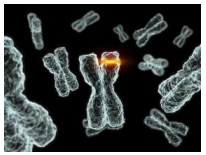
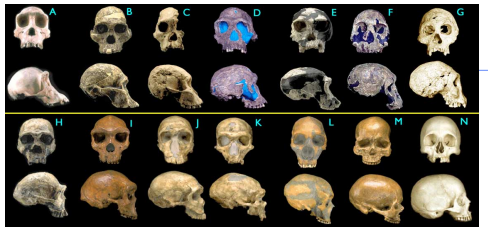
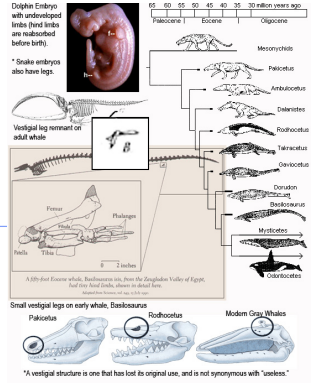
Evolution

(or is it?)



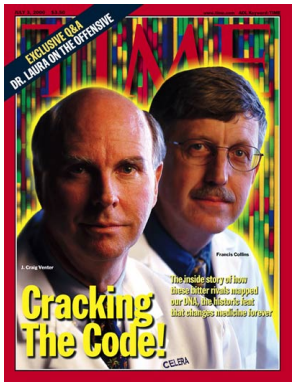
Make-up & Evolution of Genomes



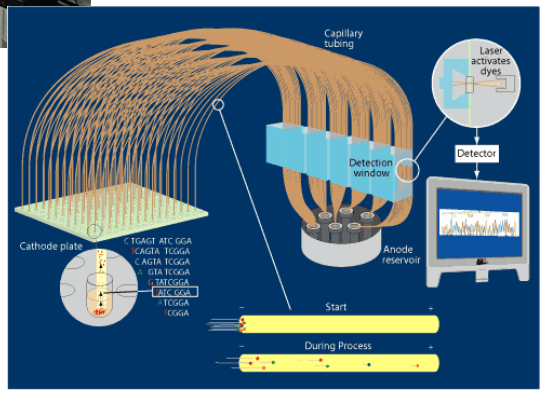


Human Genome Project

- Conducted by two competing entities:
- U.S government project (public)
  - began in 1990
    - estimated to be a 15 year project
  - D.O.E. & N.I.H.
    - initiated by Jim Watson
    - led by Francis Collins
  - goal was to sequence entire human genome
    - 3 billion base pairs
- Celera Genomics (private) set up in 1998
  - Craig Venter challenged gov't
    - Claimed they would do it faster & cheaper



- Big labs!
  - economy of scale
    - Large automated sequencing machines where used to sequence human donor DNA



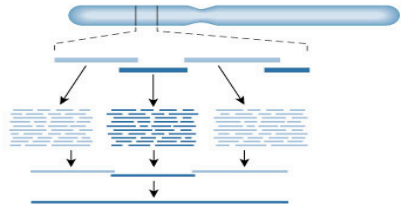
PUBLIC
- 20 large sequencing centers in 6 countries:
  - Joint Genome Institute (DOE)
  - MIT
  - Washington University of St. Louis
  - Baylor College of Medicine
  - Sanger Center (UK)

PRIVATE
- Celera Genomics

Each used Different approaches

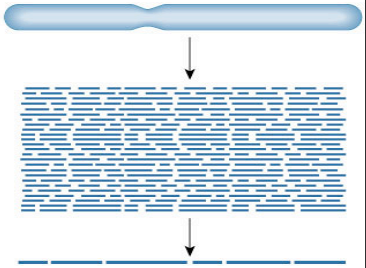
gov't
"map-based

1. Cut DNA segment into fragments which were cloned into YAC or BAC.
2. Then, arrange fragments based on overlapping nucleotide sequences to roughly estimate what the chromosome sequence must be.
2. Then, take each fragment and cut it into smaller fragments and clone each smaller fragment using phages or plasmids.
3. Then, determine the DNA sequences of these smaller fragments already with an idea of where they are located in a chromosome.
   - Carefully mapped smaller fragments whose general location was known in reference to other larger fragments
4. More precise way to figure out the exact sequence of every chromosome, but a slower method

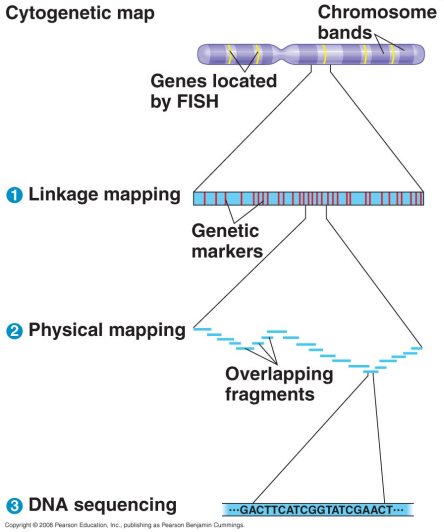


Craig Venter's method
"shotgun method"

1. Cut DNA entire chromosomes into small fragments and clone each.
2. Sequence each segment & arrange based on overlapping nucleotide sequences once sequenced.
3. Faster way of determining the sequence of a whole chromosome, but messier with increased chances of making mistakes determining the full DNA sequences of each chromosome.



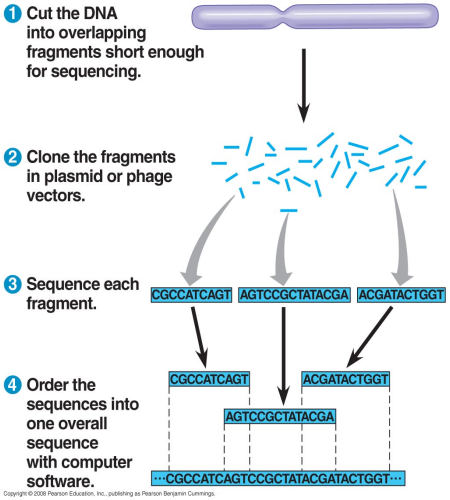
Government Human Genome Project:
Three-Step Approach



Began with Cytogenic maps
- Showed chromosome banding pattern by staining chromatin
- Showed location of some heavy gene areas

1. **Creation of Linkage Map**
   - Order genetic markers (based on recombination frequencies)
2. **Creation of Physical Map**
   - Order large cloned overlapping fragments (in BAC/YAC) - by sequencing just the ends of the fragments or using probes
     - Cut fragments one by one then into still smaller fragments and cloned in plasmids and phages
       - Ordered smaller overlapping smaller fragments
3. **Conduct DNA Sequencing**
   - Determine nucleotide sequence of each small fragment

Celera:
Whole Genome "shotgun" approach



- **Positive:**
  - More efficient and less expensive than three-step approach.
- **Negative:**
  - May cause you to miss certain duplicated sequences and underestimate the size of the genome or miss genes in these regions.

Celera relied heavily on the precise data published by the public project constantly to double-check their sequencing results and make sure they were on the right track!

Human Genome Project

The public project published all their wok throughout the project and Celera used this information to check that they were on the right track.

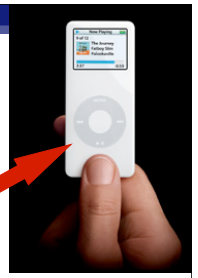
On June 26, 2001, HGP published the "working draft" of the DNA sequence of the human genome. By 2003 most of it was completed.

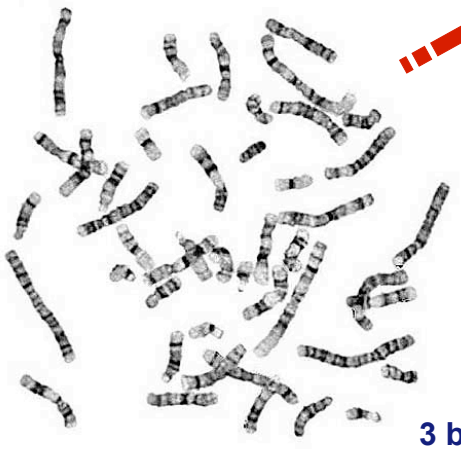
Historic Event!

- figured out the blueprint of a human
  - Now scientists, knowing the entire sequence of all chromosomes, could hunt for genes within the sequences in order to find out which genes exist, where they are located, what RNA or Proteins they code for, and what these molecules do inside cells.
    - the potential to change science & medicine


The leaders of the private and public genome projects, Craig Venter and Francis Collins.

Sequence of 46 Human Chromosomes



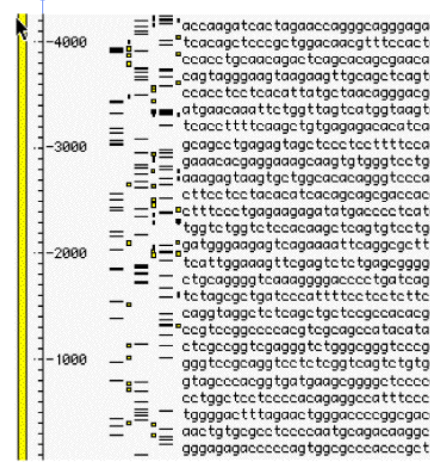


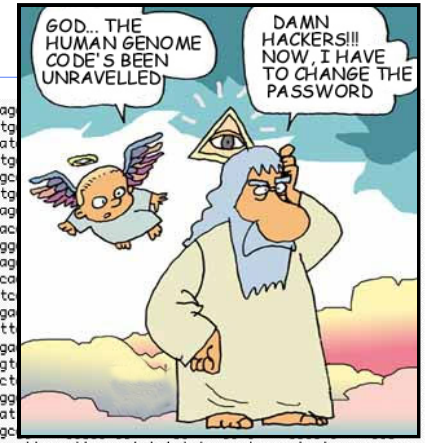
3G of data

3 billion base pairs

TACGCACATTTACGTACGCGGATGCCGCGACTATGATC
ACATAGACATGCTGTCAGCTCTAGTAGACTAGCTGACT
CGACTAGCA **human genome** AGCACACYC
GTACATCGA **3.2 billion bases** TACATGCTA
CTAGCTACTGACTCATGATCCAGATCACTGAAACCCTA
GATCGGGTACCTATTACAGTACGATCATCCGATCAGAT
CATGCTAGTACATCGATCGATACTGCTACTGATCTAGC
TCAATCAAACTCTTTTTGCATCATGATACTAGACTAGC
TGACTGATCATGACTCTGATCCCGTAGATCGGGTACCT
ATTACAGTACGATCATCCGATCAGATCATGCTAGTACA
TCGATCGATACTGCTACTGATCTAGCTCAATCAAACTC
TTTTTGCATCATGATACTAGACTAGCTGACTGATCATG
ACTCTGATCCCGTAGATCGGGTACCTATTACAGTACGA
TCATCCGATCAGATCATGCTAGTACATCGATCGATACT

Raw genome data can be overwhelming for a scientific to work with





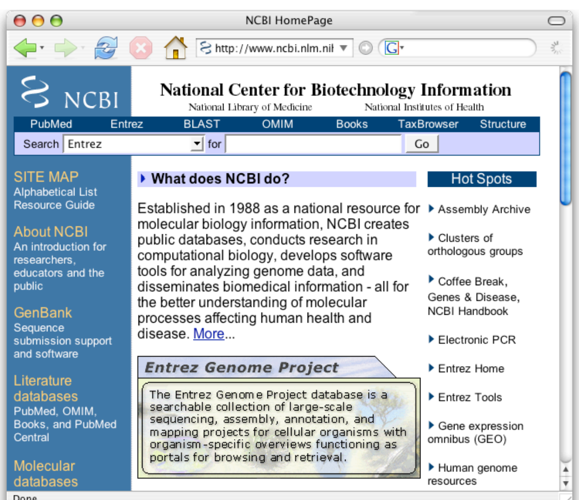
Needed a way to organize all this data in an orderly accessible manner.

Bioinformatics: application of computation methods to the storage and analysis of biological data.

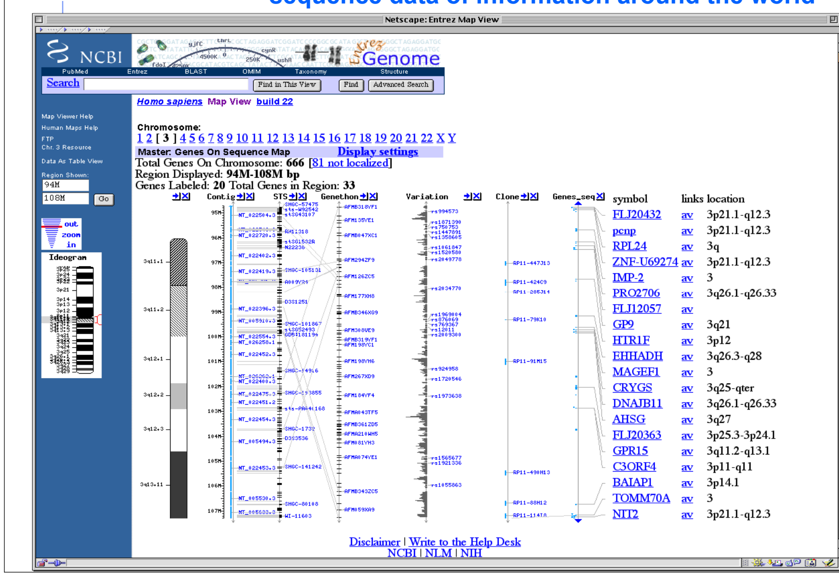
Government created the NCBI GenBank:

A database of genetic sequences gathered from research around the world and organized in an easy to use way.

Publicly available on Web!

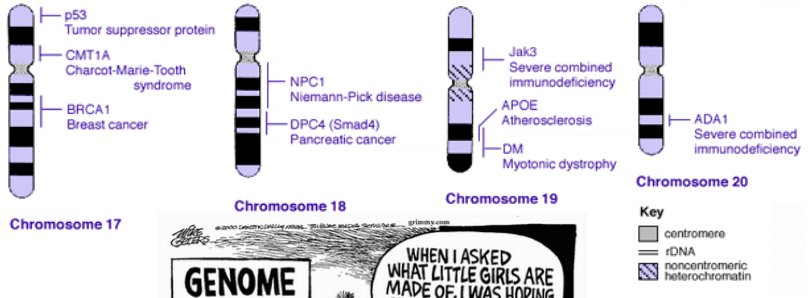


NCBI GenBank: Organizes new data & speeds up sharing of genetic sequence data of information around the world



Maps of human genes now exist.

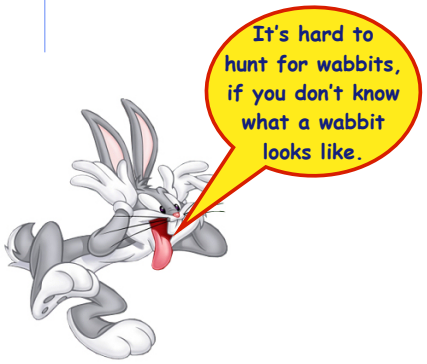
- We know exactly where the genes are…
  - and can now compare wild type & their mutant alleles



We started by defining a gene…

"Defining a gene is problematic because... one gene can code for several protein products, some genes code only for RNA, two genes can overlap, and there are many other complications."

– Elizabeth Pennisi, Science 2003



gene → protein

gene → RNA

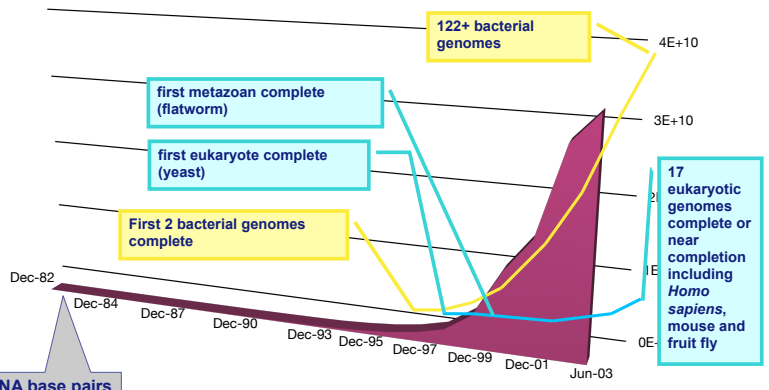
gene → polypeptide 1

gene → polypeptide 2

gene → polypeptide 3

And we didn't stop there…

The Progress - In the last 5 years or so the amount of data has grown exponentially, including the growth of online databases and resources.



Data from NCBI and TIGR (www.ncbi.nlm.nih.gov and www.tigr.org)