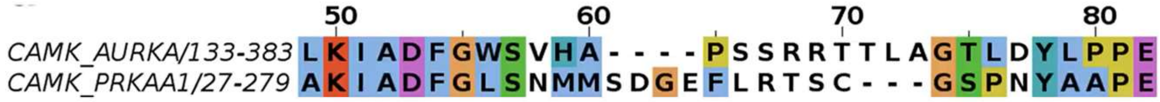
# Historical Events

## The Protein Sequence and Structure Wave

- 1955: Sanger sequenced bovine insulin

## The Protein Sequence and Structure Wave

- 1955: Sanger sequenced bovine insulin
- 1970: Needleman–Wunsch algorithm


50
60
70
80
CAMK_AURKA/133-383 LKIADFGWSVHA----PSSRRTTLAGTLDYLPPE
CAMK_PRKAA1/27-279 AKIADFGLSNMMSDGEFLRTSC---GSPNYAAPE

## The Protein Sequence and Structure Wave

- 1955: Sanger sequenced bovine insulin
- 1970: Needleman–Wunsch algorithm
- 1973: PDB

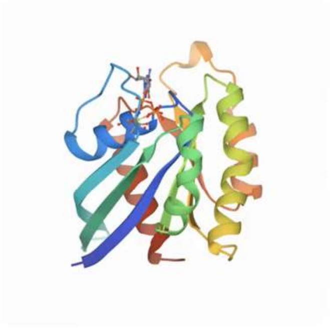
RCSB PDB PROTEIN DATA BANK

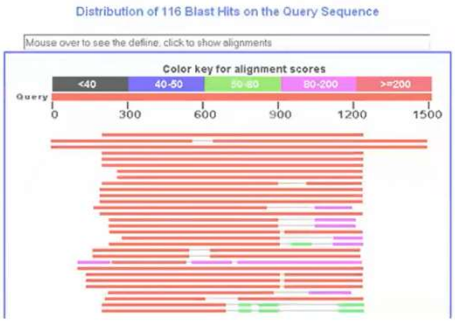
173754 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

# The Protein Sequence and Structure Wave

- 1955: Sanger sequenced bovine insulin
- 1970: Needleman–Wunsch algorithm
- 1973: PDB
- 1990: BLAST

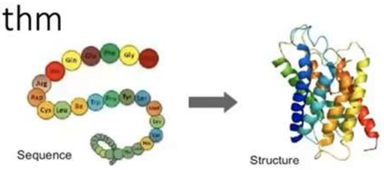
RCSB PDB PROTEIN DATA BANK 173754 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education



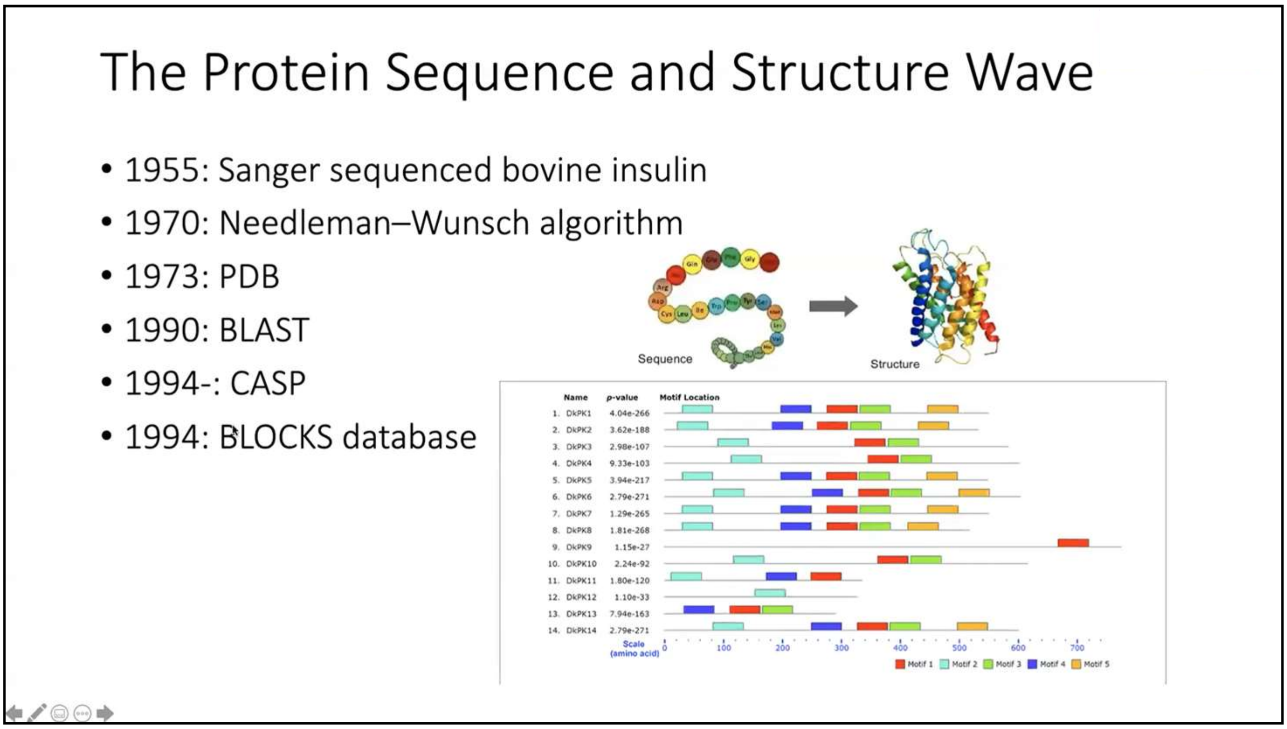

Distribution of 116 Blast Hits on the Query Sequence
Mouse over to see the define, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
300
600
900
1200
1500

# The Protein Sequence and Structure Wave

- 1955: Sanger sequenced bovine insulin
- 1970: Needleman–Wunsch algorithm
- 1973: PDB
- 1990: BLAST
- 1994-: CASP

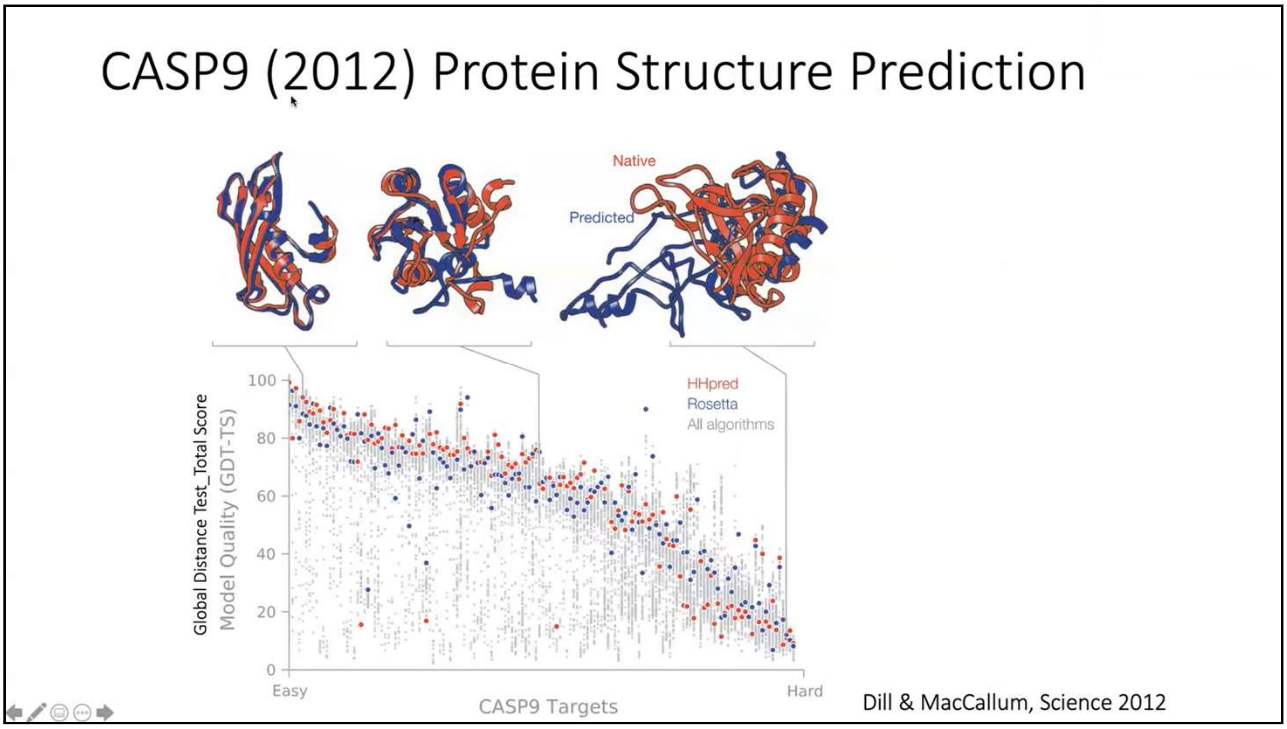


CASP (Critical Assessment of Structure Prediction) is a community wide experiment to determine and advance the state of the art in modeling protein structure from amino acid sequence. Every two years, participants are invited to submit models for a set of proteins for which the experimental structures are not yet public. In the latest CASP round, CASP15, nearly 100 groups from around the world submitted more than 53,000 models on 127 modeling targets in 5 prediction categories. Independent assessors then compare the models with experiment. Assessments and results are published in a special issue of the journal PROTEINS



# The Protein Sequence and Structure Wave

- 1955: Sanger sequenced bovine insulin
- 1970: Needleman–Wunsch algorithm
- 1973: PDB
- 1990: BLAST
- 1994-: CASP
- 1994: BLOCKS database

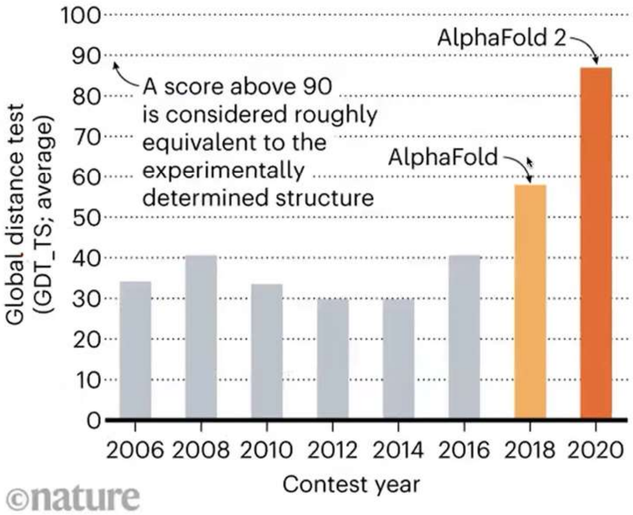


# CASP9 (2012) Protein Structure Prediction

Dill & MacCallum, Science 2012

**STRUCTURE SOLVER**

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



©nature

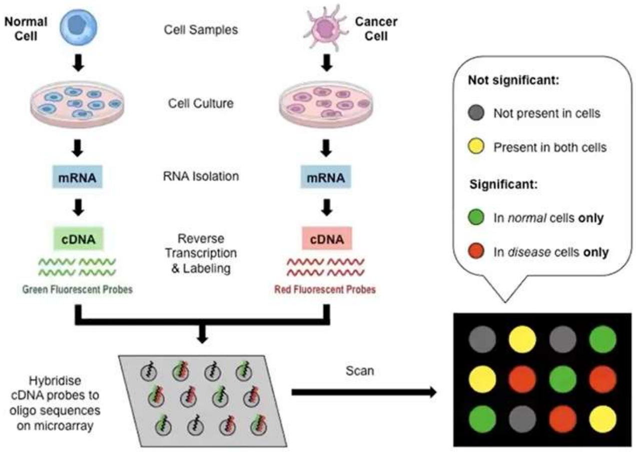
## The Gene Expression Wave

- Northern blot (1977) measures the expression of a single gene

Imagine from BioNinja

# The Gene Expression Wave

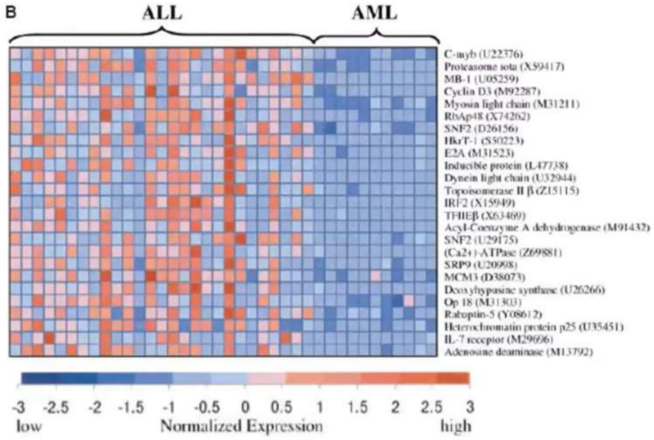
- Northern blot (1977) measures the expression of a single gene
- Microarray (1995) contains hundreds to millions of tiny probes
- Measuring many genes in a condition



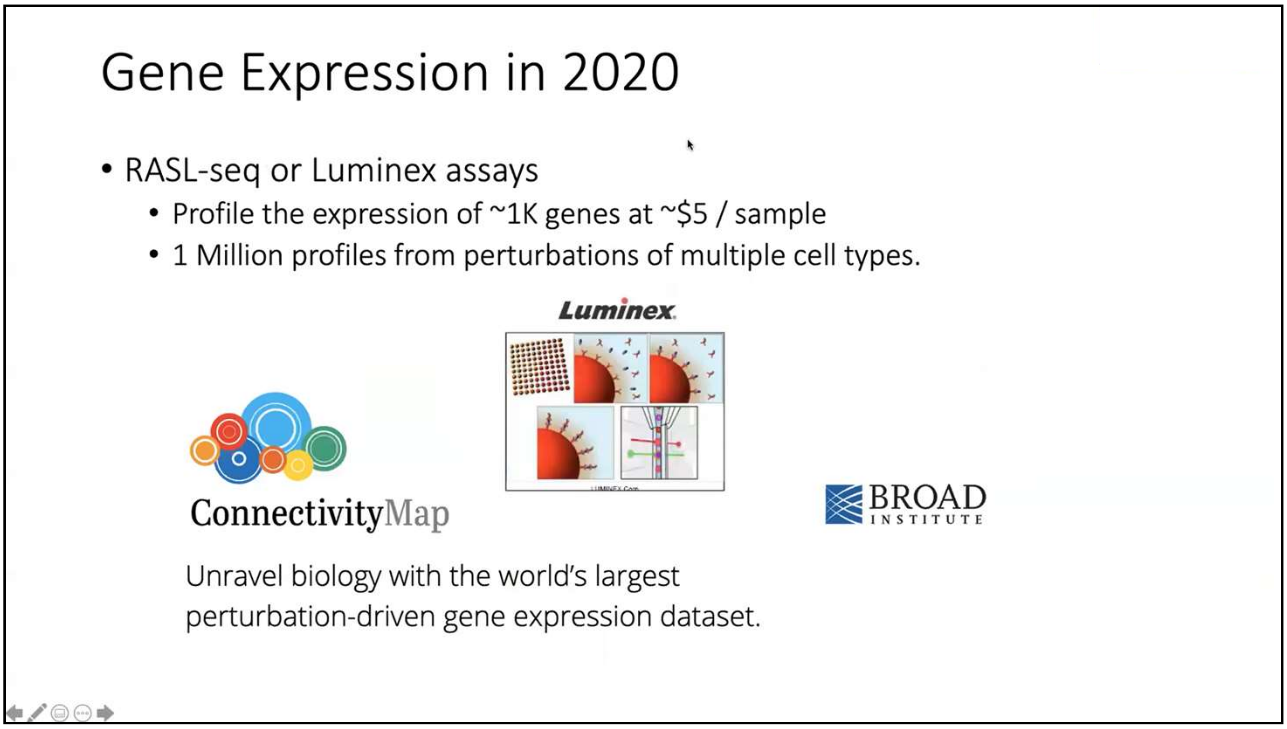
Imagine from BioNinja

# Gene Expression in ~2000

- Distinguishing between acute lymphoblastic leukemia and acute myeloid leukemia



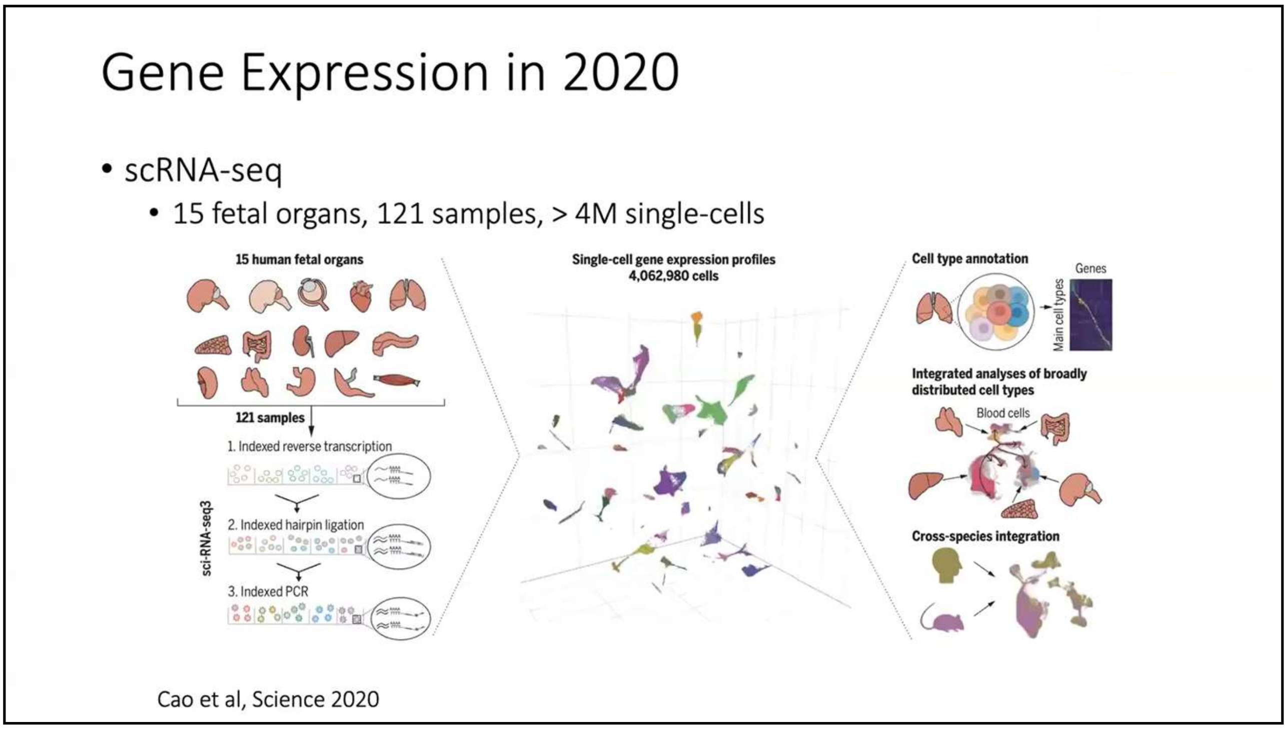
Golub et al, Science 1999.



## Gene Expression in 2020

- RASL-seq or Luminex assays
  - Profile the expression of ~1K genes at ~$5 / sample
  - 1 Million profiles from perturbations of multiple cell types.

Unravel biology with the world's largest perturbation-driven gene expression dataset.

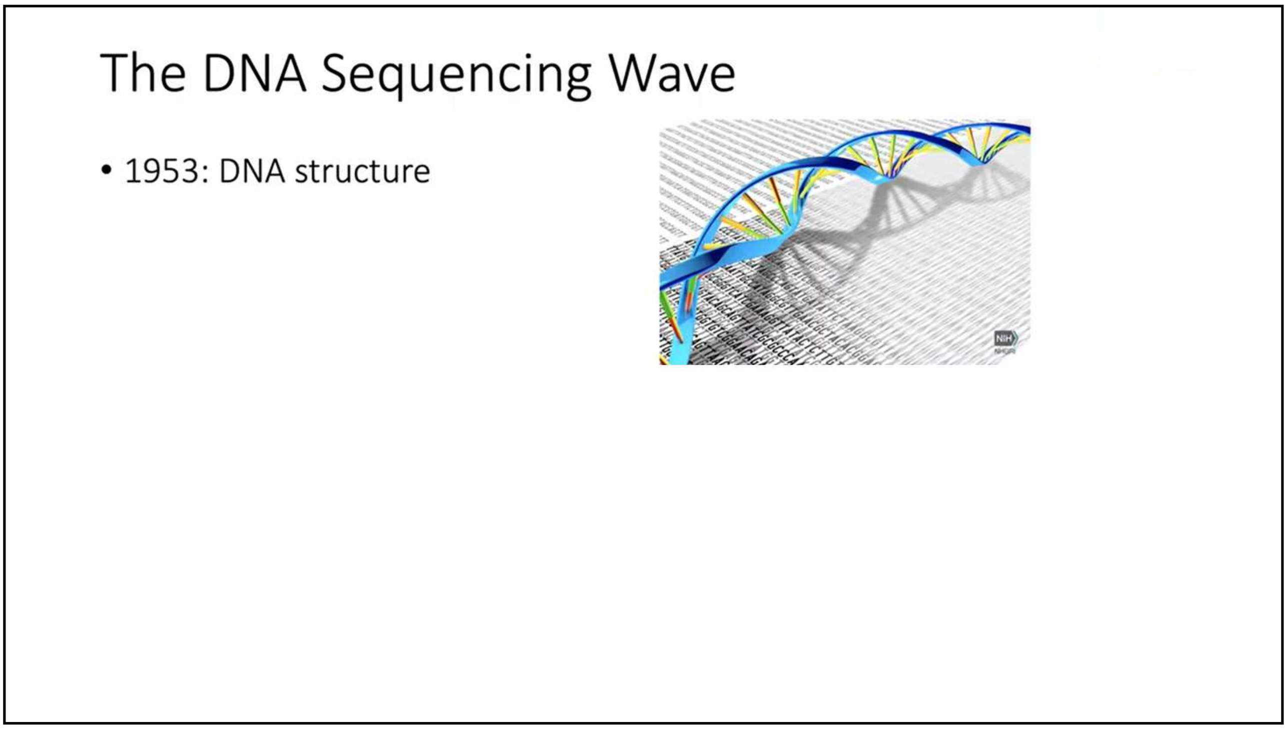


## Gene Expression in 2020

- scRNA-seq
  - 15 fetal organs, 121 samples, > 4M single-cells

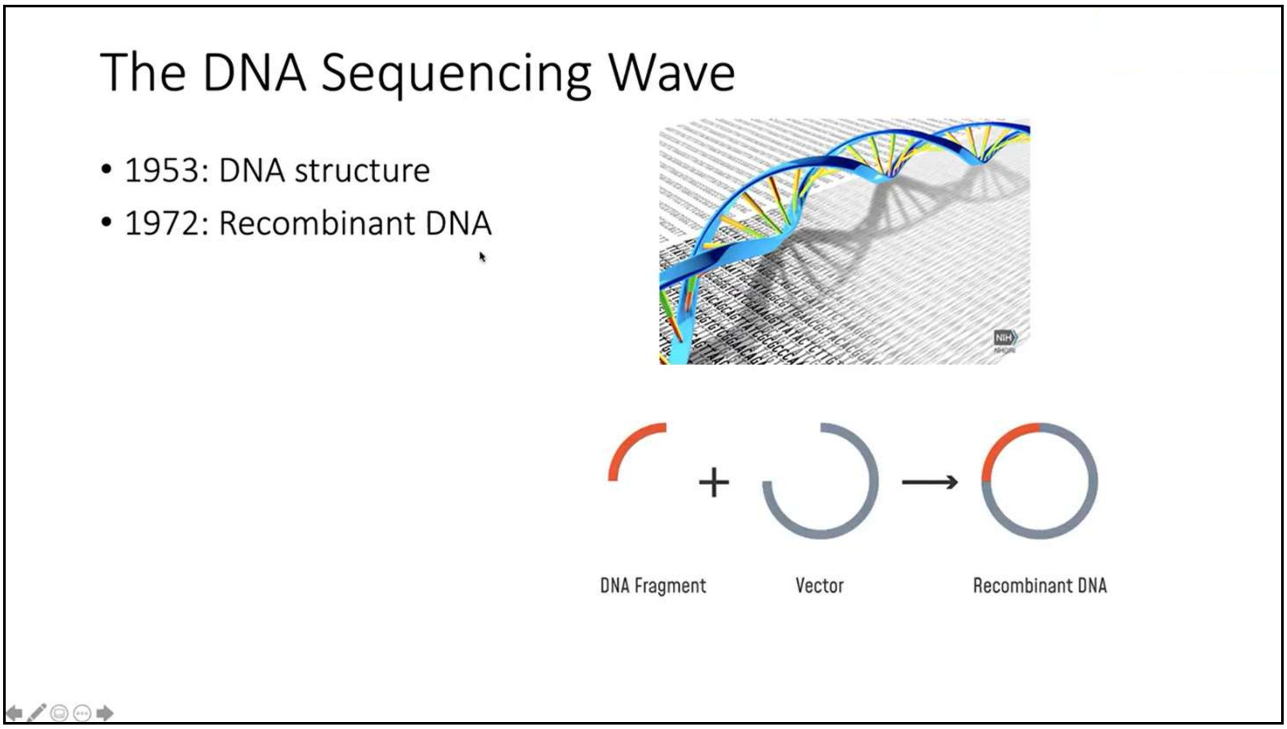
15 human fetal organs
121 samples
sci-RNA-seq3
1. Indexed reverse transcription
2. Indexed hairpin ligation
3. Indexed PCR
Single-cell gene expression profiles
4,062,980 cells
Cell type annotation
Genes
Main cell types
Integrated analyses of broadly distributed cell types
Blood cells
Cross-species integration

Cao et al, Science 2020



## The DNA Sequencing Wave

- 1953: DNA structure

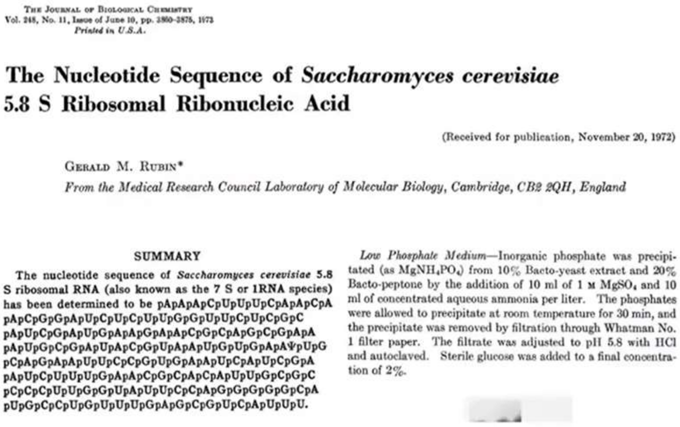


## The DNA Sequencing Wave

- 1953: DNA structure
- 1972: Recombinant DNA

# DNA Sequencing in the 1970s

- 1953: DNA structure
- 1972: Recombinant DNA
- 1977: Sanger sequencing



THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 248, No. 11, Issue of June 10, pp. 3860-3875, 1973
*Printed in U.S.A.*

**The Nucleotide Sequence of *Saccharomyces cerevisiae* 5.8 S Ribosomal Ribonucleic Acid**

(Received for publication, November 20, 1972)

GERALD M. RUBIN*
*From the Medical Research Council Laboratory of Molecular Biology, Cambridge, CB2 2QH, England*

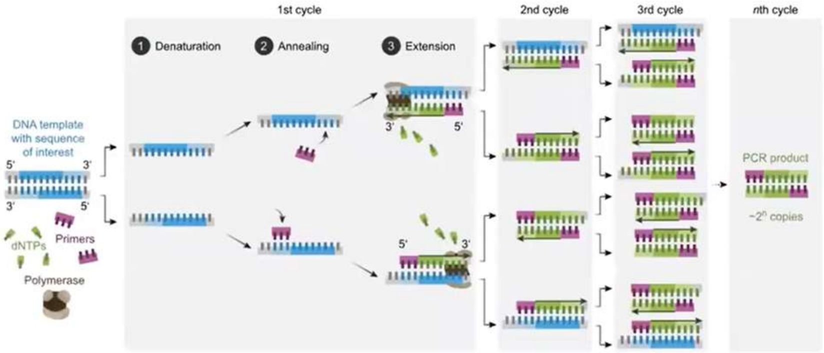
**SUMMARY**

**The nucleotide sequence of *Saccharomyces cerevisiae* 5.8 S ribosomal RNA (also known as the 7 S or 1RNA species) has been determined to be pApApApCpUpUpUpCpApApCpA pApCpGpGpApUpCpUpCpUpUpGpGpUpUpCpUpCpGpC pApUpCpGpApUpGpApApGpApApCpGpCpApGpCpGpApA pApUpGpCpGpApUpApCpGpUpApApUpGpUpGpApAΨpUpG pCpApGpApApUpUpCpCpGpUpGpApApUpCpApUpCpGpA pApUpCpUpUpUpGpApApCpGpCpApCpApUpUpGpCpGpC pCpCpCpUpUpGpGpUpApUpUpCpCpApGpGpGpGpGpCpA pUpGpCpCpUpGpUpUpUpGpApGpCpGpUpCpApUpUpU.**

*Low Phosphate Medium*—Inorganic phosphate was precipitated (as $MgNH_4PO_4$) from 10% Bacto-yeast extract and 20% Bacto-peptone by the addition of 10 ml of 1 M $MgSO_4$ and 10 ml of concentrated aqueous ammonia per liter. The phosphates were allowed to precipitate at room temperature for 30 min, and the precipitate was removed by filtration through Whatman No. 1 filter paper. The filtrate was adjusted to pH 5.8 with HCl and autoclaved. Sterile glucose was added to a final concentration of 2%.

# The DNA Sequencing Wave

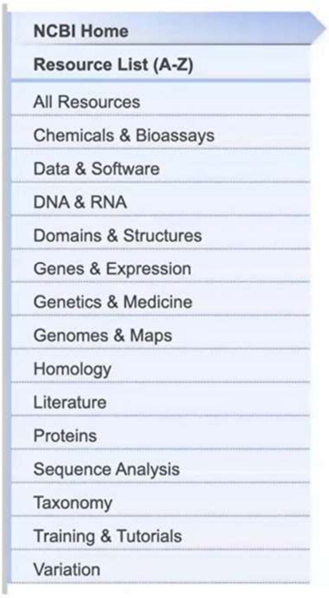
- 1953: DNA structure
- 1972: Recombinant DNA
- 1977: Sanger sequencing
- 1985: PCR



https://en.wikipedia.org/wiki/Polymerase_chain_reaction

# The DNA Sequencing Wave

- 1953: DNA structure
- 1972: Recombinant DNA
- 1977: Sanger sequencing
- 1985: PCR
- 1988: NCBI
- 1990: BLAST



**NCBI Home**

**Resource List (A-Z)**

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

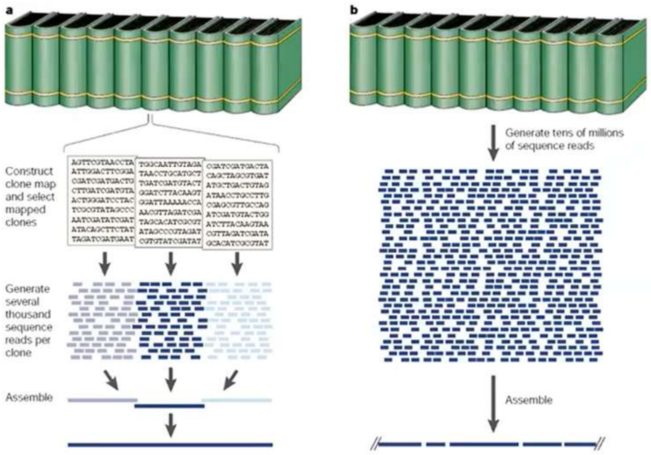
Taxonomy

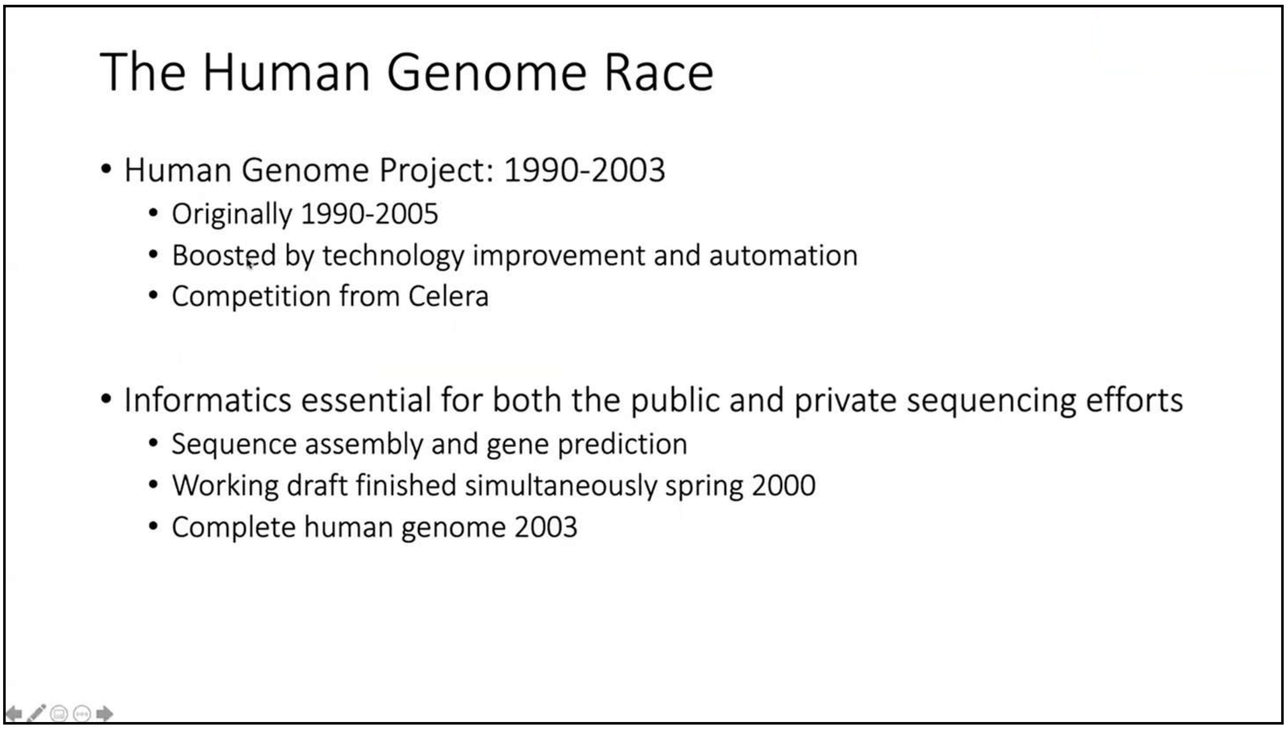
Training & Tutorials

Variation

# The Human Genome Race

- Clone-by-clone (public) vs whole-genome shotgun (private)

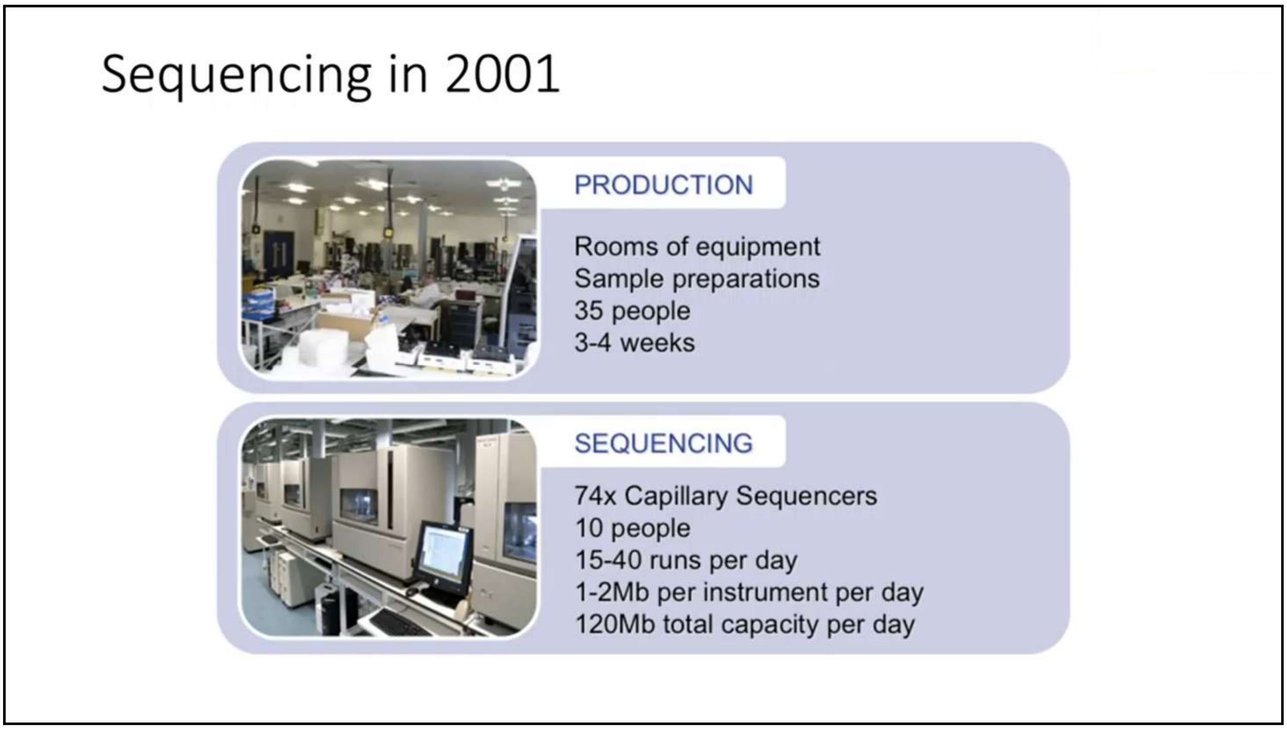

a
Construct clone map and select mapped clones
Generate several thousand sequence reads per clone
Assemble
b
Generate tens of millions of sequence reads
Assemble



## The Human Genome Race

- Human Genome Project: 1990-2003
  - Originally 1990-2005
  - Boosted by technology improvement and automation
  - Competition from Celera

- Informatics essential for both the public and private sequencing efforts
  - Sequence assembly and gene prediction
  - Working draft finished simultaneously spring 2000
  - Complete human genome 2003



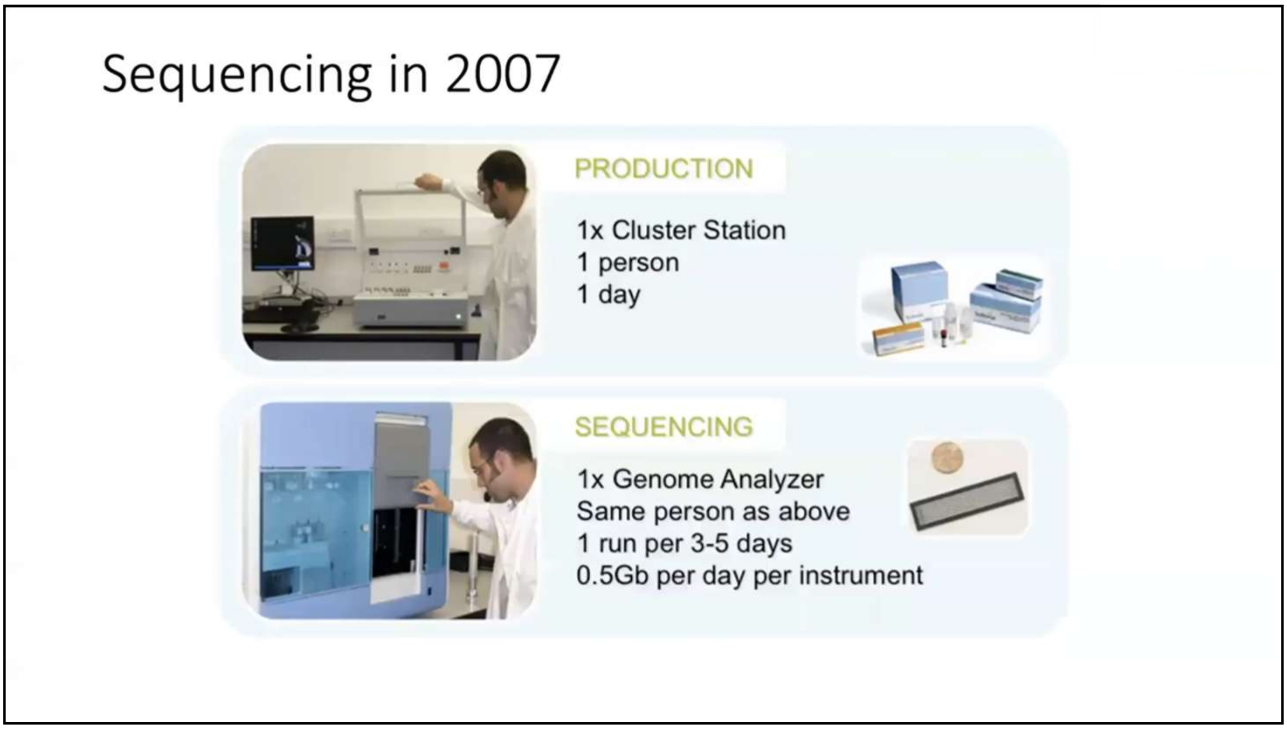
## Sequencing in 2001

PRODUCTION

Rooms of equipment  
Sample preparations  
35 people  
3-4 weeks

SEQUENCING

74x Capillary Sequencers  
10 people  
15-40 runs per day  
1-2Mb per instrument per day  
120Mb total capacity per day



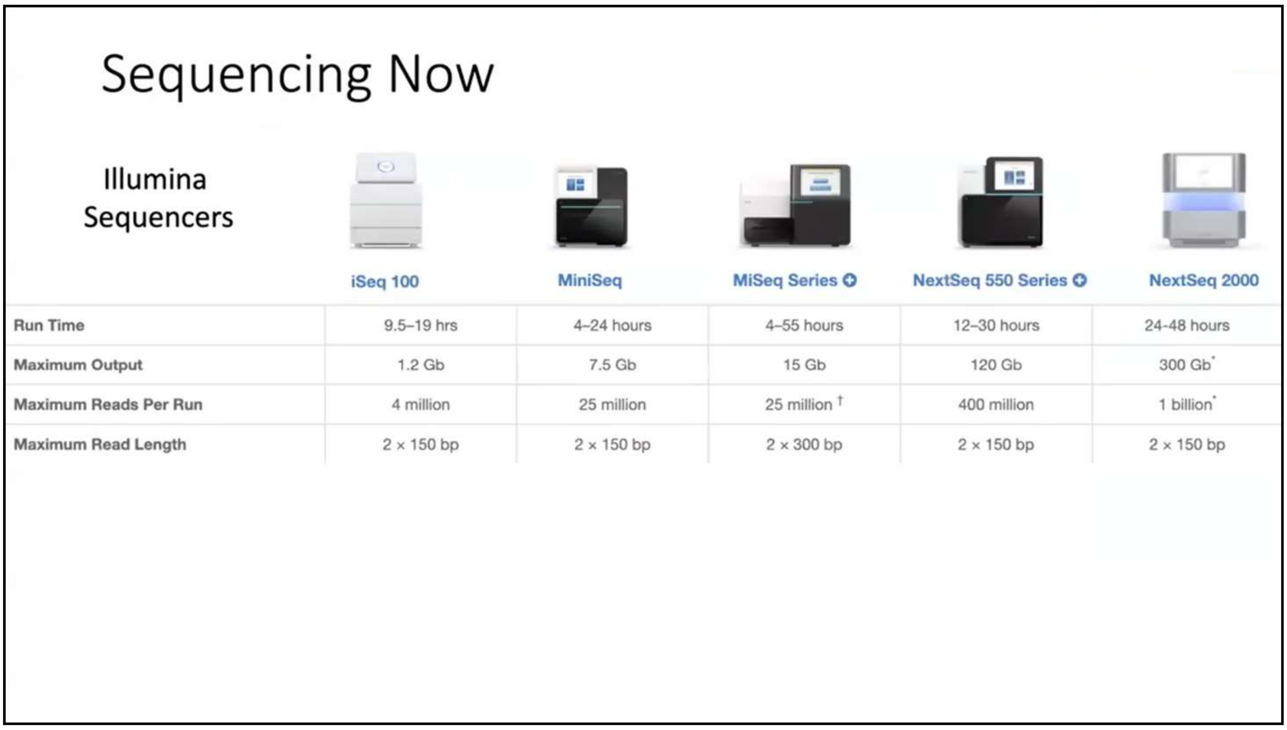
## Sequencing in 2007

PRODUCTION

1x Cluster Station
1 person
1 day

SEQUENCING

1x Genome Analyzer
Same person as above
1 run per 3-5 days
0.5Gb per day per instrument

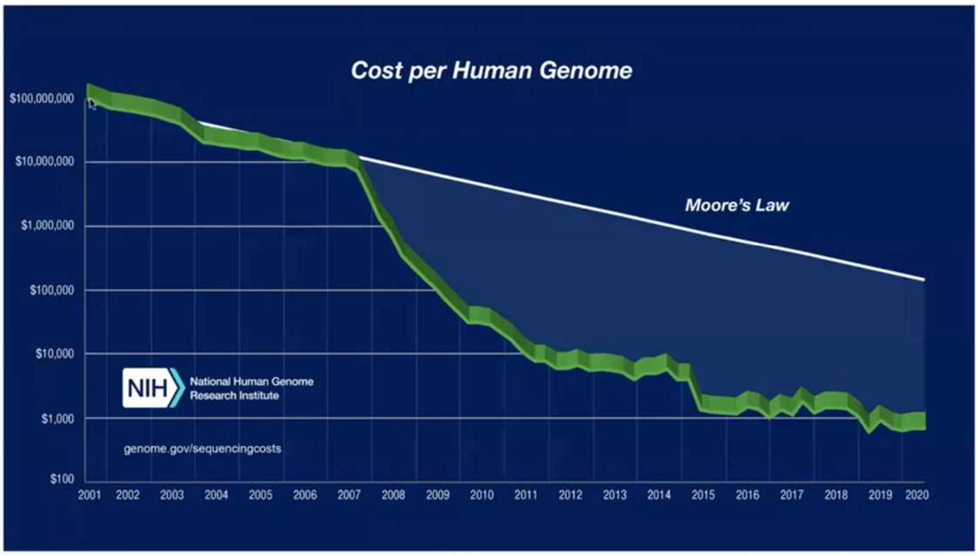


## Sequencing Now

Illumina Sequencers

| | iSeq 100 | MiniSeq | MiSeq Series ⊕ | NextSeq 550 Series ⊕ | NextSeq 2000 |
|---|---|---|---|---|---|
| Run Time | 9.5–19 hrs | 4–24 hours | 4–55 hours | 12–30 hours | 24-48 hours |
| Maximum Output | 1.2 Gb | 7.5 Gb | 15 Gb | 120 Gb | 300 Gb* |
| Maximum Reads Per Run | 4 million | 25 million | 25 million † | 400 million | 1 billion* |
| Maximum Read Length | 2 × 150 bp | 2 × 150 bp | 2 × 300 bp | 2 × 150 bp | 2 × 150 bp |

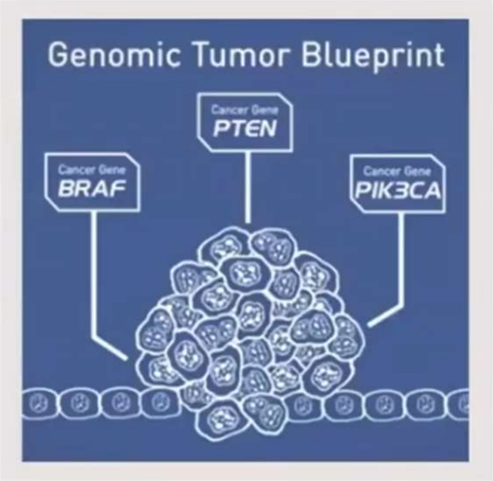
## Cost of Sequencing Human Genomes



## Personalized Disease Prevention and Treatment

**6 Years Have Passed Since Angelina Jolie's Preventative Double Mastectomy — Here's What You Need To Know About Inherited Risk For Breast Cancer**

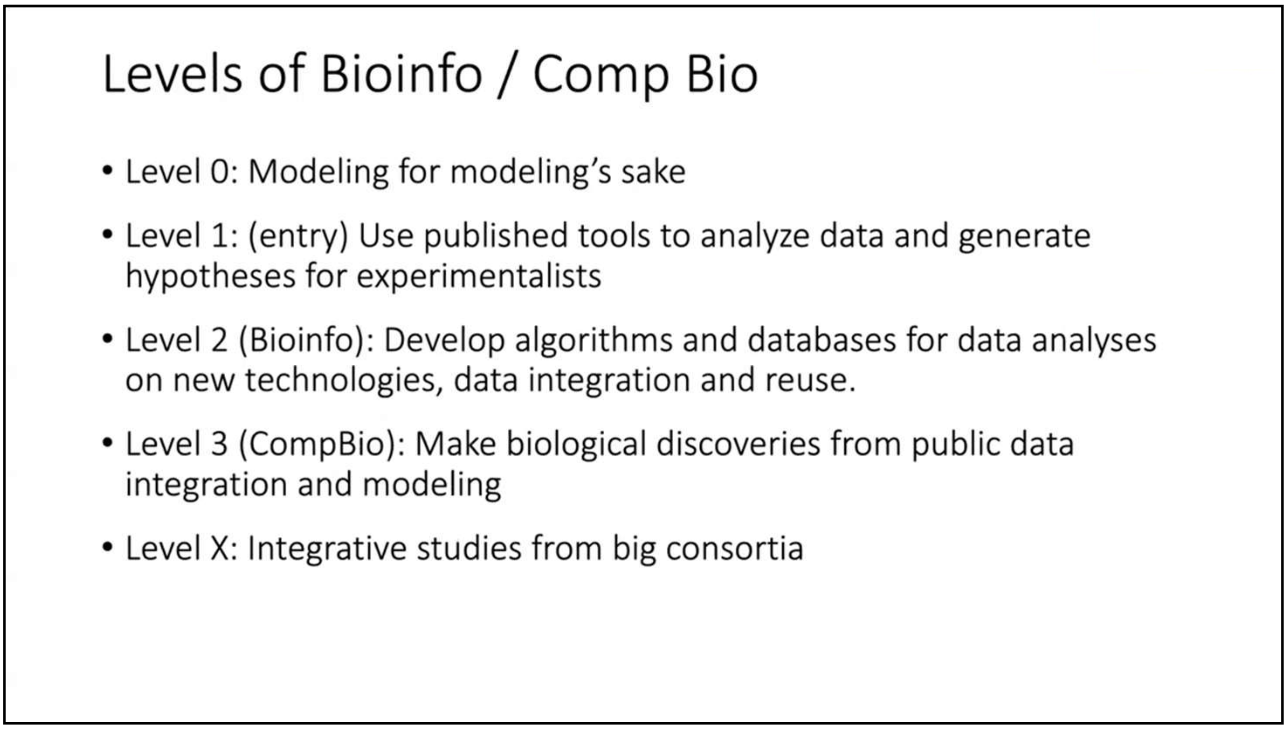
Published Oct 17, 2019



## Big Data Challenges

## Bioinformatics vs Computational Biology?

- Bioinformatics = the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.
- Computational biology = the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about discovery.
- Used interchangeably in this course



## Levels of Bioinfo / Comp Bio

- Level 0: Modeling for modeling’s sake
- Level 1: (entry) Use published tools to analyze data and generate hypotheses for experimentalists
- Level 2 (Bioinfo): Develop algorithms and databases for data analyses on new technologies, data integration and reuse.
- Level 3 (CompBio): Make biological discoveries from public data integration and modeling
- Level X: Integrative studies from big consortia