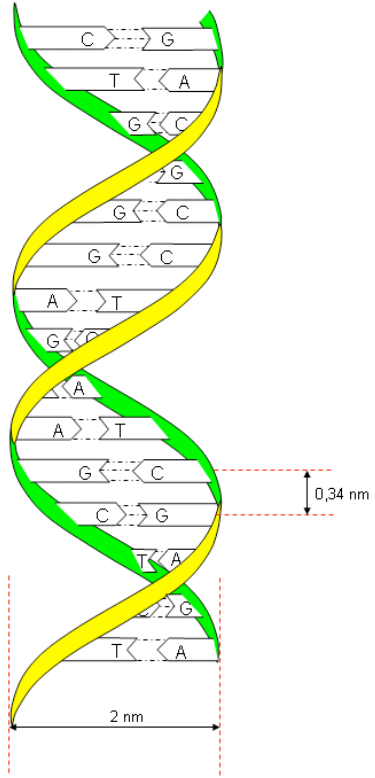


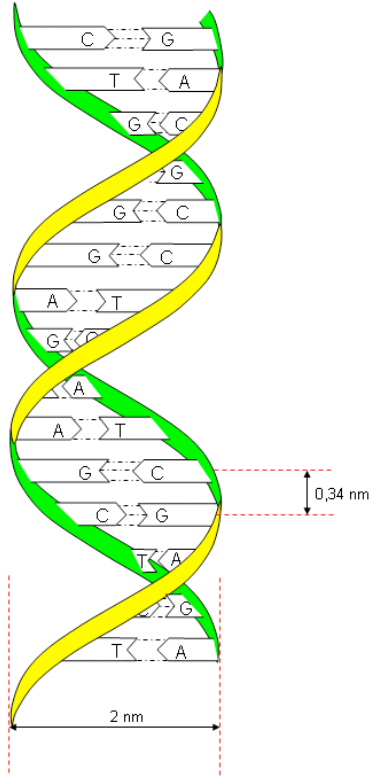
2. Gènes et protéines

- La séquence est-elle un bon modèle de l'ADN ?
- Les gènes, de Mendel à la biologie moléculaire
- Le code génétique
- Un algorithme de traduction
- Implémenter le code génétique
- Algorithmes + structures de données = programmes
- Les compromis de la conception d'algorithmes
- **Les technologies de séquençage de l'ADN**
- Le séquençage de génomes complets
- Comment trouver les gènes ?

Le séquençage de l'ADN

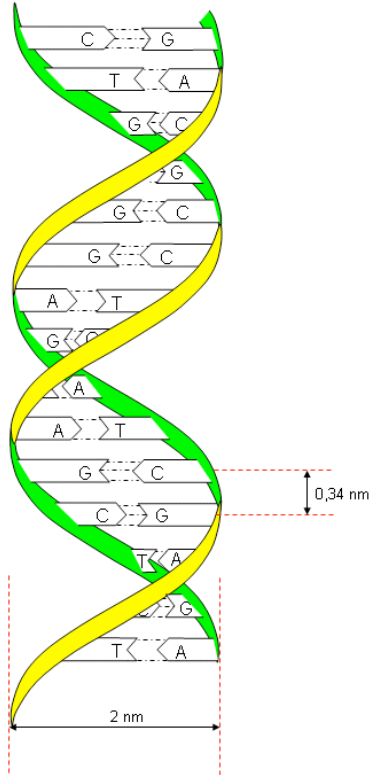


Le séquençage de l'ADN



...
CACCTCACTACGGGTCAGGGGGGAAA
GGGGA ACTGAGAGATTTGCAGTGTGA
GAAGCAGTCCCAGGAGTTAGAAGTAG
TGGCTCCATGACTCACAAATTA ACTT
CCCTTTCAGGCAGGGCTTCTTATTTT
CCTTAGCATCCCTGTCTTGATCC CAG
CCTGCTCAGACCCCTGCCTCTCACTG
CAAGATGTGCTTGAGTATGAGAGTCA
GGAATGTTACTTCTCAGAGGCGCCAA
ATGGCAGTTGTACAGGGTCATCATA
GAGGGTATATGTTTACTGCACTGGGC
TCTGAGGCTTGCTTGTGAAGAAACAG
AAGCTAAGGGATCCAGGGAGTCCCAA
CTTAGAGAGTCCCACAGGCCCACT
CTGGTTCTGTTGGCAGGAAAATTTGC
TGAATTGGGGCAGGAAGTTGTGTAAC
AAAACGATTACATCCATTTTTTGCAAG
GCAAGAGTGAGCTATTCACCTCCATG
TTGGTGATATTTTTTTGCCATATAAGC
AGCTAATTCCTTTCAGTAATTCTACT
CTAAACTAGTCTTAATGTGACTTCTA
TATAAATTCTGAACTGAATAATTTTG
GGAACGTTGTAAAAA...

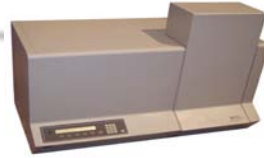
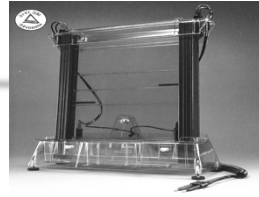
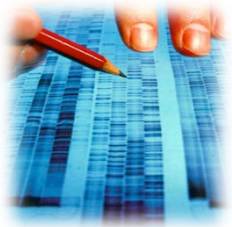
Le séquençage de l'ADN



...
CACCTCACTACGGGTCAGGGGGGAAA
GGGGA ACTGAGAGATTTGCAGTGTGA
GAAGCAGTCCCAGGAGTTAGAAGTAG
TGGCTCCATGACTCACAAATTA ACTT
CCCTTTCAGGCAGGGCTTCTTATTTT
CCTTAGCATCCCTGTCTTGATCCAG
CCTGCTCAGACCCCTGCCTCTCACTG
CAAGATGTGCTTGAGTATGAGAGTCA
GGAATGTTACTTCTCAGAGGCGCCAA
ATGGCAGTTGTACAGGGTCATCATA
GAGGGTATATGTTTACTGCACTGGGC
TCTGAGGCTTGCTTGTGAAGAAACAG
AAGCTAAGGGATCCAGGGAGTCCCAA
CTTAGAGAGTCCCACAGGCCCACT
CTGGTTCTGTTGGCAGGAAAATTTGC
TGAATTGGGGCAGGAAGTTGTGTAAC
AAAACGATTACATCCATTTTTTGCAAG
GCAAGAGTGAGCTATTCACCTCCATG
TTGGTGATATTTTTTTGCCATATAAGC
AGCTAATTCTTTTCAGTAATTCTACT
CTAAACTAGTCTTAATGTGACTTCTA
TATAAATTCTGAACTGAATAATTTTG
GGAACGTTGTAAAAA...

Le séquençage, une technologie dite « exponentielle »

- Les premières séquences ont été obtenues au début des années 70
- Apparition d'une nouvelle génération de séquenceurs aux alentours de 2008 : NGS, *Next Generation Sequencing*



1990: 10^3 bases/jour

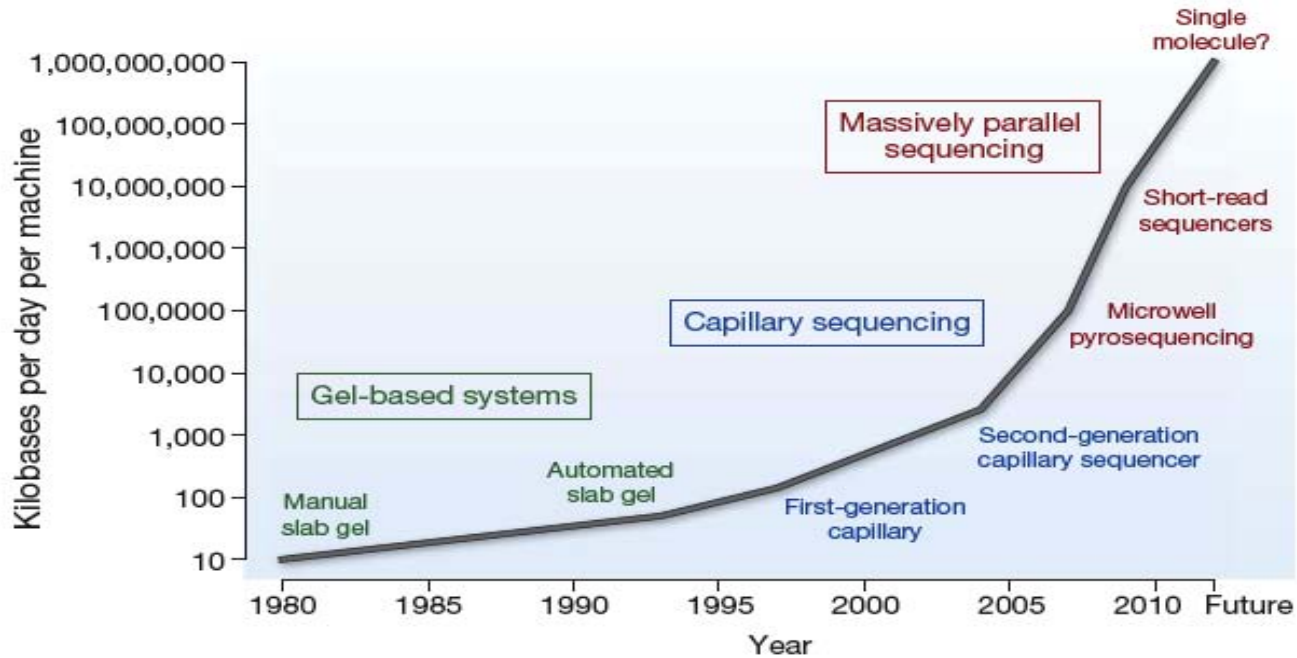


2000: 10^6 bases/jour



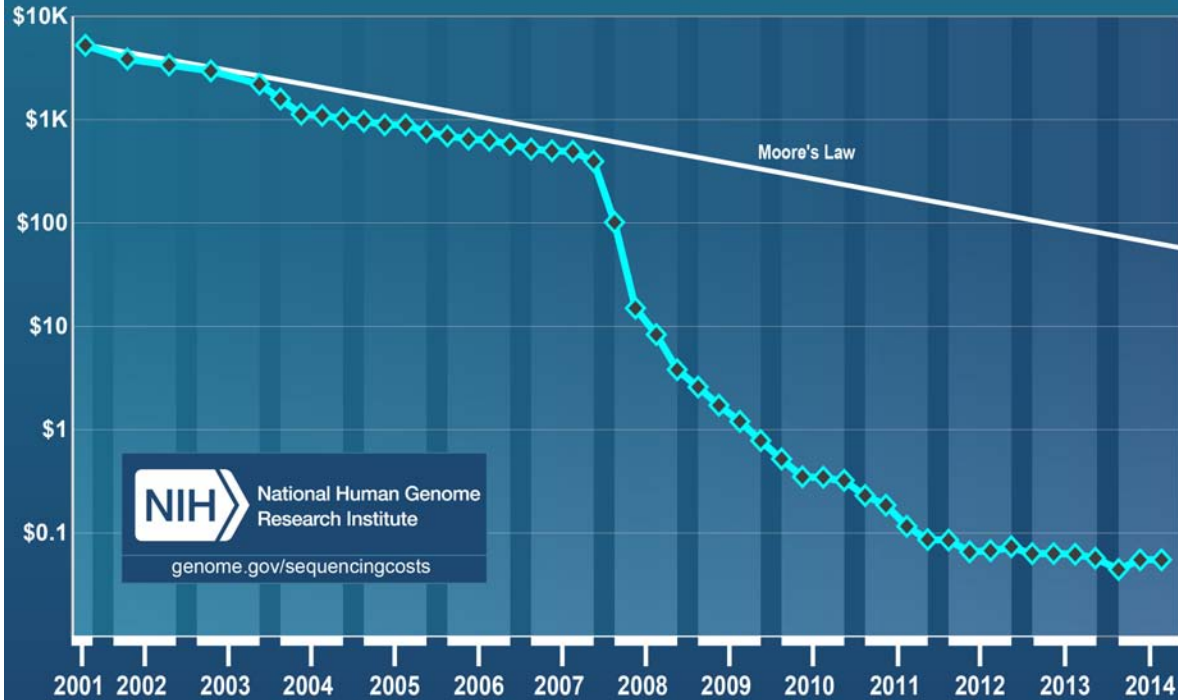
2010: 10^9 bases/jour





“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/>

Les séquenceurs NGS produisent des « reads »

- **Données brutes** (*raw data*)
 - Des *reads* recouvrants
 - Une étape d'assemblage requise
 - Des algorithmes dédiés



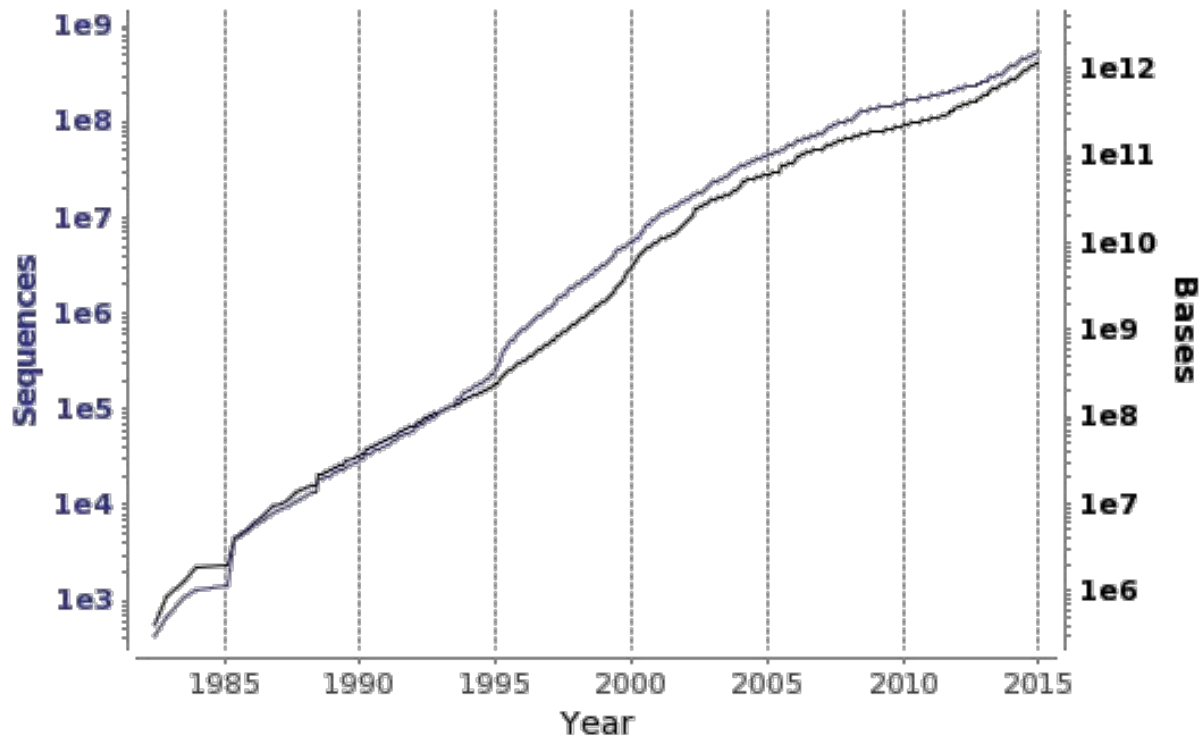
- **Profondeur de séquençage**
 - Le nombre de *reads* alignés sur une position de la séquence assemblée
- **Couverture**
 - La proportion du génome couverte par des *reads*

International Nucleotide Sequence Database Collaboration

- Une collection annotée de toutes les **séquences d'ADN disponibles publiquement**
 - GenBank au NCBI
 - DNA DataBank of Japan (DDBJ)
 - European Molecular Biology Laboratory (EMBL)
- Les séquences y sont déposées directement par les **laboratoires**
- Ces trois organisations s'échangent quotidiennement le contenu de leur **base de données**
 - <http://www.insdc.org/>

Assembled/annotated sequence growth

19-Jan-2015



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

Illustrations & photos : crédits

p. 2, 3 et 4 : Nature Education / Ressource utilisable dans un contexte éducatif

p. 4 : [By Konrad Förstner](#) – Domaine Public – Via Wikipedia

p.6 : Droits réservés