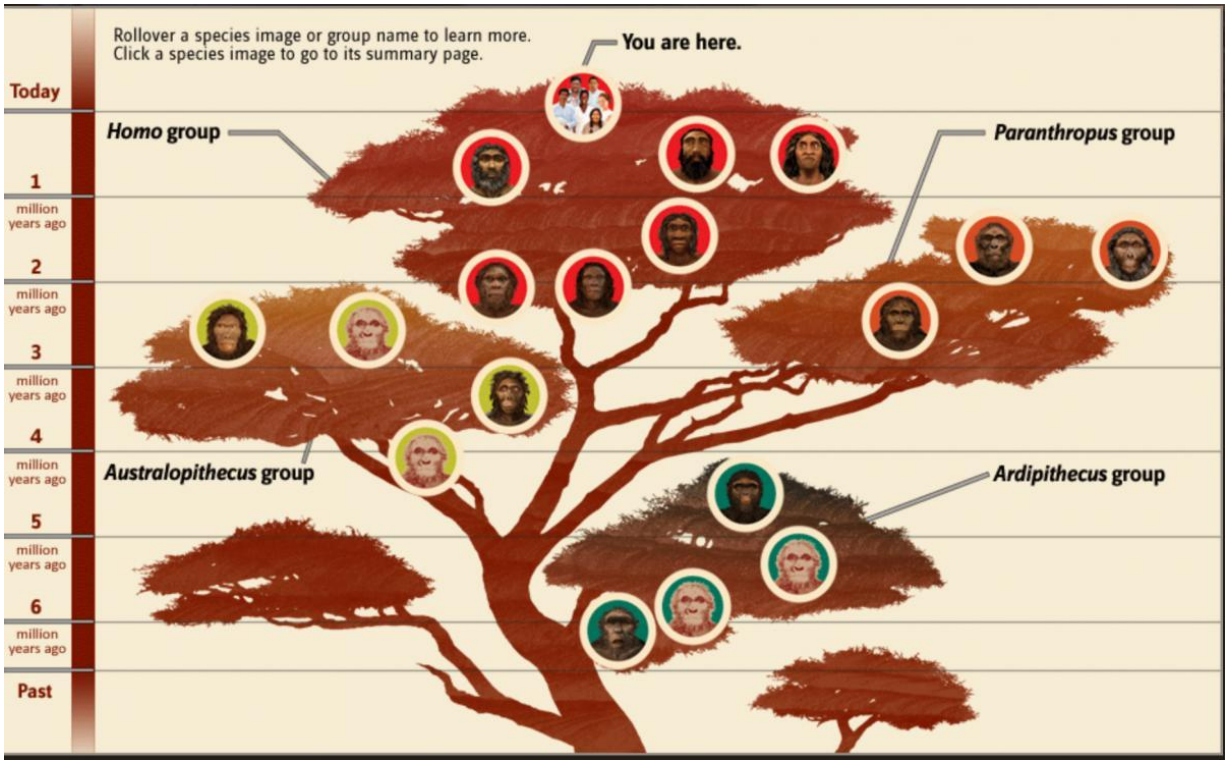
Genome sequencing and assembly

Research findings

- http://www.thehindu.com/news/international/neanderthals-humans-interbred-a-lakh-years-ago-study/article8252757.ece

- http://humanorigins.si.edu/evidence/human-family-tree



Recap

The story of four-year-old Georgia Walburn-Green and her family.

http://www.itv.com/news/2016-01-12/georgias-story-gives-hope-to-many-families-after-dna-sequencing-breakthrough/

The story of four-year-old Georgia Walburn-Green and her family.

Born seemingly healthy, her parents Amanda and Matt told ITV News they only had 20 minutes of “worry-free cuddling” with their brand new baby before doctors drew the curtains around the bed to tell them something was not quite right - they could see that Georgia’s head was bigger than normal.

News on ITV on 12-jan-2016

Since then, Georgia has had a range of problems with her eyes, kidneys and brain.

Her physical and mental development has been delayed. Despite years of tests, though, doctors couldn't pinpoint what was wrong with this otherwise lively, smiling girl.

That changed a few weeks ago, when Amanda and Matt got a phone call from scientists at Great Ormond Street Hospital.

Georgia's condition, they were told, was due to a mutation in a gene called Kdn5b.

Fragment assembly of DNA

- Biological background
- Models
- Algorithms
- Heuristics

Fragment assembly of DNA

- Fred Sanger developed the first technique for sequencing DNA.
- DNA is replicated in the presence of chemically altered versions of the A, C, G, and T bases.
- These bases stop the replication process when they are incorporated into the growing strand of DNA, resulting in varying lengths of short DNA.
- These short DNA strands are ordered by size, and by reading the end letters from the shortest to the longest piece, the whole sequence of the original DNA is revealed.

Fragment assembly of DNA

- Automated method of DNA sequencing, built upon the chemistry of PCR and the sequencing process developed by Frederick Sanger in 1977.

- https://scholar.google.co.in/scholar?q=Sanger+method+of+DNA+sequencing&btnG=&hl=en&as_sdt=0%2C5
- http://smcg.ccg.unam.mx/enp-unam/03-EstructuraDelGenoma/animaciones/secuencia.swf
- https://www.dnalc.org/view/15479-Sanger-method-of-DNA-sequencing-3D-animation-with-narration.html
- http://www.yourgenome.org/video/sanger-dna-sequencing
- https://www.youtube.com/watch?v=KTstRrDTmWI

Limitations to sequencing

- You must have a primer of known sequence to initiate PCR

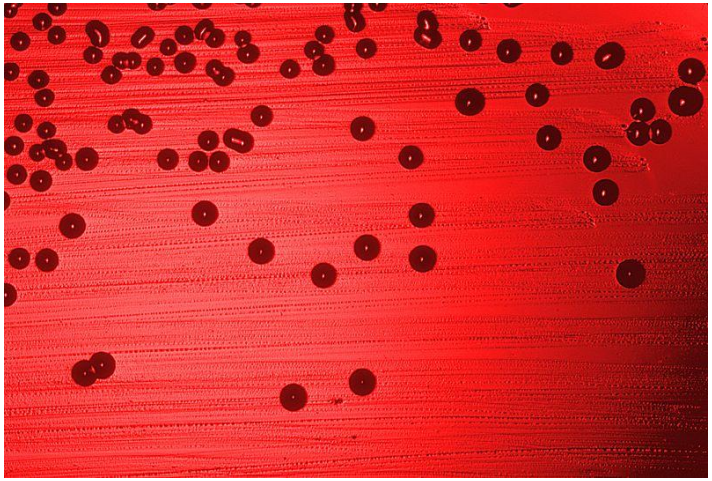
- Only about 1000nts can be sequenced in a single reaction

Limitations to sequencing

- Human Genome sequencing ?

Whole genome sequencing

- ***Haemophilus influenzae (1995)***
  - commensal bacterium which resides in the human respiratory tract was the first organism to have its entire genome sequenced



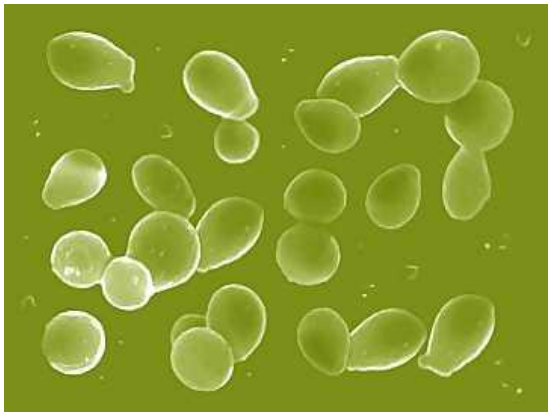
Haemophilus influenzae

http://science.sciencemag.org/content/269/5223/496

1,830,140 base pairs of DNA

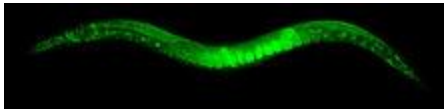
Whole genome sequencing

- **yeast *Saccharomyces cerevisiae* (1996)**
  - The first eukaryotic genome was sequenced.
- Has a genome of only around 12 million nucleotide pairs
- https://dx.doi.org/10.1126%2Fscience.274.5287.546



Whole genome sequencing

- The worm ***Caenorhabditis elegans*** ***(1998)***
  - The first animal to have its whole genome sequenced.
- Has a genome of only around 100,258,171 nucleotide pairs.



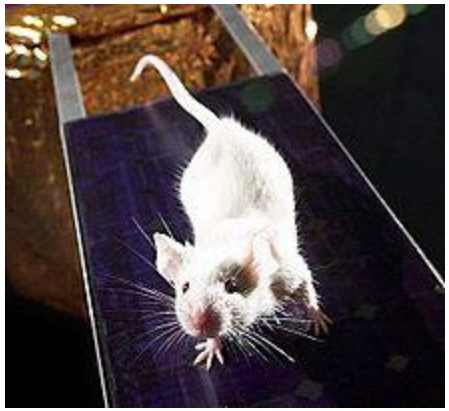
- https://dx.doi.org/10.1126%2Fscience.282.5396.2012

Whole genome sequencing

- In 1999, the entire DNA sequence of human chromosome 22.
- By the year 2000, the second animal and second invertebrate (yet first insect) genome was sequenced - that of the fruit fly *Drosophila melanogaster*
- The first plant genome - that of the model organism *Arabidopsis thaliana* - was also fully sequenced by 2000

Whole genome sequencing

- The genome of the laboratory mouse *Mus musculus* was completed in 2002



Whole genome sequencing

In 2004, the Human Genome Project published the human genome.

articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

**A list of authors and their affiliations appears in the* Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

Whole Genome Sequencing

- Human Genome sequencing ?

Fragment assembly of DNA

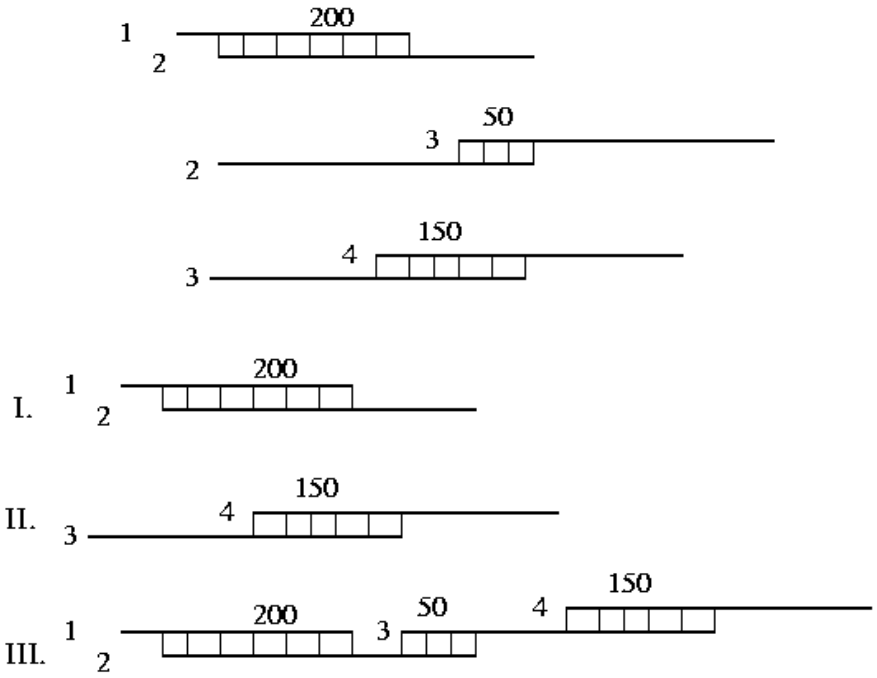
A typical approach to sequencing long DNA molecules is to sample and then sequence fragments from them.

Fragment assembly of DNA

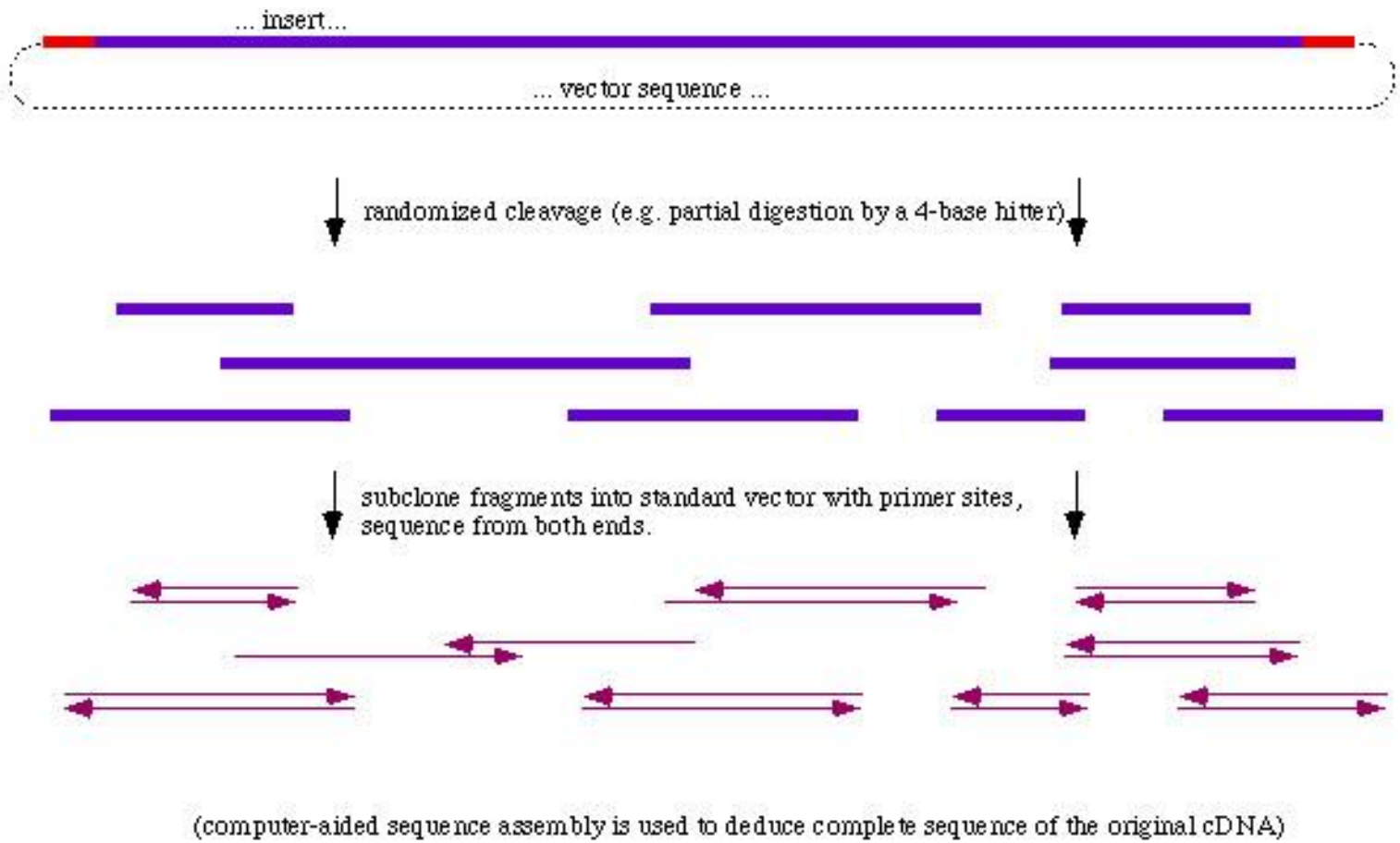
- In the 1980s, two key developments allowed researchers to believe that sequencing the entire genome could be possible.
- The first was a technique called polymerase chain reaction (PCR) that enabled many copies of DNA sequence to be quickly and accurately produced.
- The second, an automated method of DNA sequencing, built upon the chemistry of PCR and the sequencing process developed by Frederick Sanger in 1977.

Greedy

- Build a rough map of fragment overlaps
- Pick the largest scoring overlap
- Merge the two fragments
- Repeat until no more merges can be done



Shotgun Sequencing



The Ideal Case

Find maximal overlaps between fragments: