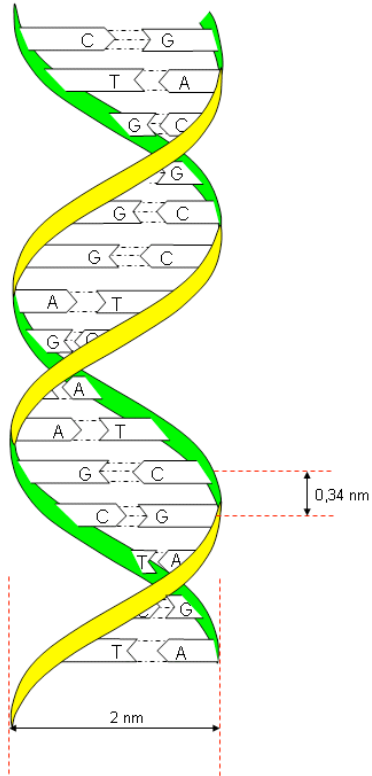


## 2. Genes and proteins

- The sequence as a model of DNA
- Genes: from Mendel to molecular biology
- The genetic code
- A translation algorithm
- Implementing the genetic code
- Algorithms + data structures = programs
- The algorithm design trade-off
- **DNA sequencing**
- Whole genome sequencing
- How to find genes?

# DNA sequencing

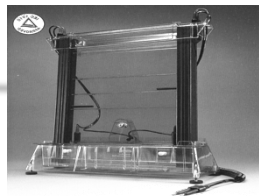
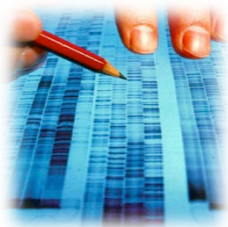
# DNA sequencing



...  
GATCACCTCACTACGGGTCAGGGGAA  
GGAAAGGGGAAGTGAAGATTTGCAG  
TGTGAGAAGCAGTCCCAGGAGTTAGA  
AGTAGTGGCTCCATGACTCACAAATT  
AACTTCCCTTTTCAGGCAGGGCTTCTT  
ATTTTCCTTAGCATCCCTGTCTTGAT  
CCCAGCCTGCTCAGACCCCTGCCTCT  
CACTGCAAGATGTGCTTGAGTATGAG  
AGTCAGGAATGTTACTTCTCAGAGGC  
GCCAAATGGCAGTTGTCACAGGGTCA  
TCATAGAGGGTATATGTTTACTGCAC  
TGGGCTCTGAGGCTTGCTTGTGAAGA  
AACAGAAGCTAAGGGATCCAGGGAGT  
CCCAACTTAGAGAGTCCCACAGGCC  
ACACTCTGGTTCTGTTGGCAGGAAAA  
TTTGGCTGAATTGGGGCAGGAAGTTG  
TGTAACAAAACGATTACATCCATTTT  
TGCAAGGCAAGAGTGAGCTATTACACC  
TCCATGTTGGTGATATTTTTTGCCAT  
ATAAGCAGCTAATTCCTTTTCAGTAAT  
TCTACTCTAAACTAGTCTTAATGTGA  
CTTCTATATAAATTCTGAACTGAATA  
ATTTTGGGAACGTTGTAAAAA...

# Sequencing is a so-called “exponential technology”

- First sequences obtained in the early 70's
- Next Generation Sequencers (NGS) around 2008



1990:  $10^3$  bases/day

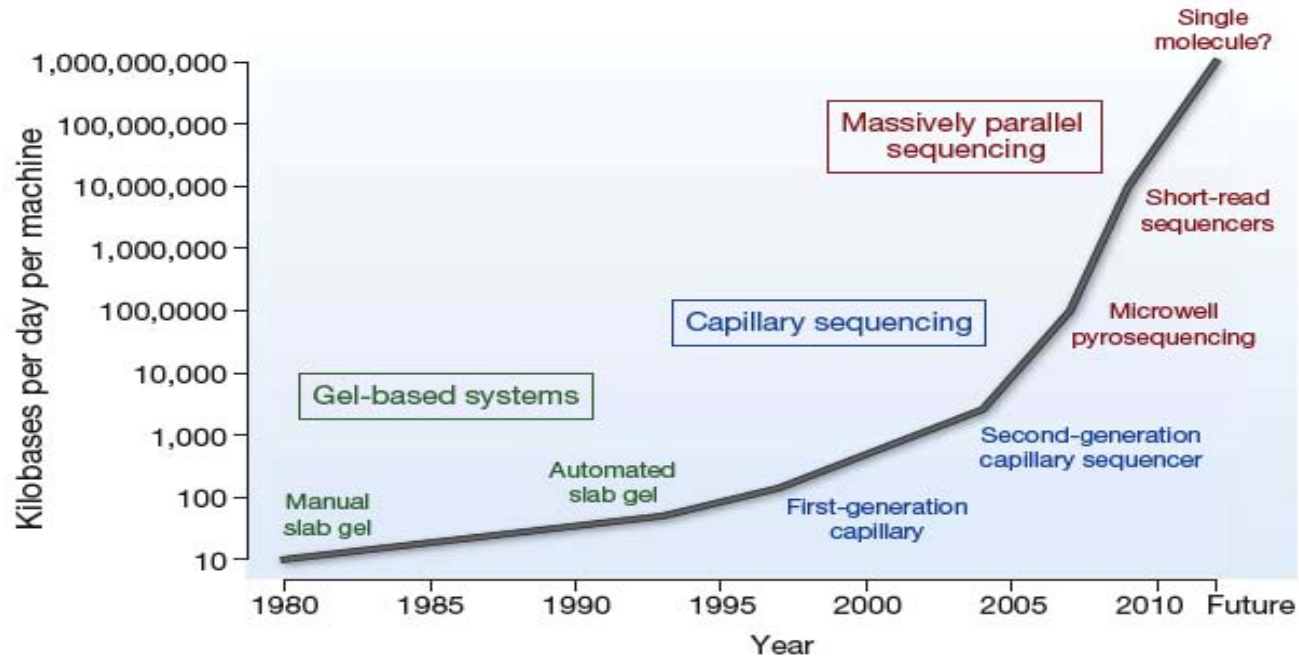


2000:  $10^6$  bases/day



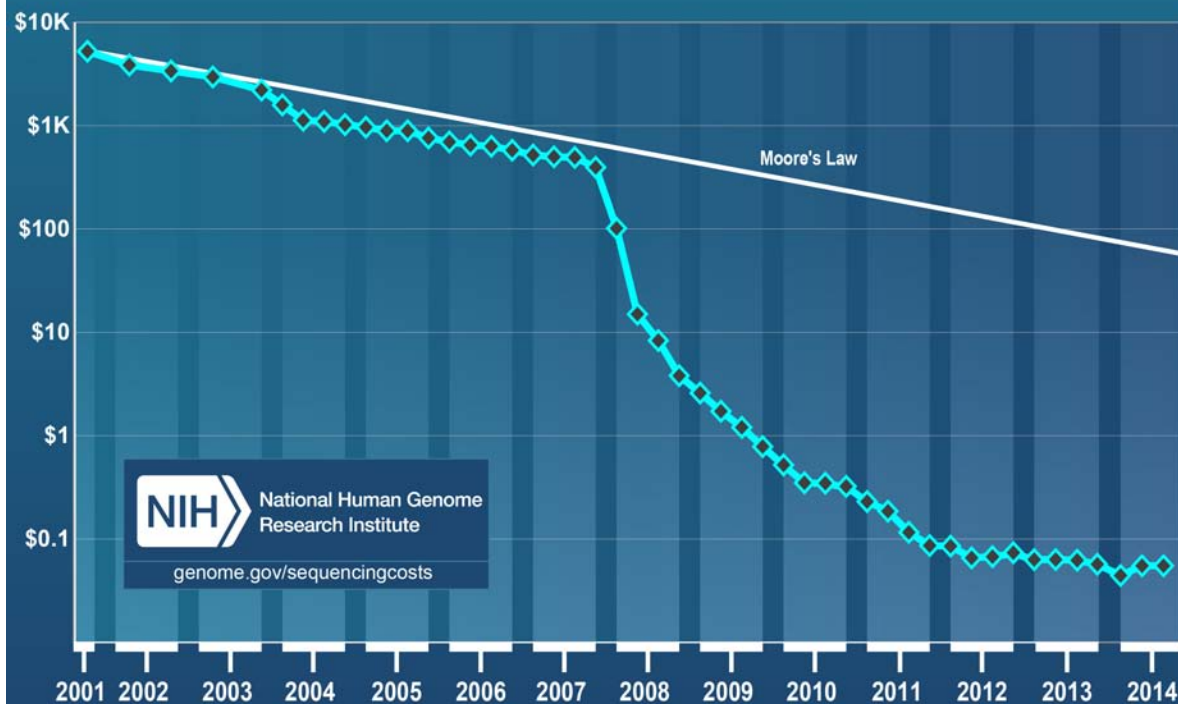
2010:  $10^9$  bases/day





“The cancer genome”, Michael R. Stratton, Peter J. Campbell, P. Andrew Futreal, *Nature*, Vol. 458, Avril 2009

## Cost per Raw Megabase of DNA Sequence



<http://www.genome.gov/sequencingcosts/>

# Next Generation Sequencers generate “reads”

- Raw data
  - Overlapping reads to be assembled
  - Assembly algorithms
- Sequencing depth
  - The number of reads aligned over a given position of the resulting sequence
- Coverage
  - The length of the regions which are covered by reads as a percentage of the expected length of the whole genome

The diagram illustrates three overlapping sequencing reads (CGCTGGAACGG, AACGGCTAGTTA, TAGTTAGAA) aligned to a reference sequence (CGCTGGAACGGCTAGTTAGAA). The reads are represented as dark red rounded rectangles with white text, positioned above the reference sequence which is a black line with black text. The first read covers the first 9 bases, the second read covers bases 4-12, and the third read covers bases 8-16.

```
CGCTGGAACGG
      AACGGCTAGTTA
                TAGTTAGAA
CGCTGGAACGGCTAGTTAGAA
```

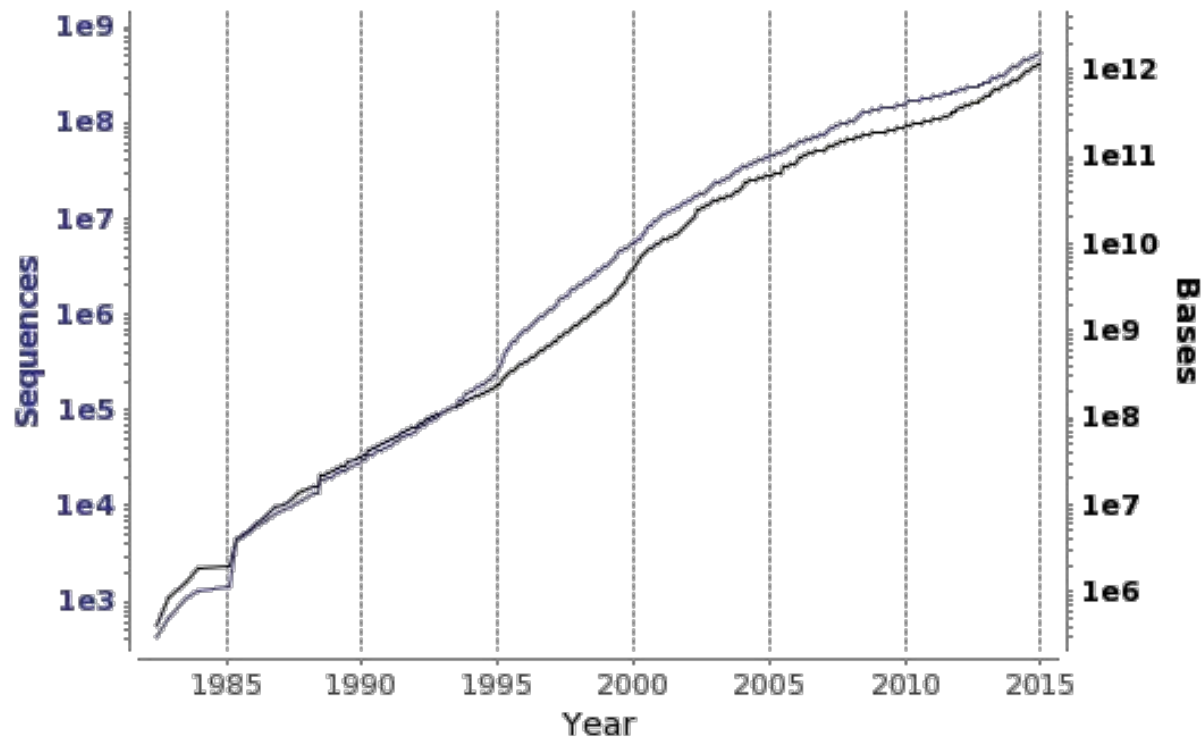


# The International Nucleotide Sequence Database Collaboration

- An annotated collection of all publicly available DNA sequences
  - GenBank at NCBI
  - DNA DataBank of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL)
- Submission of sequences by labs
- These three organizations exchange data on a daily basis
- <http://www.insdc.org/>

## Assembled/annotated sequence growth

19-Jan-2015



— Sequences (520.0 millions) — Bases (1,133.9 billions)

# Pictures & movies : material licensing

p. 3 : Nature Education / Conditions of use for academic research : users may view, print, copy, download and text and data-mine the content

p. 3 : [By Konrad Förstner](#) – Public domain – From Wikipedia