BIO306: Bioinformatics

Lecture 1
Introduction to Bioinformatics

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech

Goals of this course

- Familiarity with current areas of bioinformatics
- Programing and coding for short reads alignment
- Next generation sequencing data analysis
  - RNA-seq
  - ChIP-seq
  - DNA-seq (Variants calling and gene mapping)
- Other Bioinformatics problems

Pre-requirements

- Major in bioinformatics is preferred
- Familiar with linux
- Learned one programing language

Themes throughout the course: Literature references

You are encouraged to read original source articles. Although articles are not required, they will enhance your understanding of the material.

You can obtain articles through PubMed and Web of Science.

Grading

- Come to lectures
- Participate in lab sessions
- Submit exercises
- Programming (“mid-term exam”)
- Apply learned skills to a conduct bioinformatics project (final project)

Bioinformatics/Computational biology

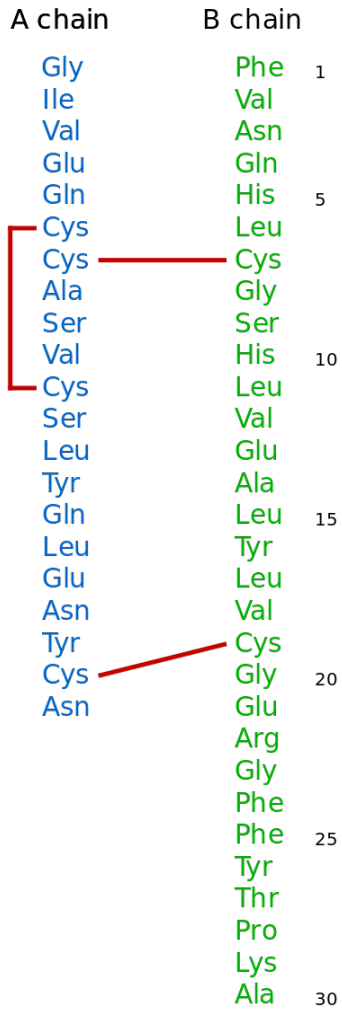
- Interdisciplinary
  - Biology, Mathematics, Statistics, Computer Science
  - Develops methods and software tools for understanding biological data.
- Broad-sense concept
  - bioinformatics is applied statistics and computing to biological science.

Emergence of bioinformatics

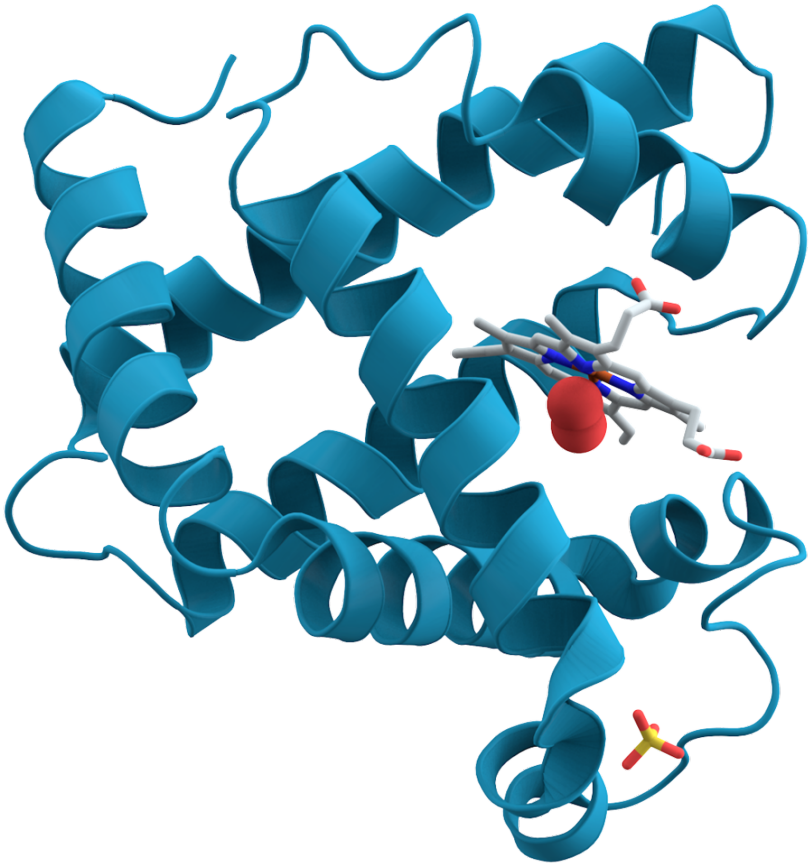
- 1951-1953: Sanger sequenced insulin
- 1960s: Margaret Oakley Dayhoff applied computer to analyze protein data

- 1990: BLAST (NCBI)
- 1990s: rise of bioinformatics
  - The protein sequence and Structure
  - Microarray
  - DNA/RNA sequencing

Protein sequence and Structure

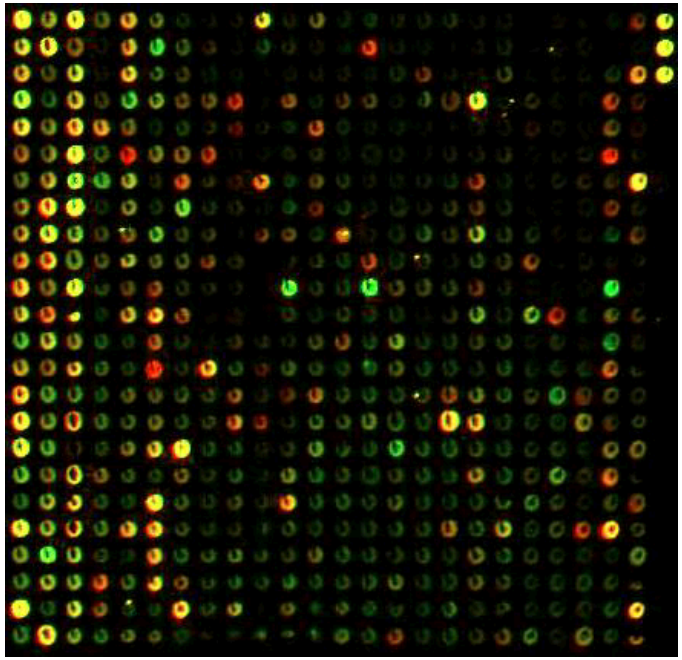


Insulin

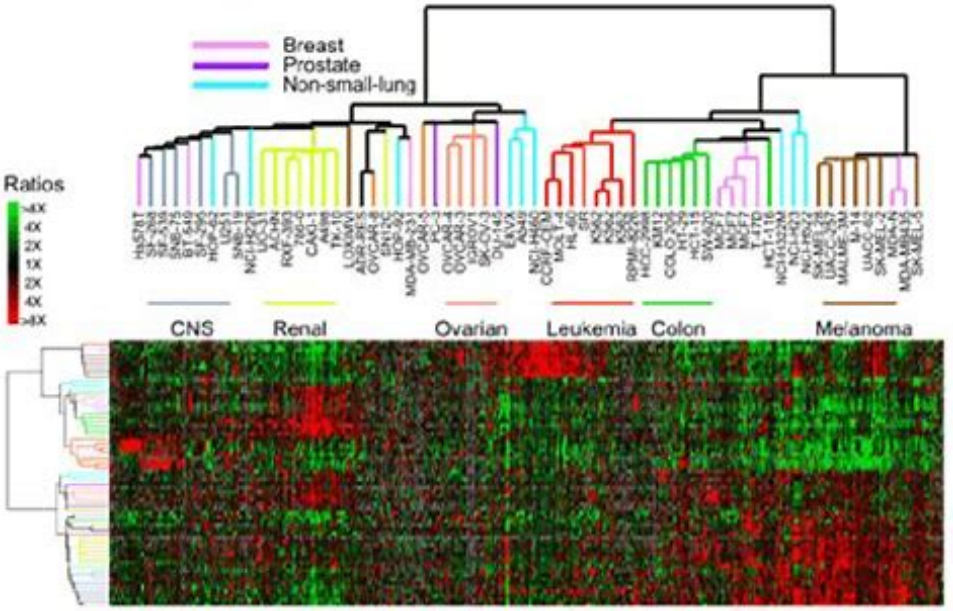


Myoglobin

Microarray



Microarray image

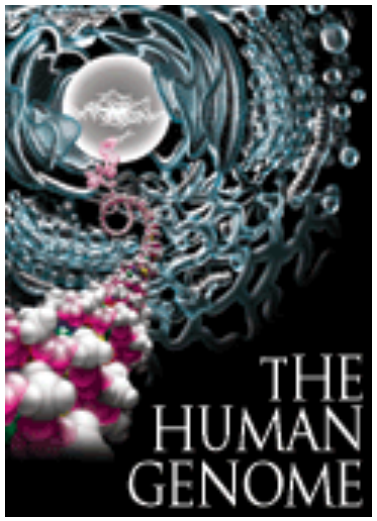


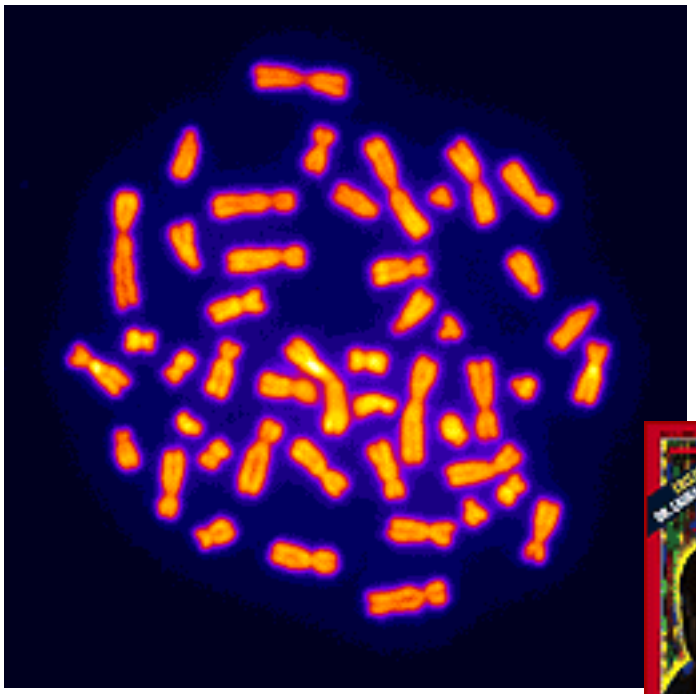
Gene expression clustering

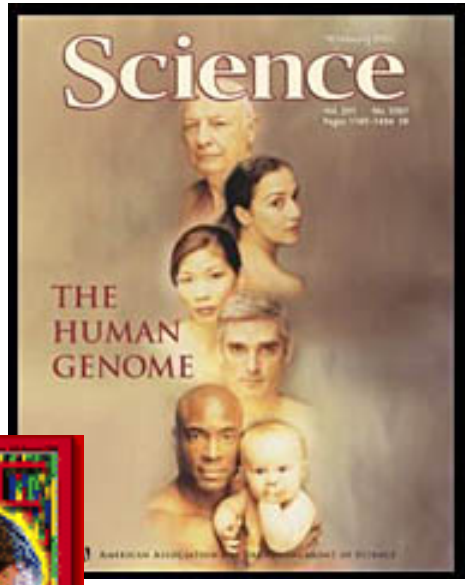
Sequence and genome

The Human Genome sequence is complete

approximately 3.2 billion base pairs

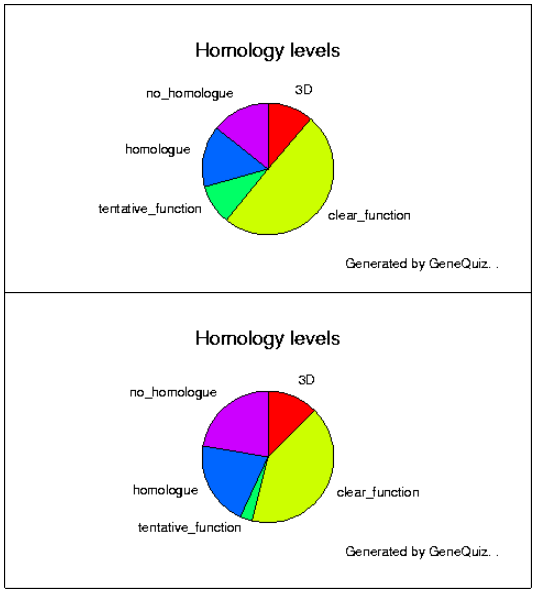


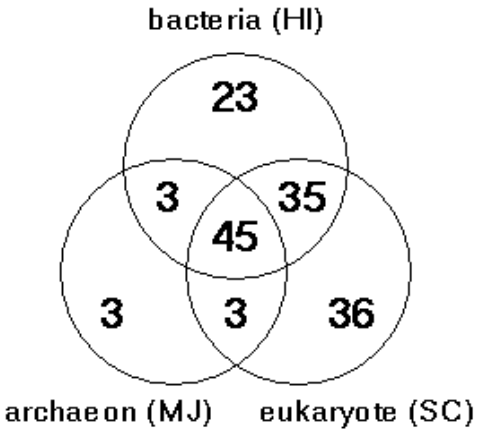




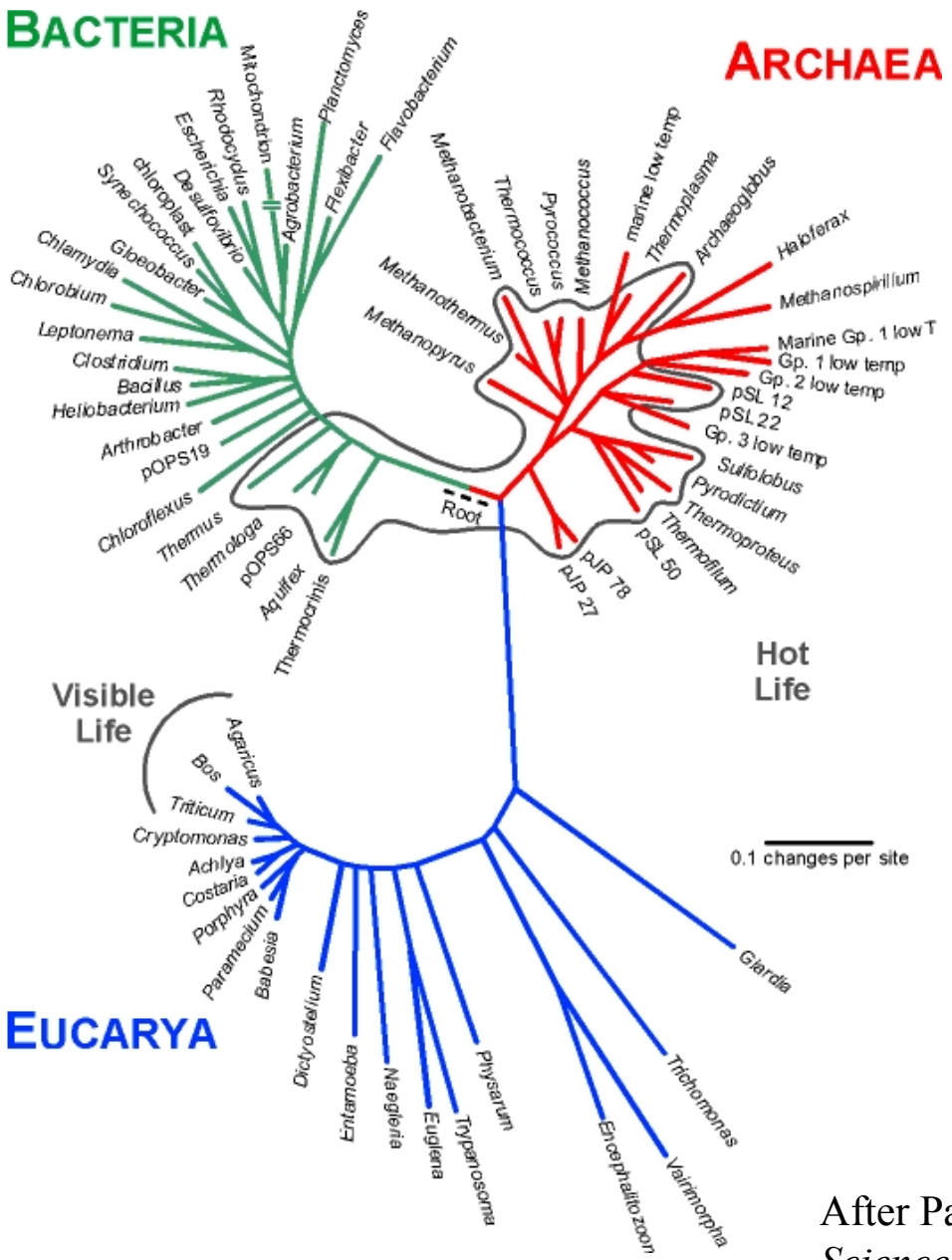
Major Application: Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics





(Clock figures, yeast v. Synechocystis,
adapted from GeneQuiz Web Page, Sander Group, EBI)



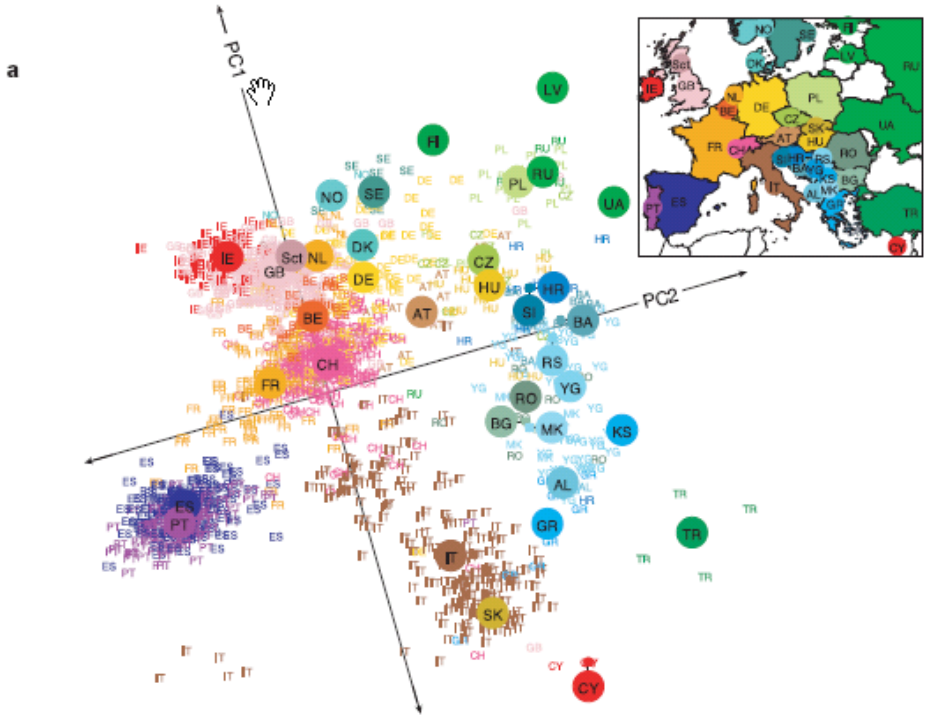
After Pace NR (1997)
Science 276:734

doi:10.1038/nature07331

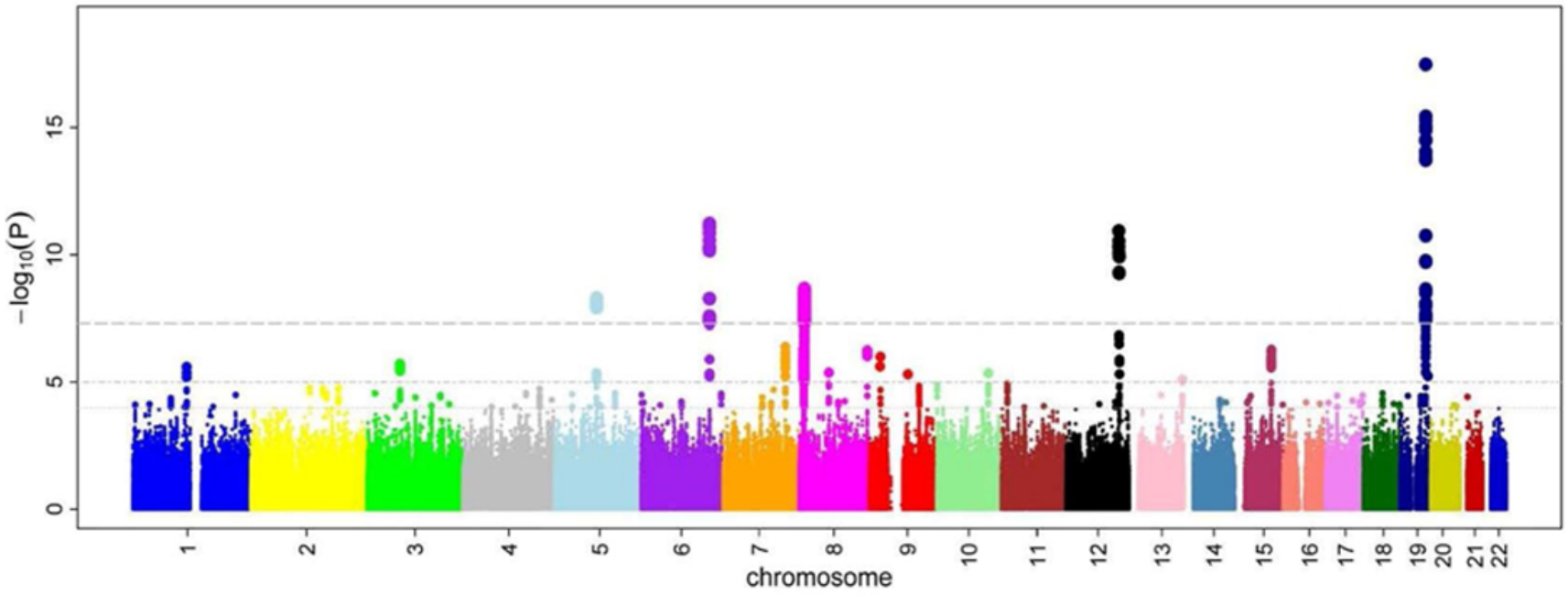
LETTERS

Genes mirror geography within Europe

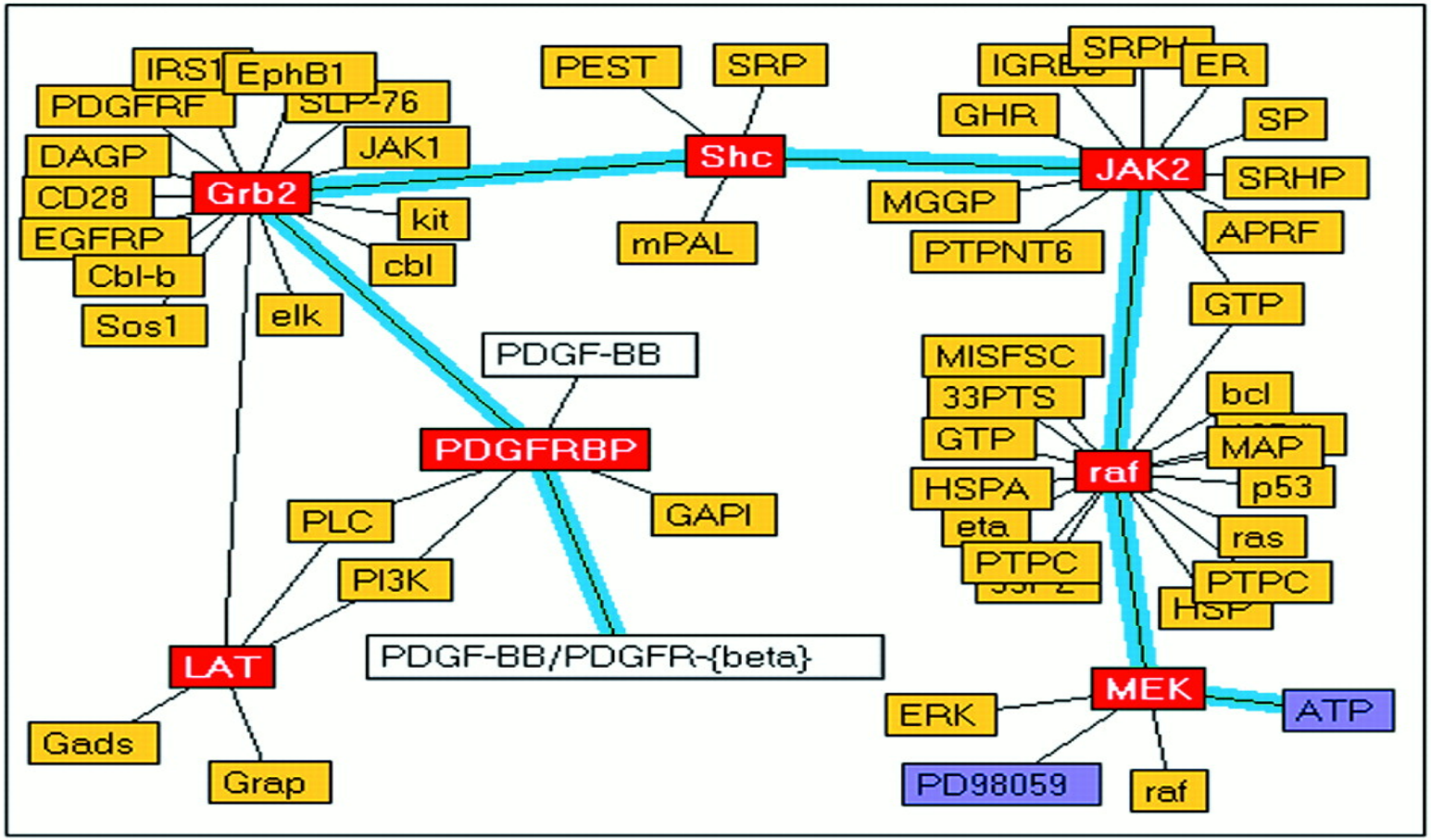
John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]



Gene mapping



Visually Navigate BIND - Microsoft Internet Explorer



Info: protein-tyrosine kinase JAK1 Relax ☐ Show IIDs SeqHound ON

Org: Homo sapiens v1.0.1

Help

Reload

Done Internet

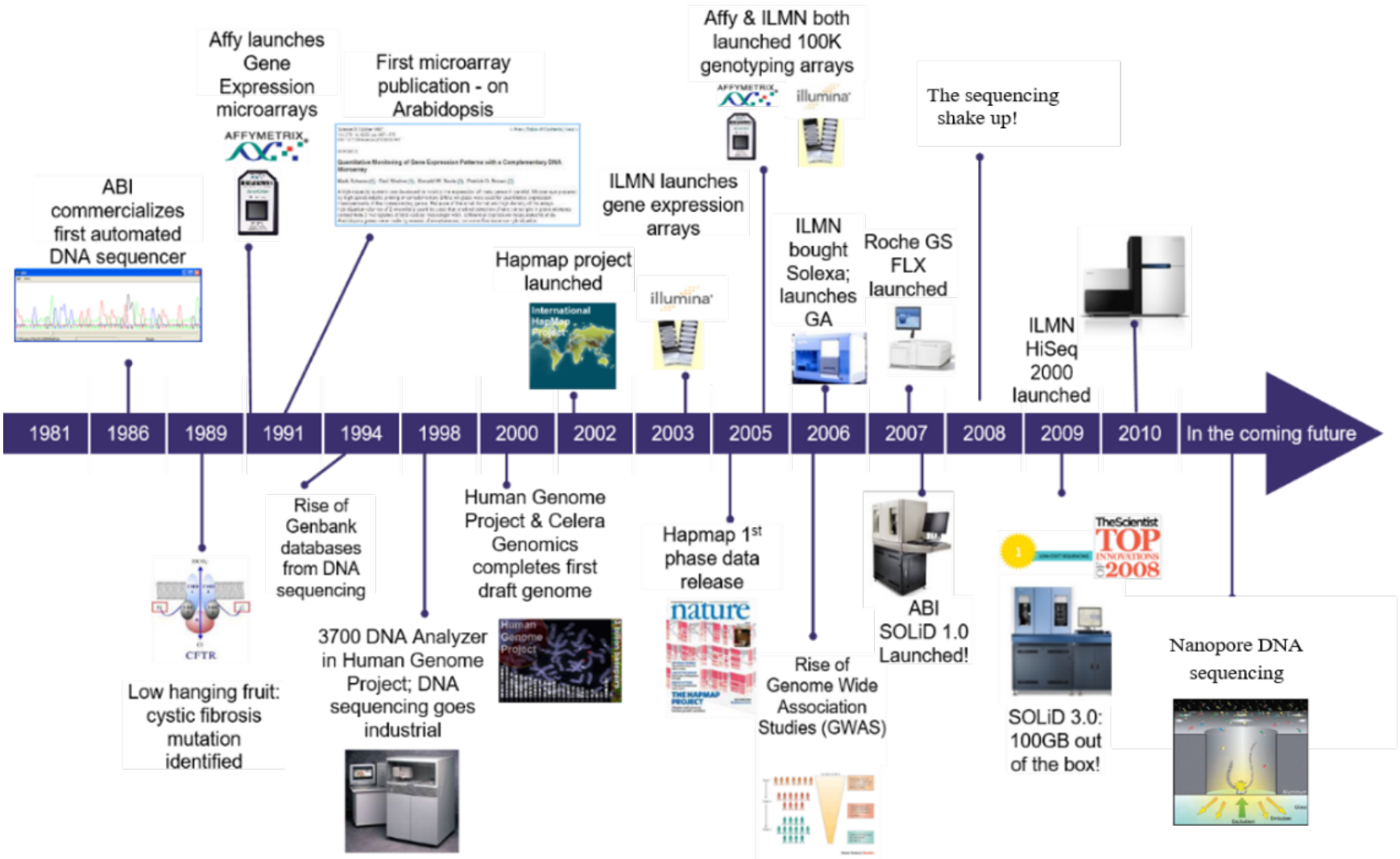
Major fields in bioinformatics

- Sequence analysis
- Gene and protein expression
- Bioimaging and structure prediction
- Network and systems biology

Let’s focus on genome research

----Especially Next generation sequencing

The milestones in genome research



A quick history of sequencing

1869 – Discovery of DNA

1909 – Chemical characterisation

1953 – Structure of DNA solved

1977 – Sanger sequencing invented

– First genome sequenced – ΦX174 (5 kb)

1986 – First automated sequencing machine

1990 – Human Genome Project started

2003 – Completion of Human Genome Project (3 Gb)

– 13 years, $2.7 bn

2005 – First “next-generation” sequencing instrument

2013– >10,000 genome sequences in NCBI database

Sanger sequencing: chain termination method

- Uses DNA polymerase
- All four nucleotides, plus one dideoxynucleotide (ddNTP)
- Random termination at specific bases
- Separate by gel electrophoresis

Sanger sequencing: chain termination method

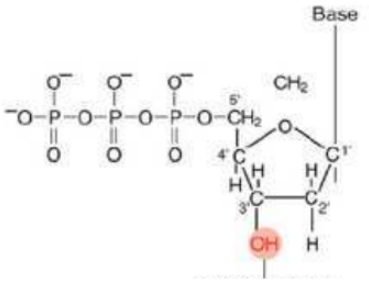
TCTGATGCAT*

TCTGATGCATGAACT*

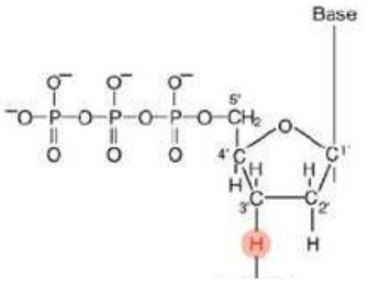
TCTGATGCATGAACTGCT*

TCTGATGCATGAACTGCTCAT*

AGACTACGTACTTGACGAGTAC.....



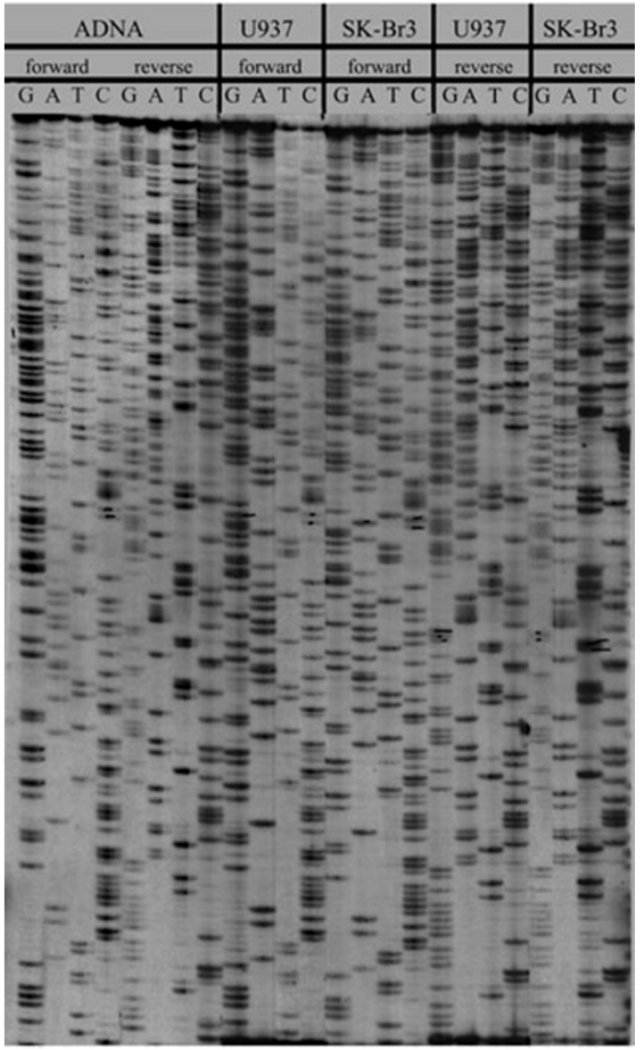
deoxynucleotide



dideoxynucleotide

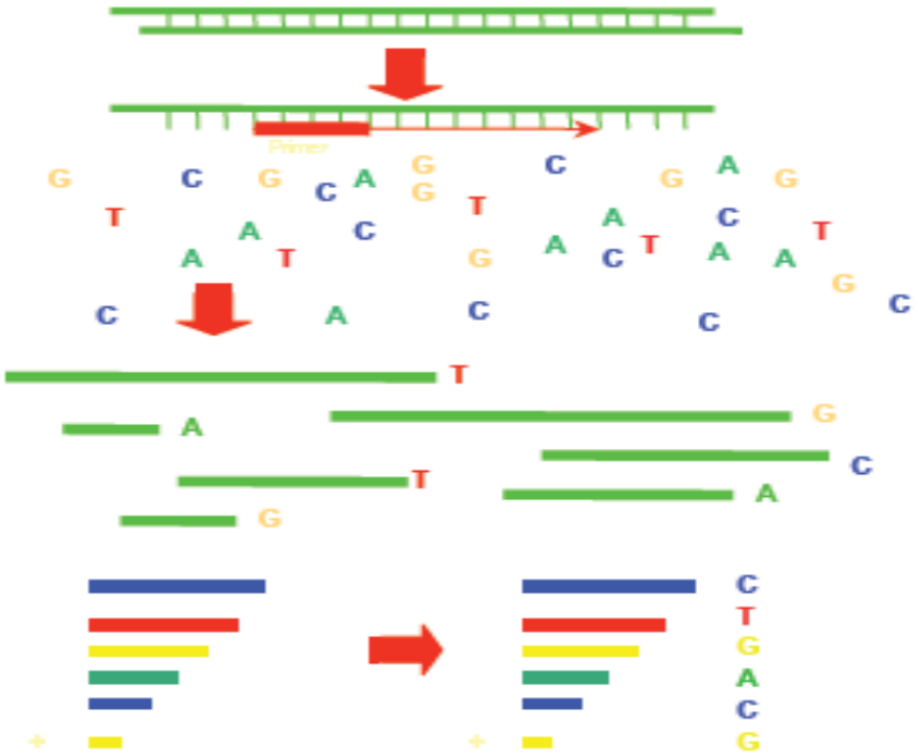
Sanger sequencing: chain termination method

Separation of fragments by gel electrophoresis

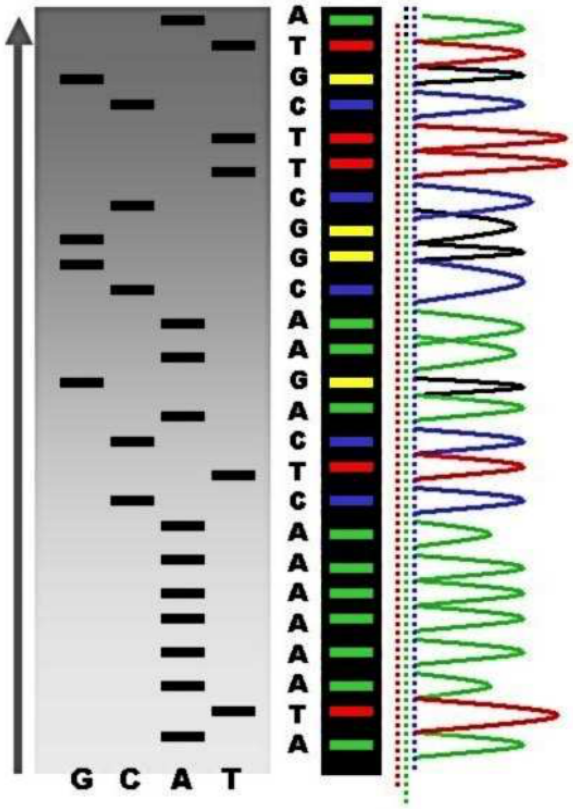


Sanger sequencing: dye-terminator sequencing

1986: 4 Reactions to 1 Lane
fluorescently labelled ddNTPs



Sequencing Reaction Products



Progression of Sequencing Reaction

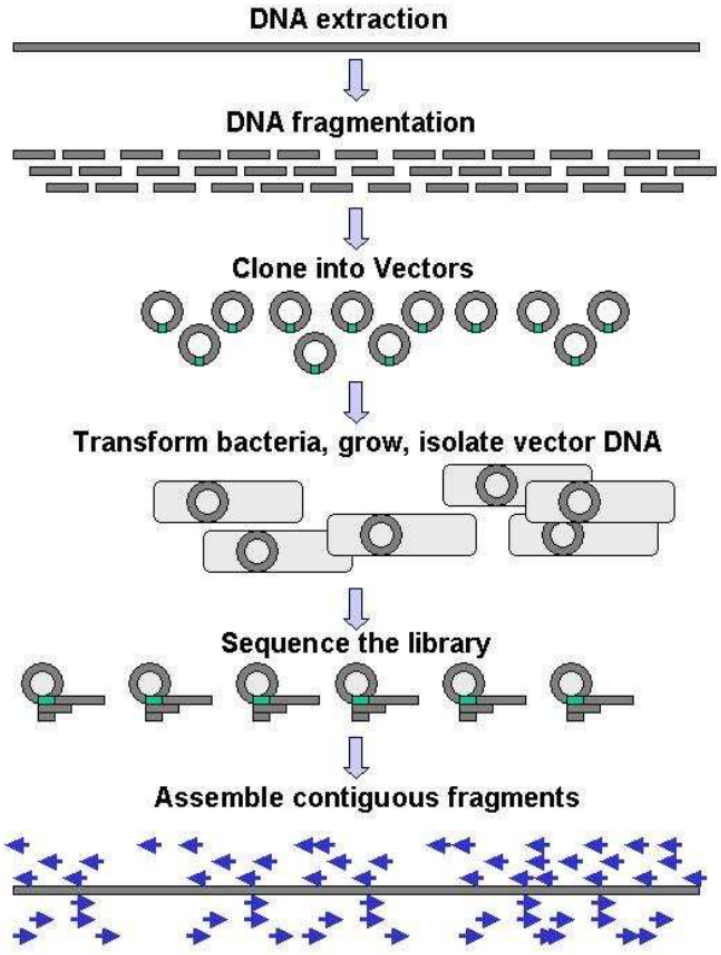
Sanger sequencing: dye-terminator sequencing

Automated DNA Sequencers

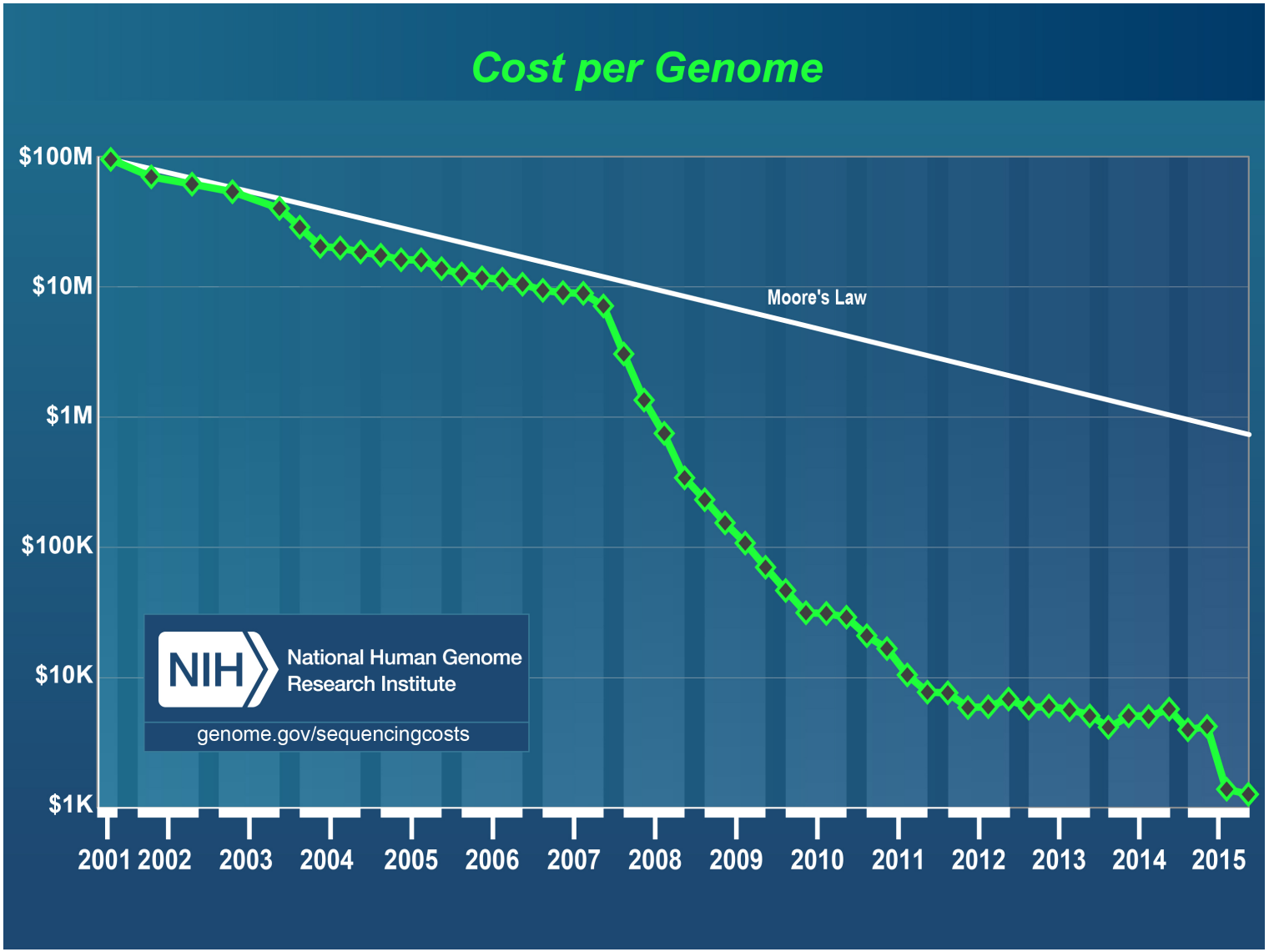
ABI 377 Plate Electrophoresis

ABI 3730 xl Capillary Electrophoresis

Sanger sequencing: shotgun library preparation



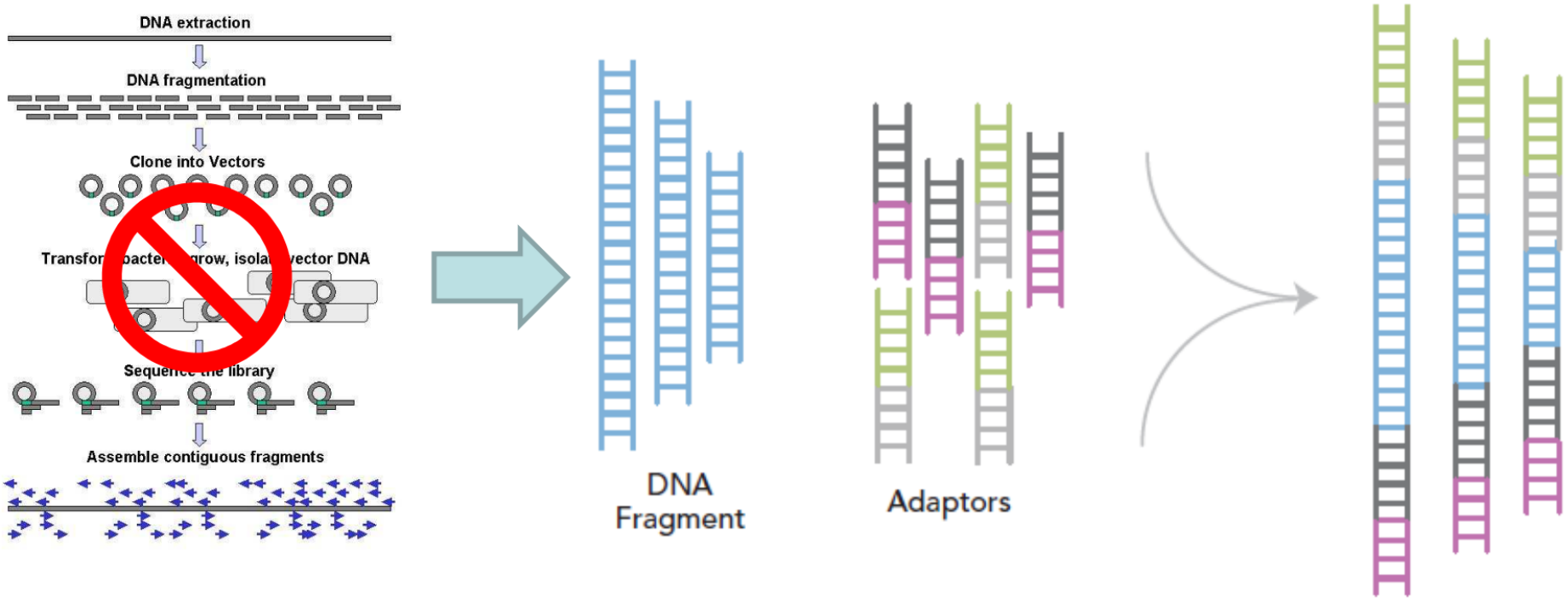
Cost of sequencing decrease quickly



Next-gen sequencing technologies

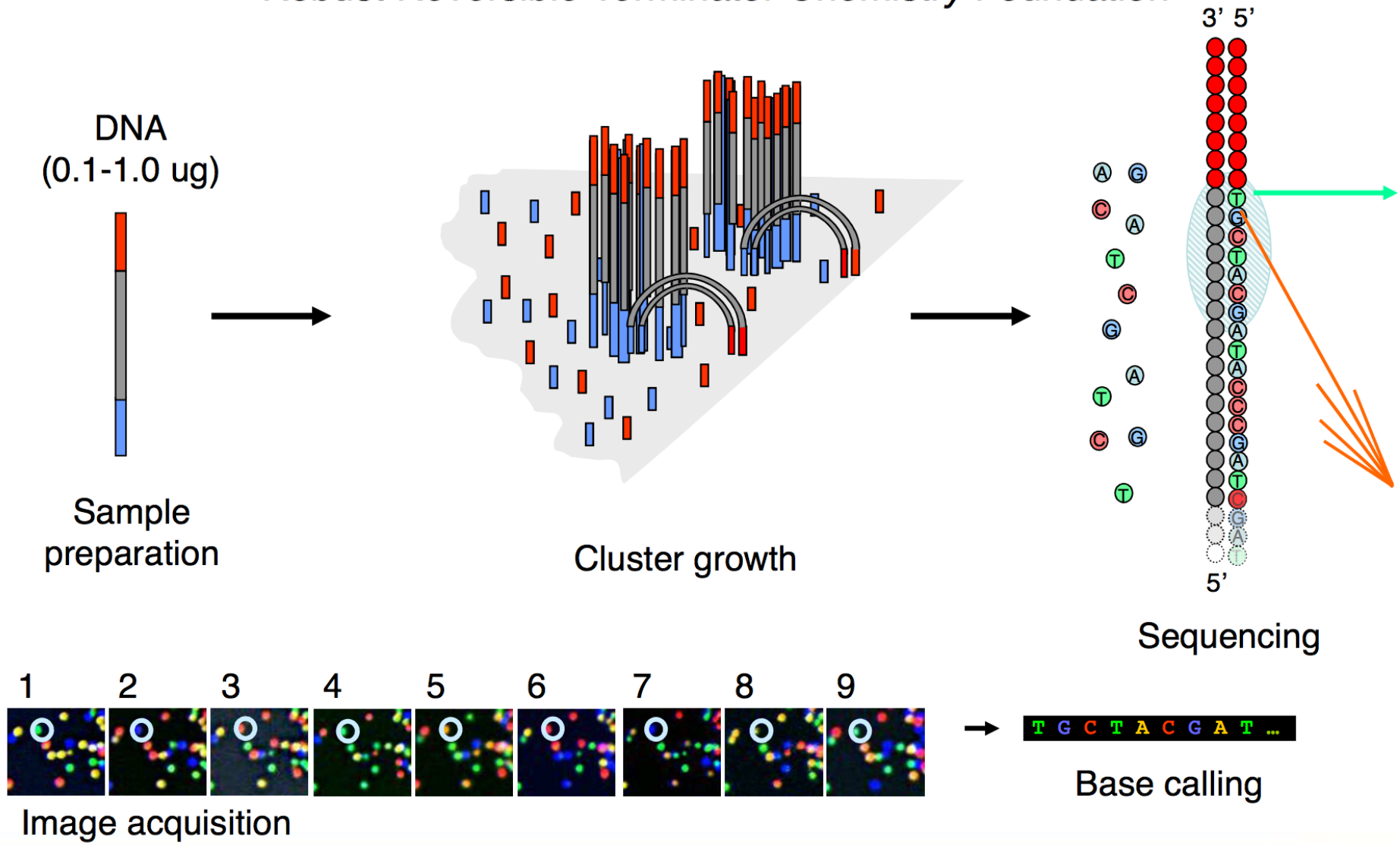
- With massively parallel sequencing new methods for sequencing template preparation is required
- Current NGS platforms utilize clonal amplification on solid supports via two main methods:
  - *emulsion PCR (emPCR)*
  - *bridge amplification (DNA cluster generation)*

Next-gen sequencing: shotgun library preparation



Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation



Illumina: Data Processing

Nucleotide Flows

↓

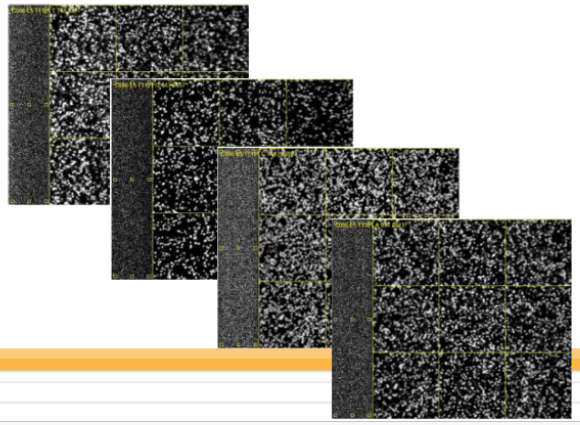
Raw Images

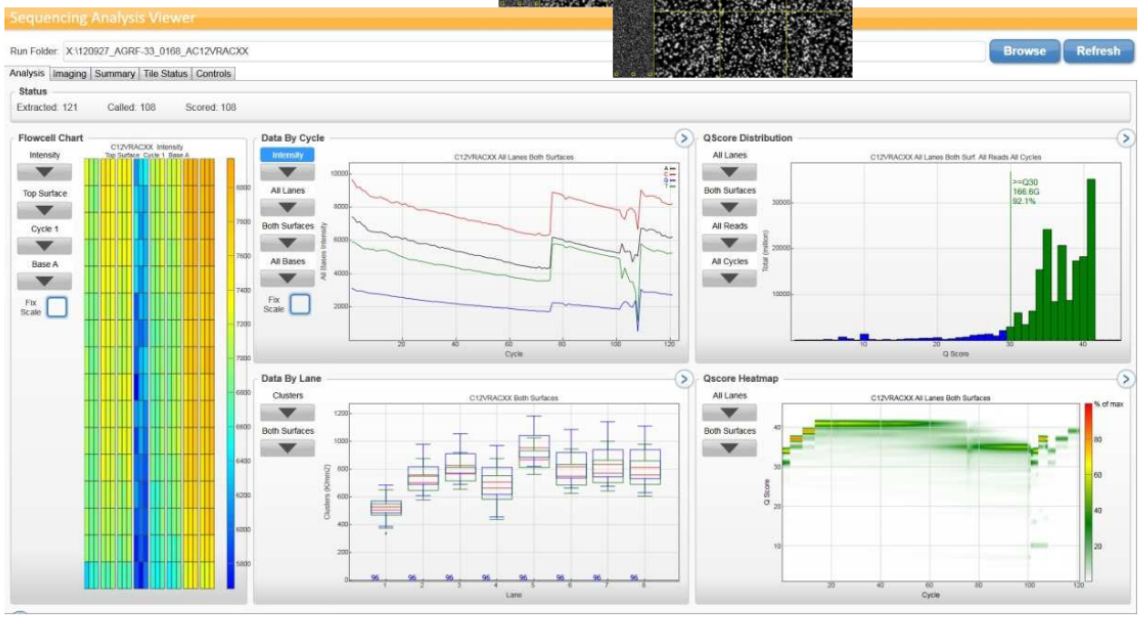
↓

Image Processing → Base-calling → Quality Filtering

↓

.bcl





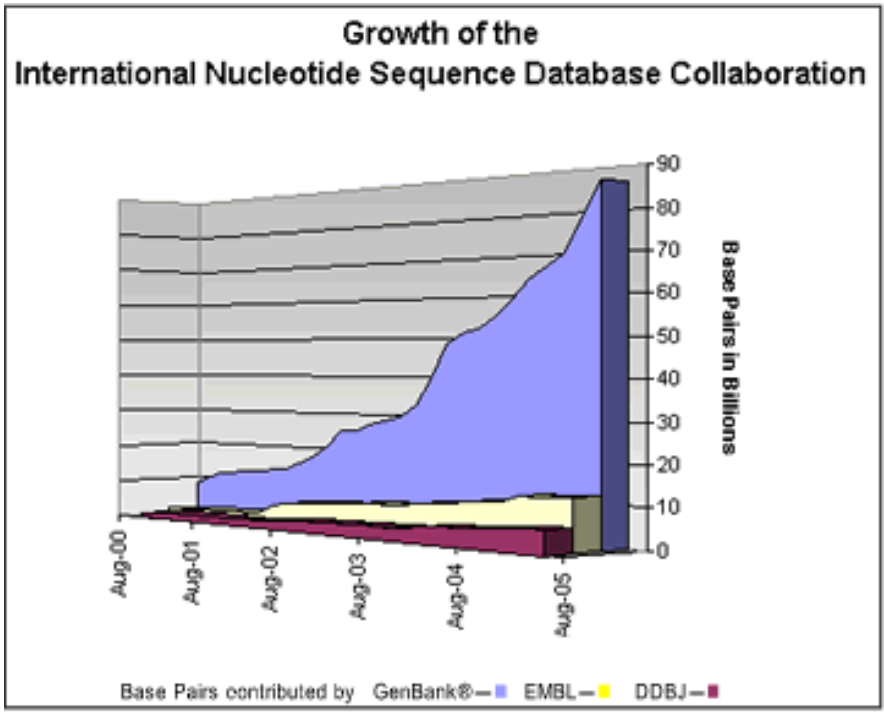
Bioinformatics Challenges

The huge dataset

100 Gigabases

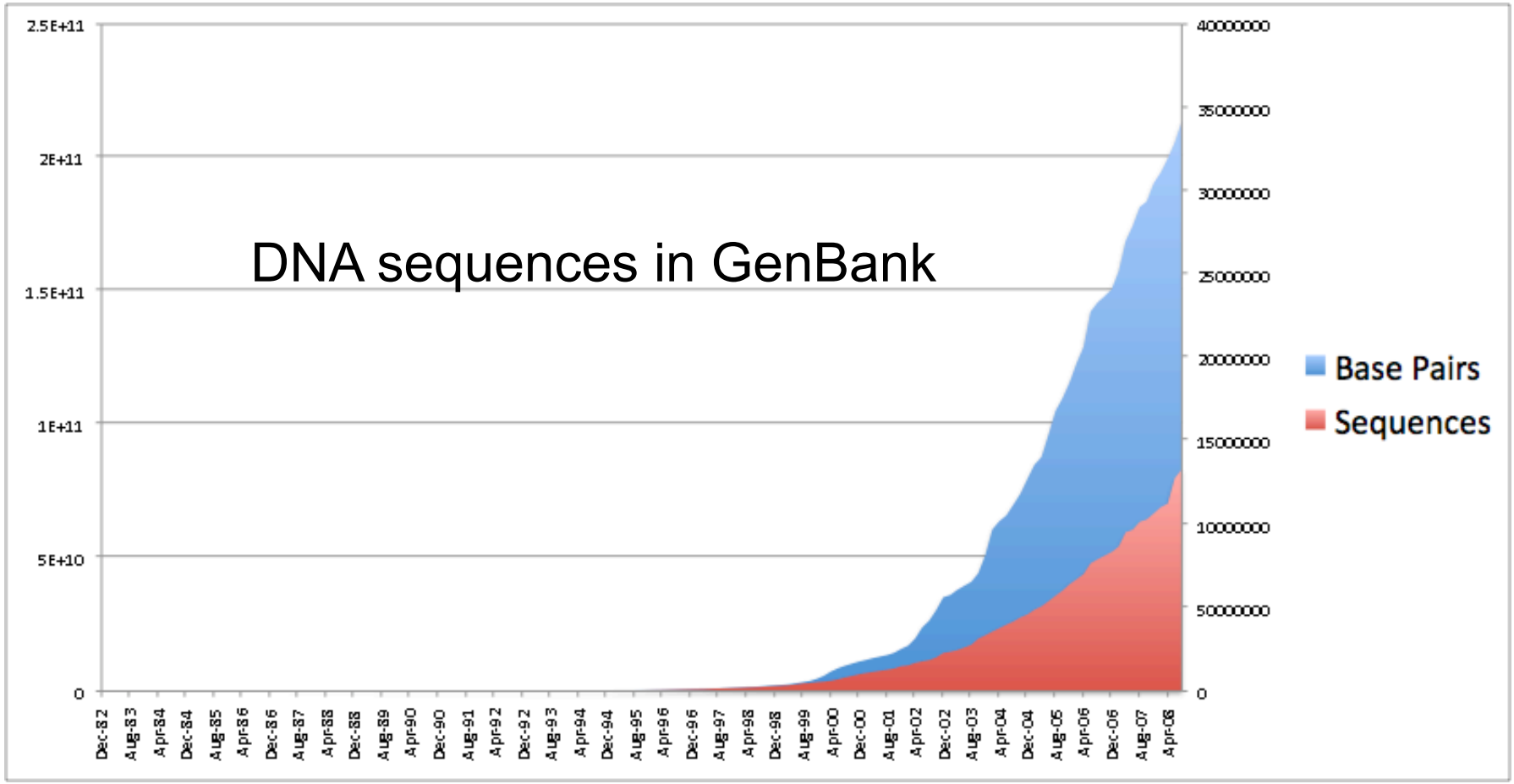
GenBank and its collaborating databases, the European Molecular Biology Laboratory and the DNA Data Bank of Japan, have reached a milestone of 100 billion bases from over 165,000 organisms. See the press release or find more information on GenBank.

- Lots of new sequences being added
  - automated sequencers
  - genome sequencing
  - EST sequencing
  - environmental/metagenomic sequencing



- GenBank has over 100 **Billion** bases and is doubling every year!!
  - problem of exponential growth
  - how can computers keep up?
  - hard drives are cheaper, but processor speeds are not keeping up

DNA Sequencing capability has grown exponentially



Doubling time = 18 months

Big Data

Thank you for your attention!