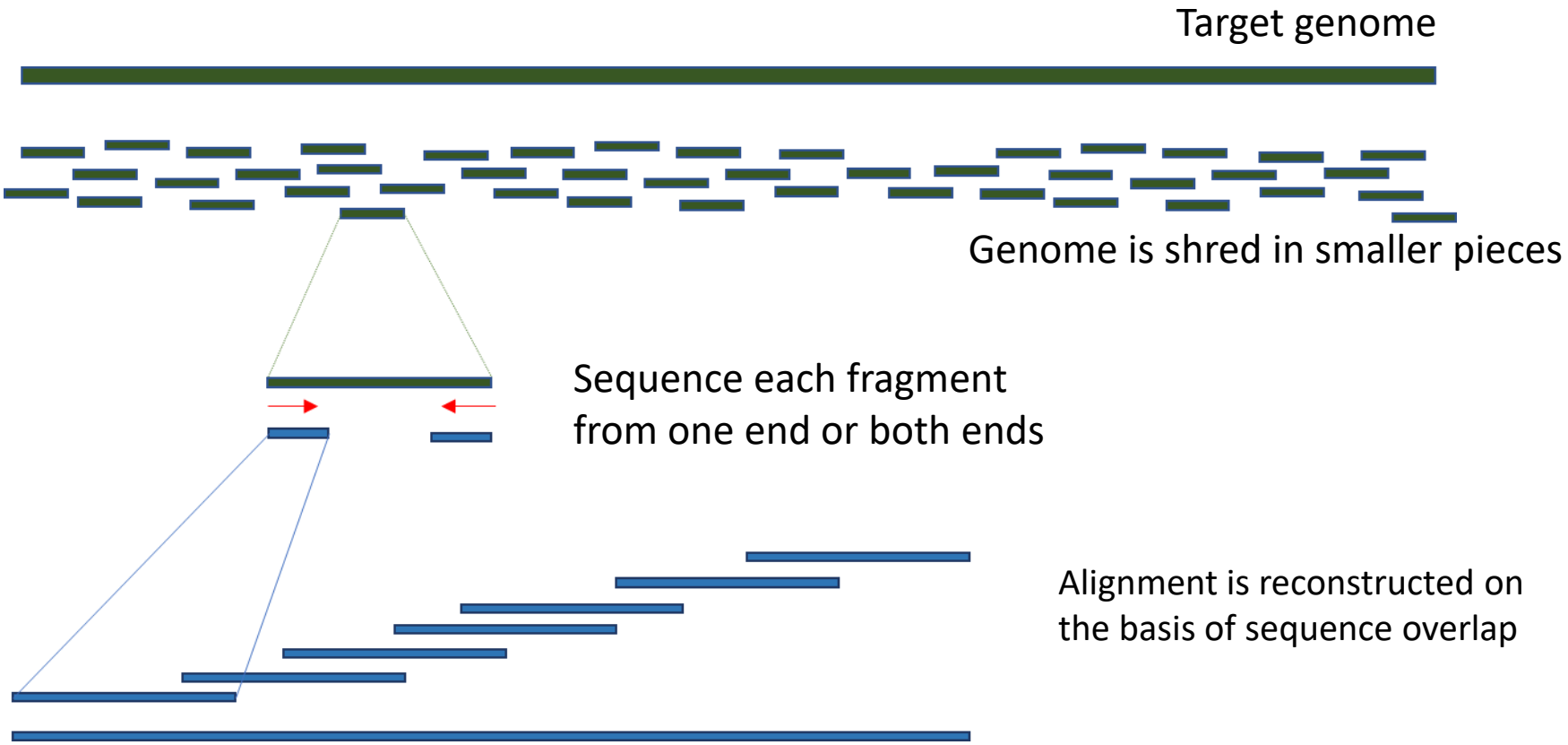
Shotgun sequence ...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG..

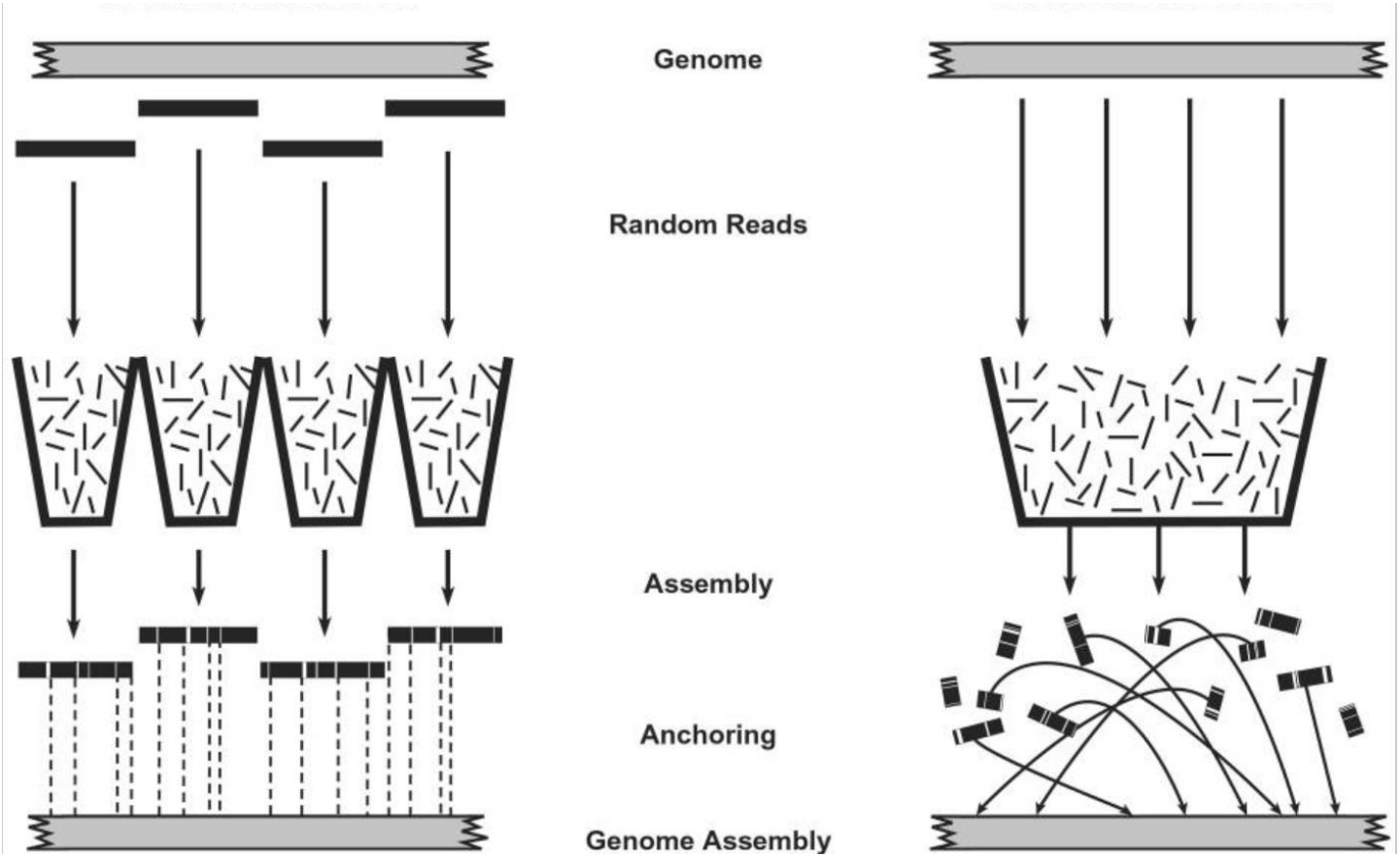
Assembly ...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG..

- The race was on with Celera Genomics (Craig Venter), which used shotgun sequencing, a cruder and faster method based on sequencing fragments of the whole genome (Whole Genome Shotgun, WGS) and assemble them with supercomputers
- At the time of the HGP, map construction and library production accounted for a minor fraction (<10%) of the total sequencing costs
- Celera started in 1998 and spent only $300 M, but was facoured by the open access policy of the HGP

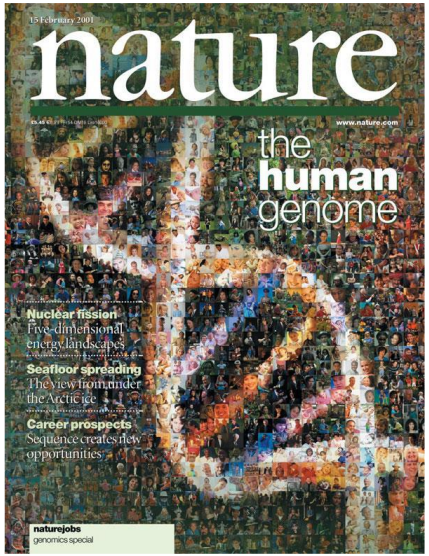


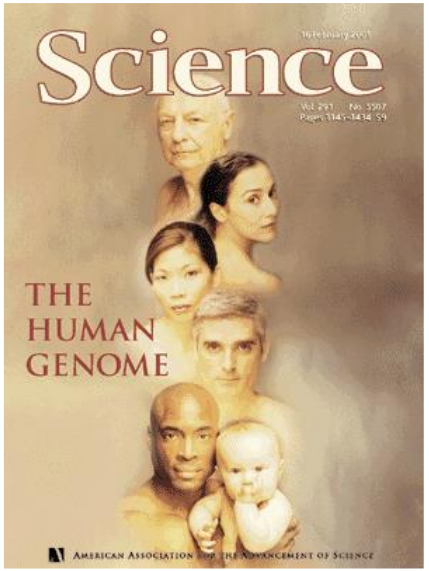
The shotgun sequencing strategy




HGP
CELERA
Genome
Random Reads
Assembly
Anchoring
Genome Assembly

Eventually, first genome drafts were published in 2001 on Nature (HGP) and Science (Celera)



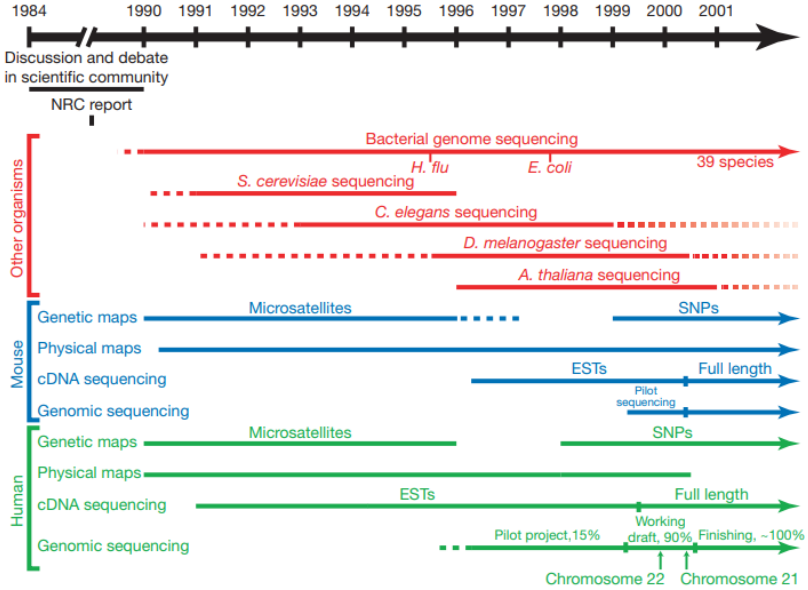


Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

- A paper of 62 pages!
- The estimated number of genes is about 30,000, one-fourth as great as previously thought. Less than that for the plant *Arabidopsis thaliana*.
- The average gene consists of about 40,000 bases, but gene sizes vary greatly
- The DNA sequence in any two individuals is 99.9 percent identical.
- Less than 2 percent of the genome codes for proteins. The human genome has a much greater portion (50 percent) of repeat sequences than the mustard weed (11 percent), the nematode worm (7 percent), and the fruit fly (3 percent).
- The average human gene produces three different proteins.
- Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA in between



First Generation

Sanger Sequencing
Maxam and Gilbert
Sanger Chain-termination

- Infer nucleotide identity using dNTPs then visualize with electrophoresis
- 500-1000 bp fragments

Second Generation
(Next Generation Sequencing)

454, Solexa,
Ion Torrent
Illumina

- High throughput from the parallelization of sequencing reactions
- ~50-500 bp fragments

Third Generation

PacBio
Oxford Nanopore

- Sequence native DNA in real time with single-molecule resolution
- Tens of kb fragments, on average

Short-read sequencing

Long-read sequencing