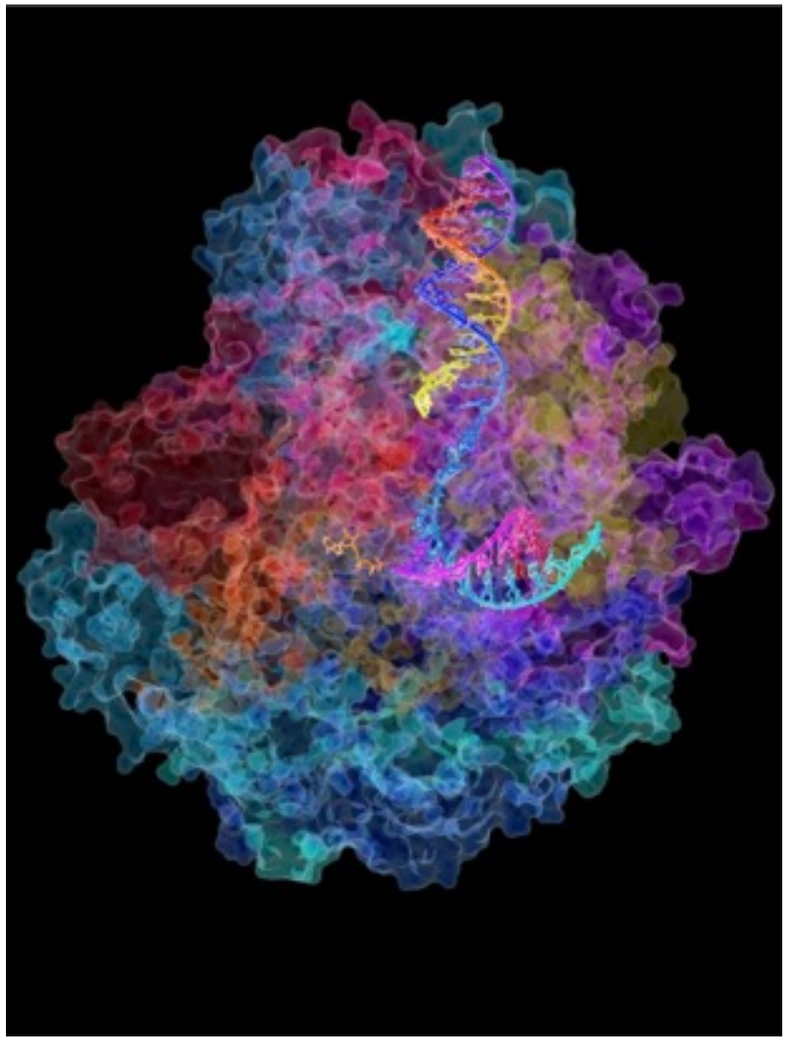
# Overview of Bioinformatics

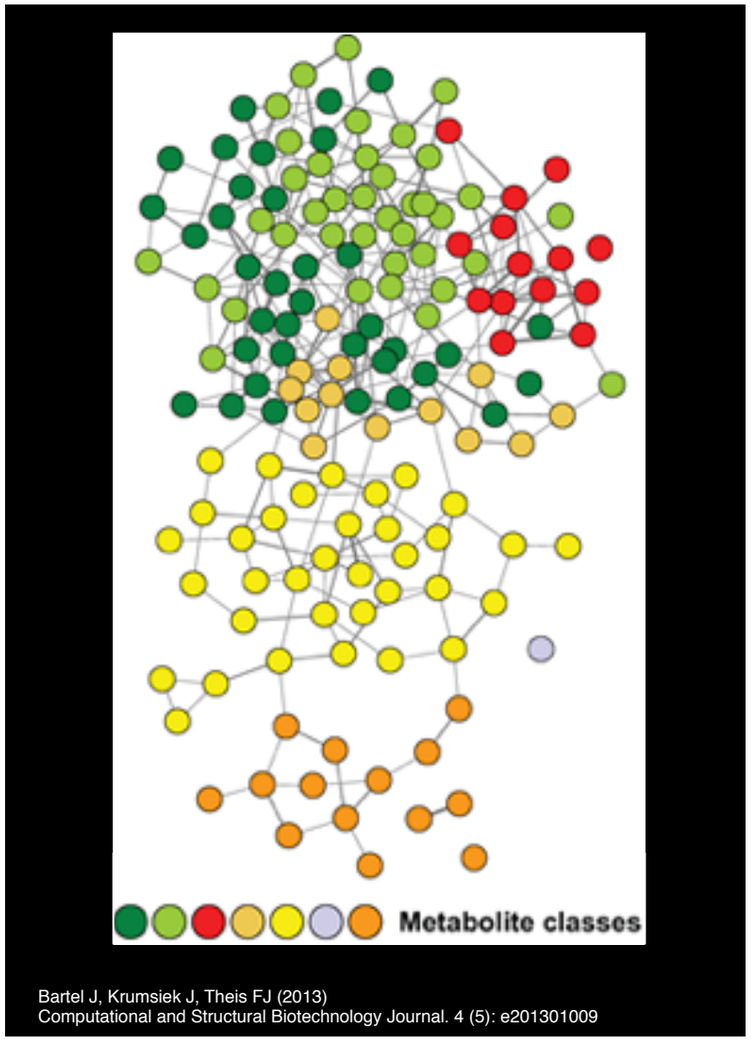
Harvard Chan Bioinformatics Core

NGS Data Analysis Course 2016

# What is Bioinformatics?

**The use of computer science, mathematics, and information theory to organize and analyze complex biological data.**






Bartel J, Krumsiek J, Theis FJ (2013)
Computational and Structural Biotechnology Journal. 4 (5): e201301009

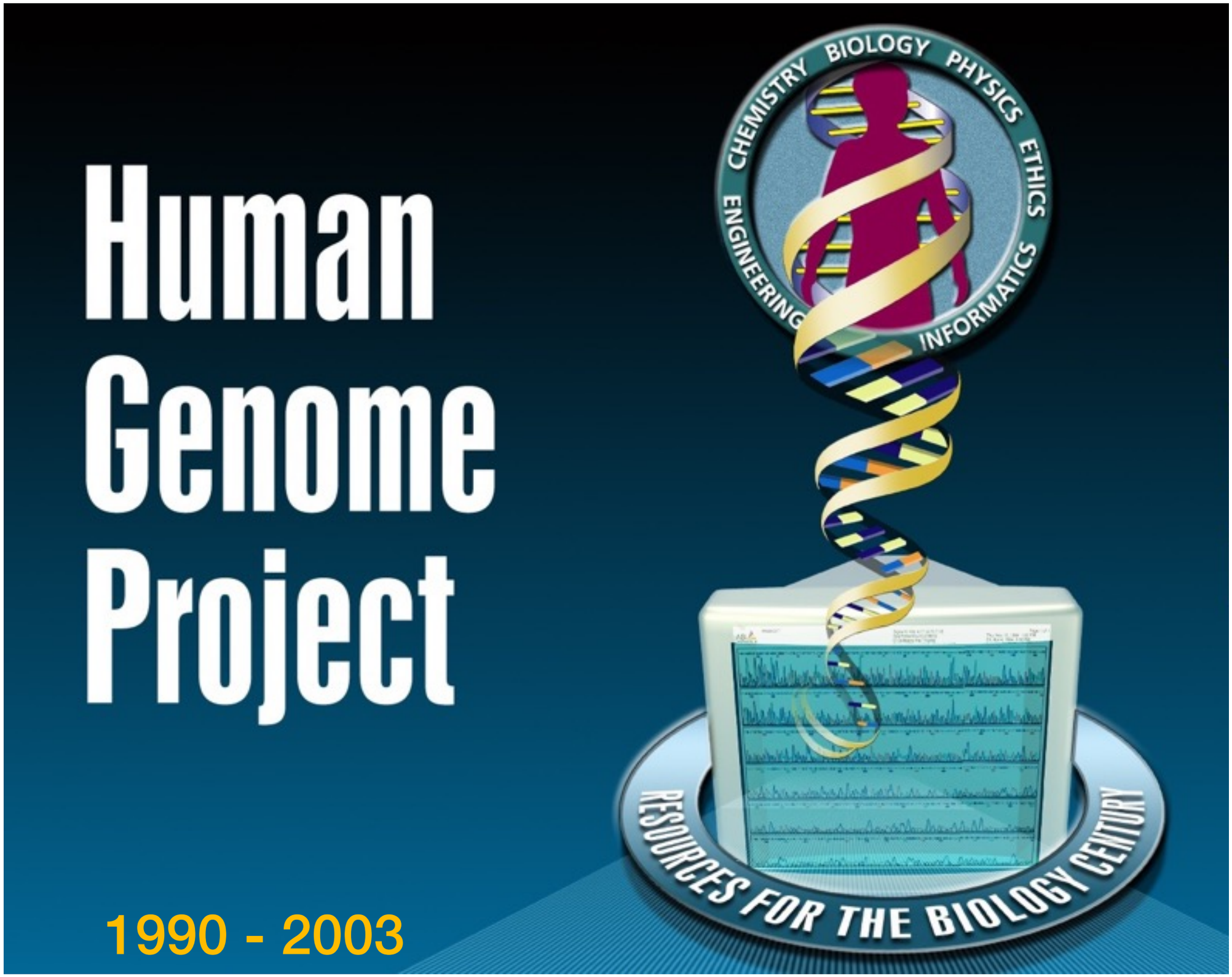
# Bioinformatics in the Omics Era

# Why Genomics?


**High Throughput**

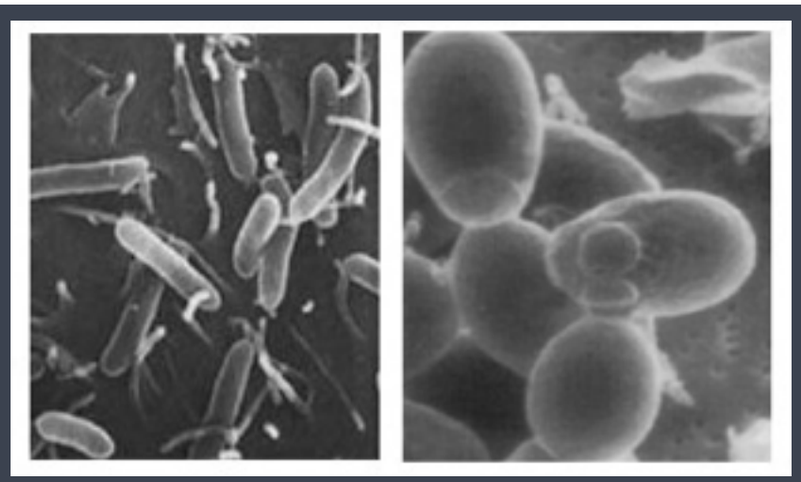
**Comprehensive**

**Exploratory**




http://blogs.swa-jkt.com/swa/10318/2014/03/18/biology-human-genome-project-prompt/

# Sequenced organisms (2000)



**38 bacteria**

***S. cerevisiae***

***C. elegans***

***D. melanogaster***

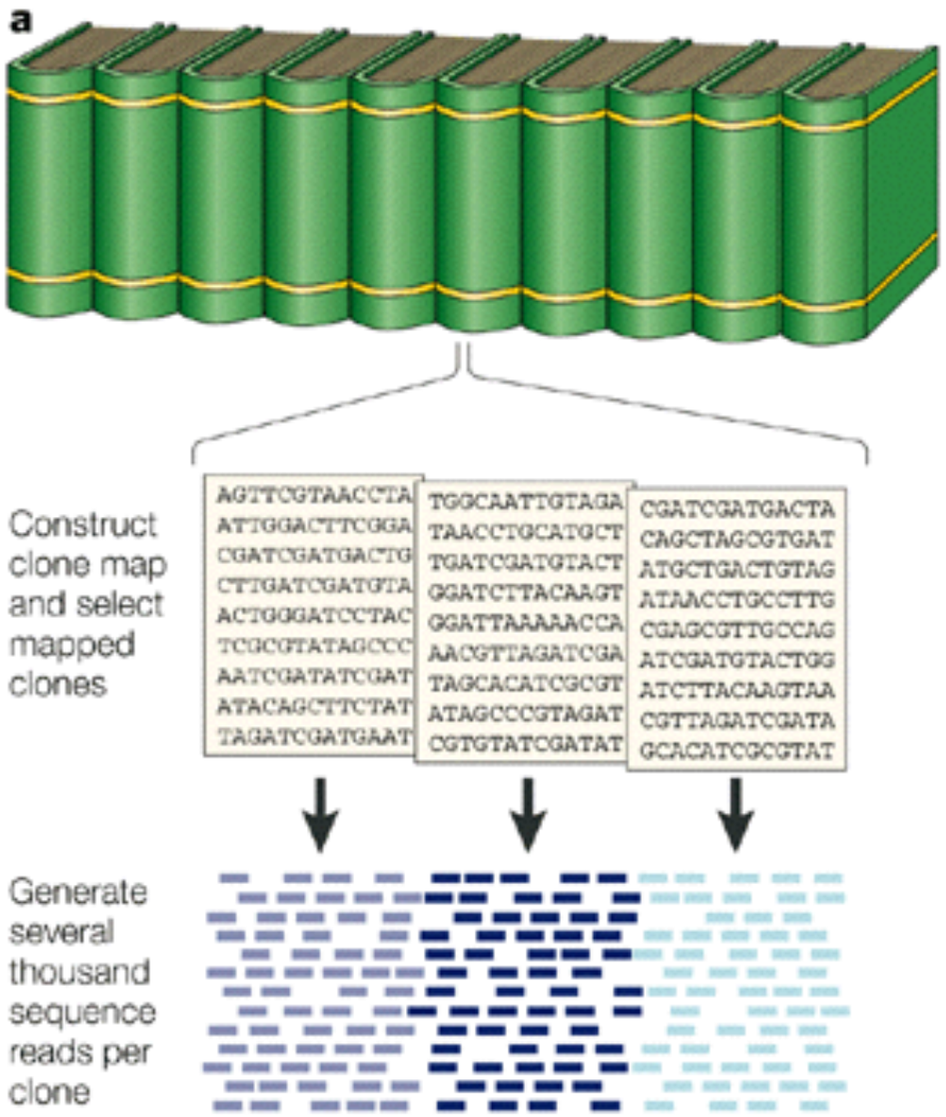
***A. thaliana***

image credit: http://www.genomesize.com/rgregory/genome.html

# Sequencing of the Human Genome

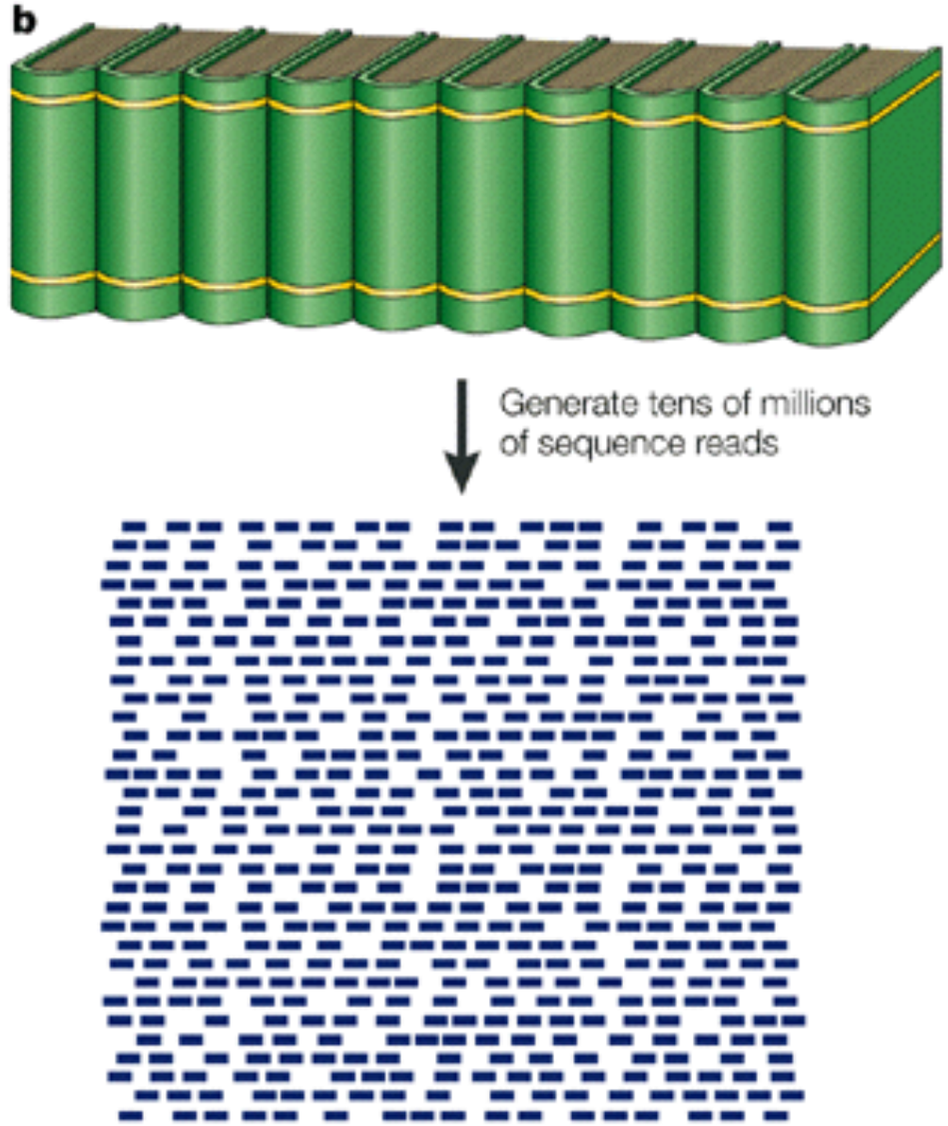
## Human Genome Project

### Heirarchical Shotgun Sequencing



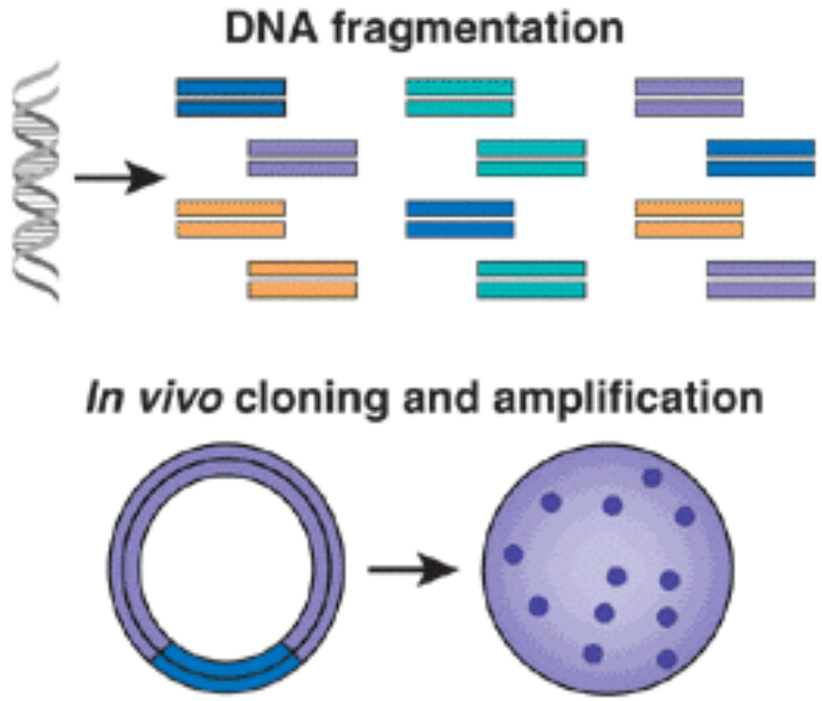

## Celera Genomics

### Whole Genome Shotgun Sequencing



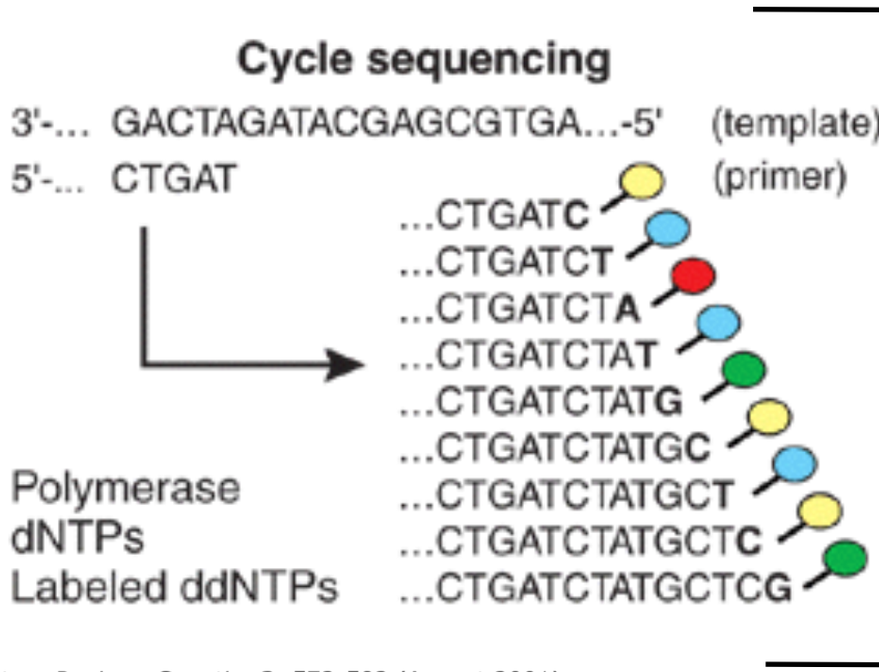

Nature Reviews Genetics 2, 573-583 (August 2001)

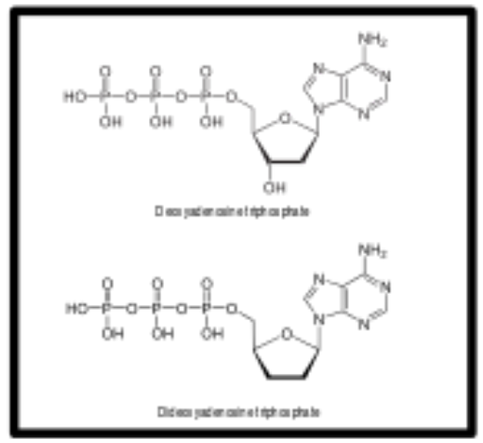
# Sequencing of the Human Genome



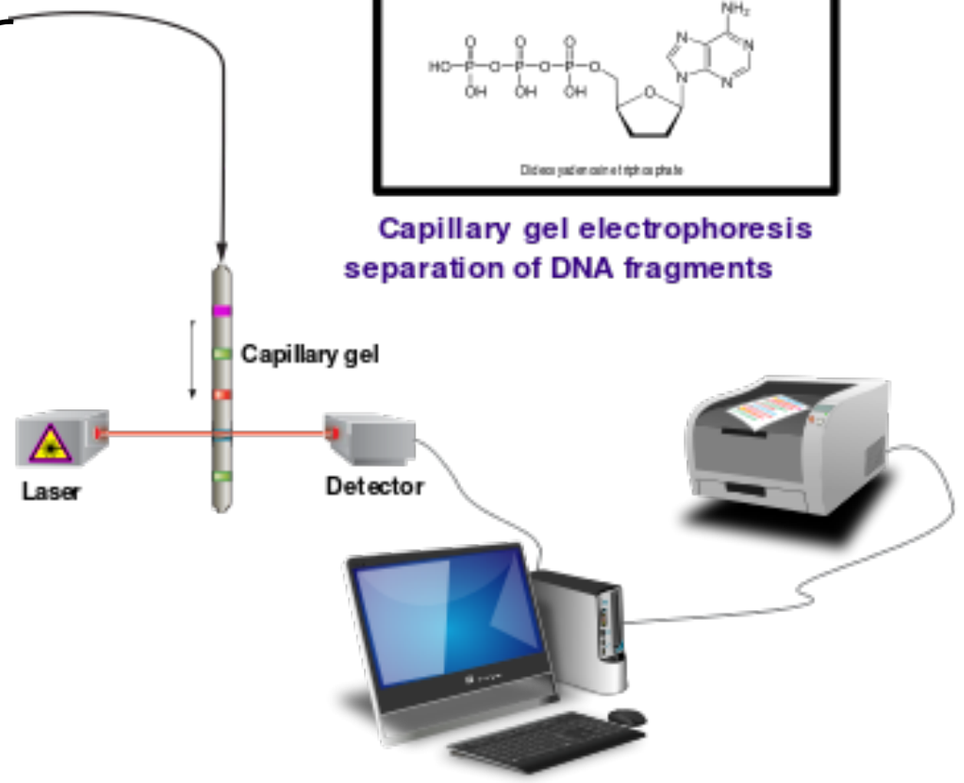


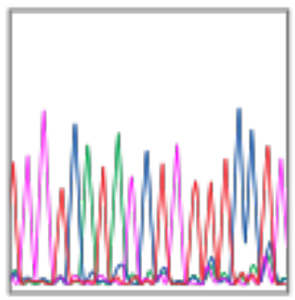




Capillary gel electrophoresis separation of DNA fragments

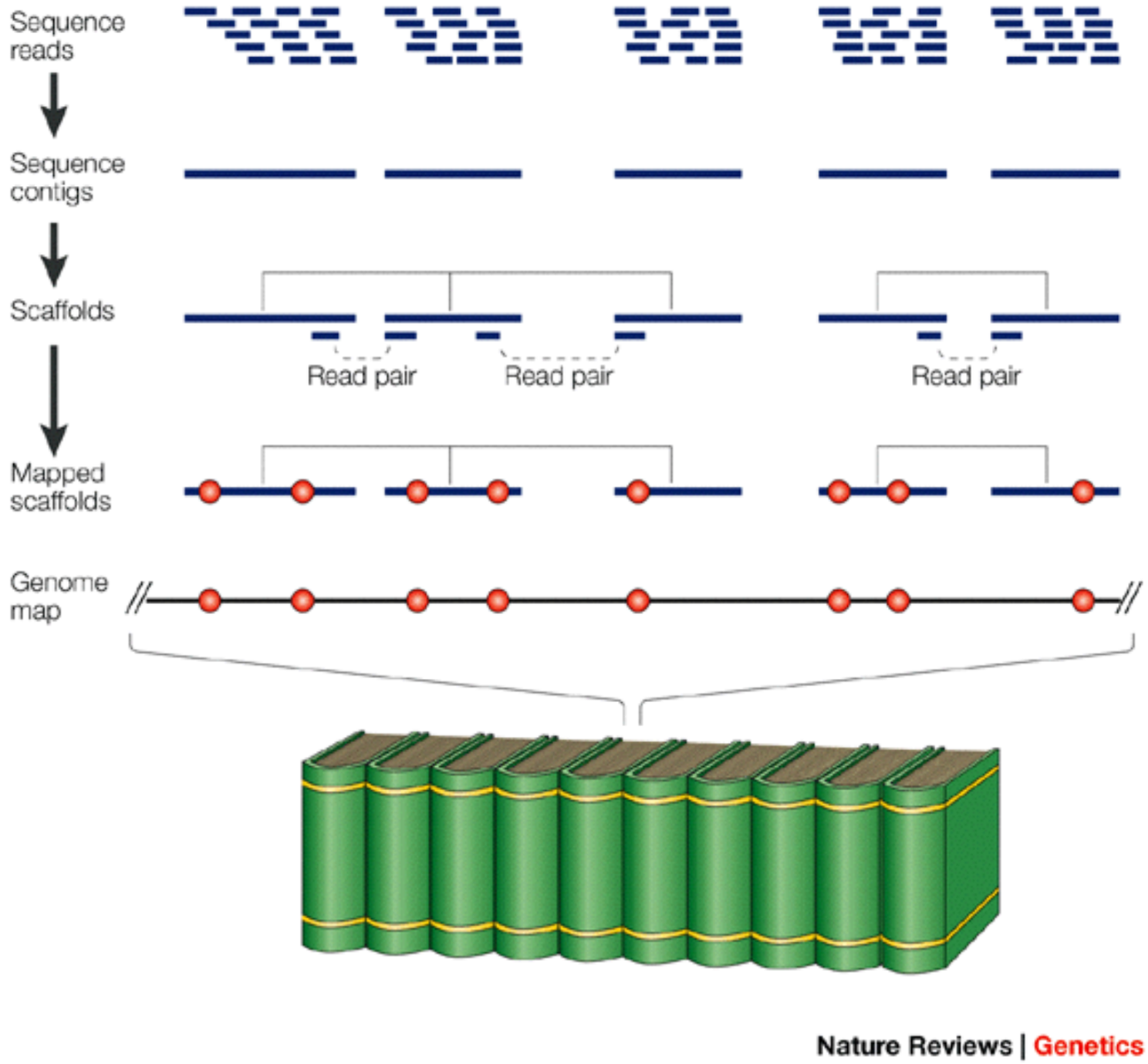



Laser detection of flourochromes and computational sequence analysis

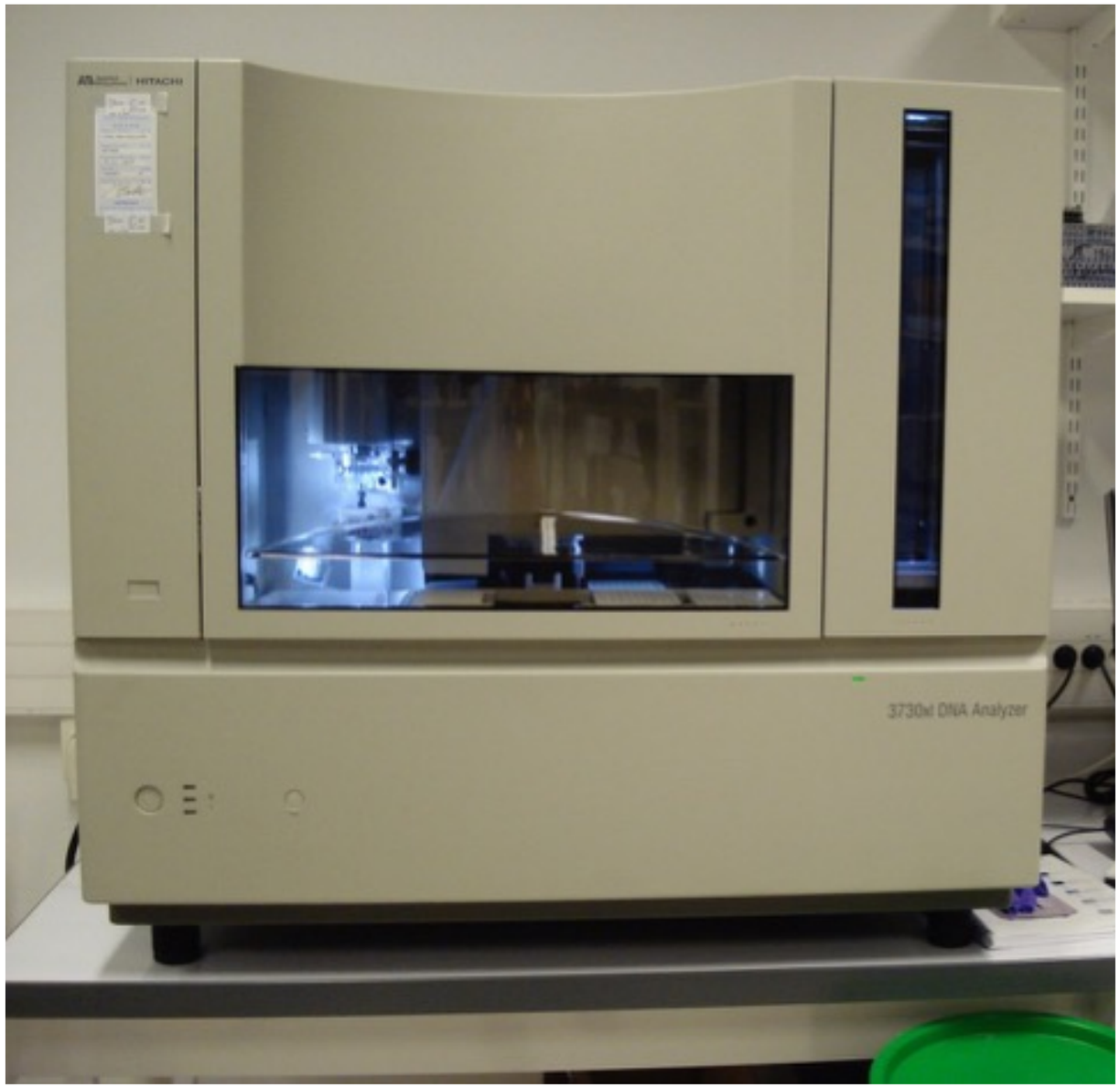
*Nature Reviews Genetics* **2**, 573-583 (August 2001)

# Sequencing of the Human Genome




Nature Reviews Genetics 2, 573-583 (August 2001)

# Sequencing of the Human Genome

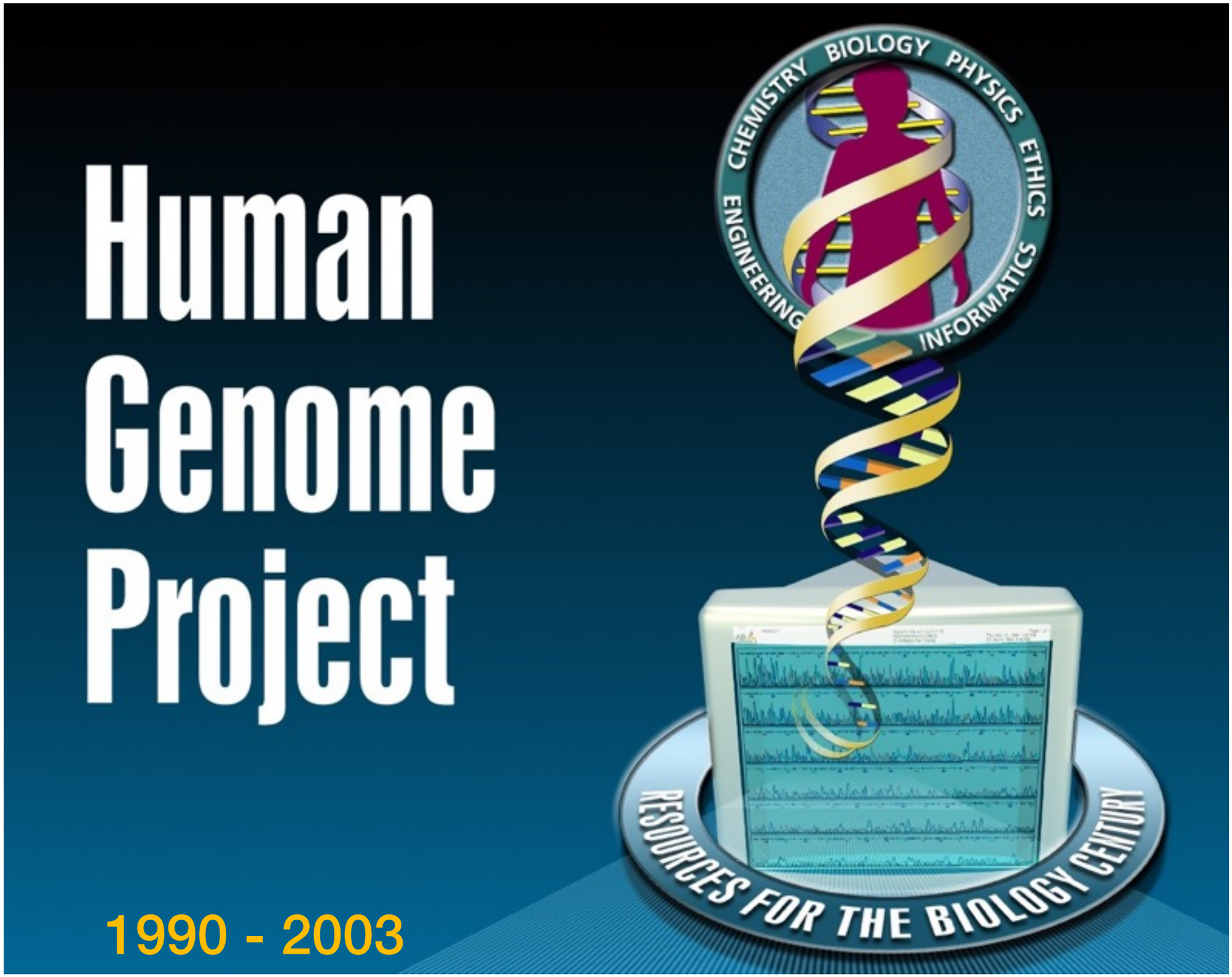



http://www.pasteur.fr/ip/portal/action/WebdriveActionEvent/oid/01s-00001u-01p

**Capillary (cycle) sequencing generated:**

- 500-700 bases per reaction (96)
- 115,000 bp / day

**Sequence production was rate limiting, not analysis**

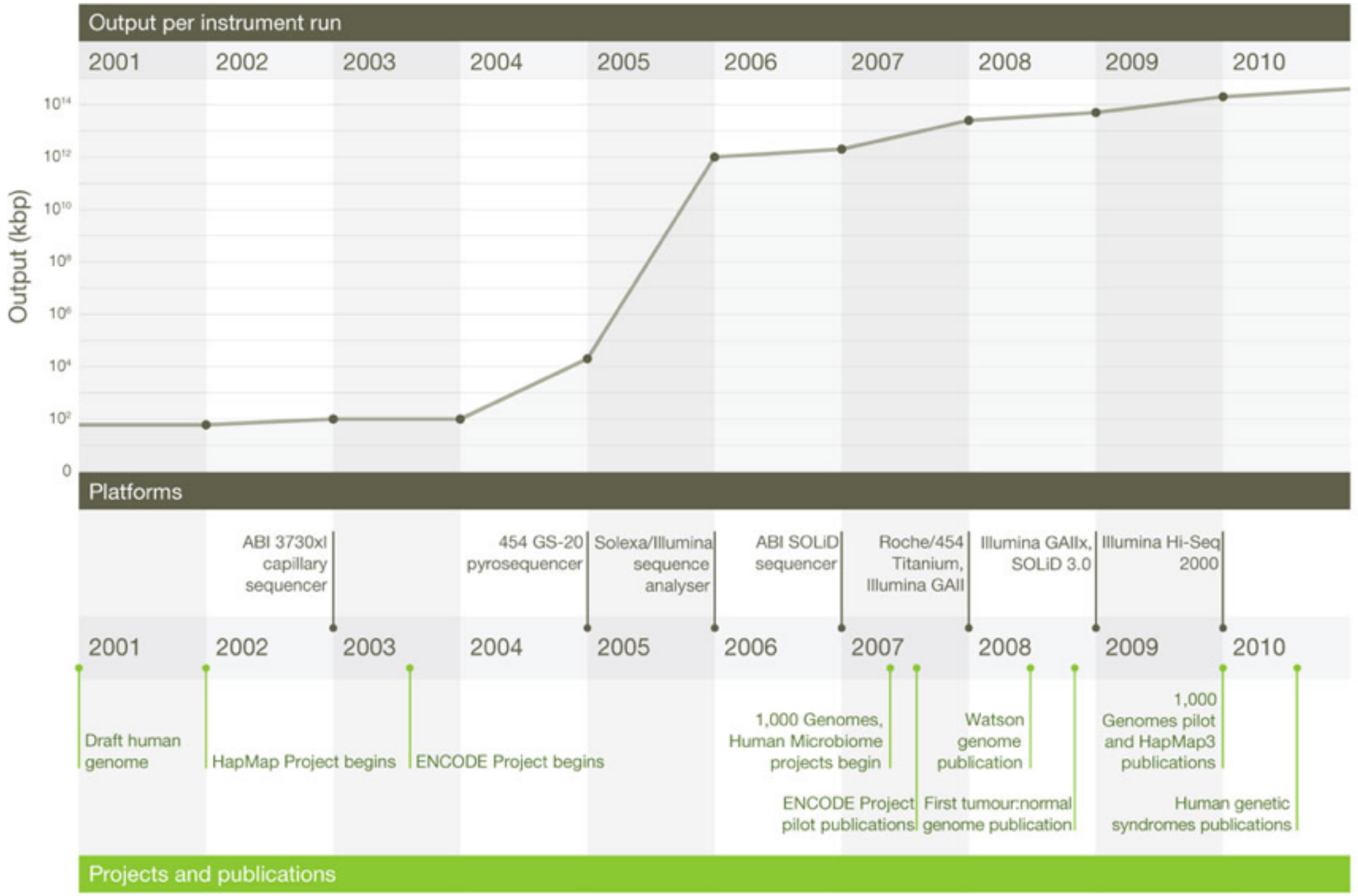



http://blogs.swa-jkt.com/swa/10318/2014/03/18/biology-human-genome-project-prompt/

# Comparative Genomics Research

Highly conserved regions of DNA likely to be functional

# Advancements in Sequencing Technology




*Nature* **470**, 198–203 (10 February 2011)

# NGS Technologies


Ion torrent

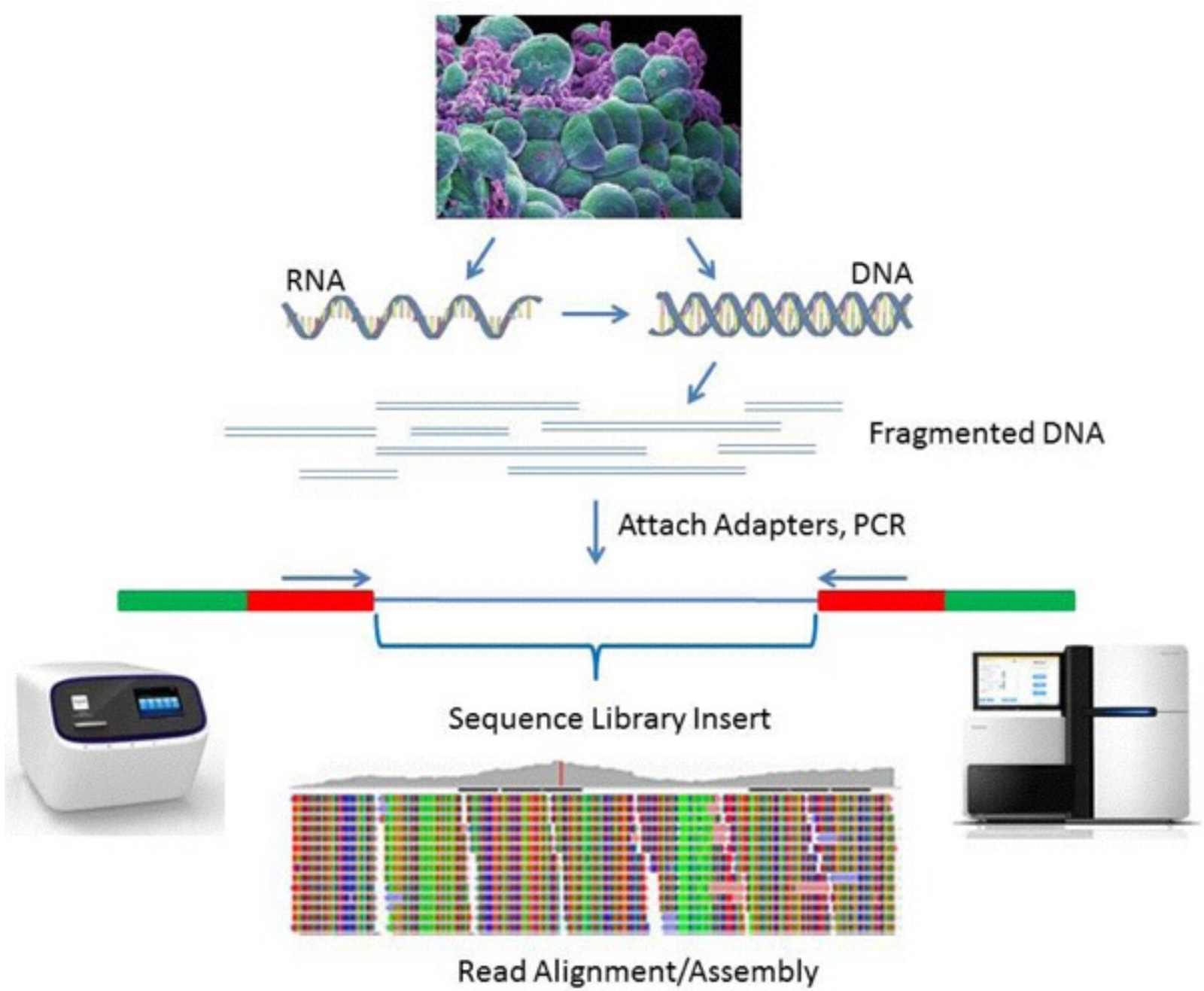
Ion Proton


MiSeq

PacBio RS

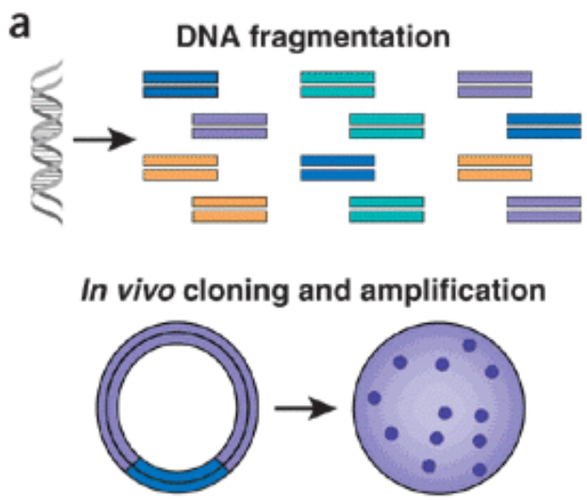
| Platform | Chemistry | Read Length | Run Time | Gb/Run | Advantage | Disadvantage |
|---|---|---|---|---|---|---|
| 454 GS FLX+ (Roche) | Pyro-sequencing | 700 | 23 hrs. | 0.7 | Long Read Length | High error rate in homopolymer |
| HiSeq (Illumina) | Reversible Terminator | 2*100 | 2 days (rapid mode) | 120 (rapid mode) | High-throughput / cost | Short reads Long run time (normal mode) |
| SOLiD (Life) | Ligation | 85 | 8 days | 150 | Low Error Rate | Short reads Long run time |
| Ion Proton (Life) | Proton Detection | 200 | 2 hrs. | 100 | Short Run times | New* |
| PacBio RS | Real-time Sequencing | 3000 (up to 15,000) | 20 min | 3 | No PCR Longest Read Length | High Error Rate |

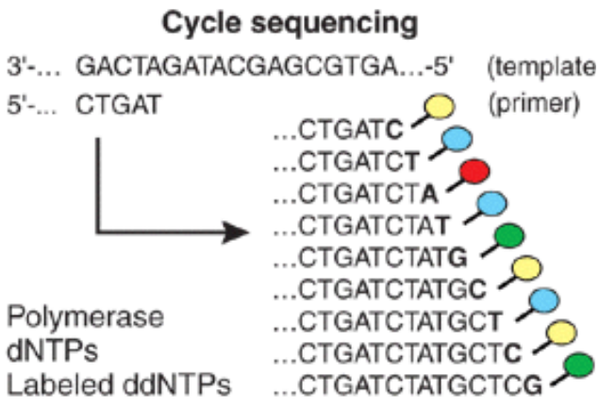
# NGS Technologies

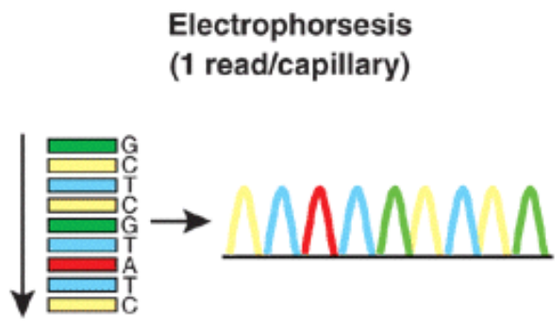


# NGS Technologies

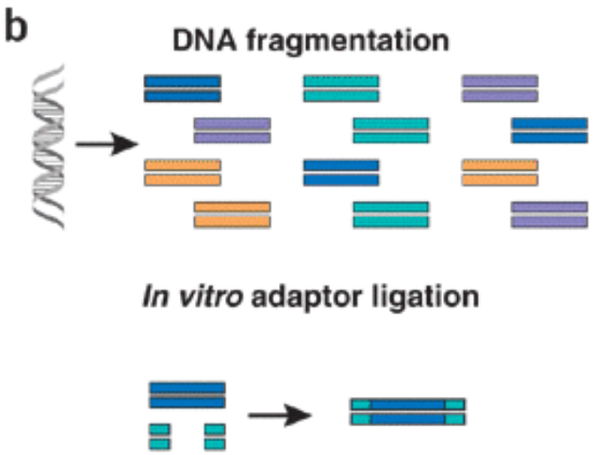
## Cycle Sequencing

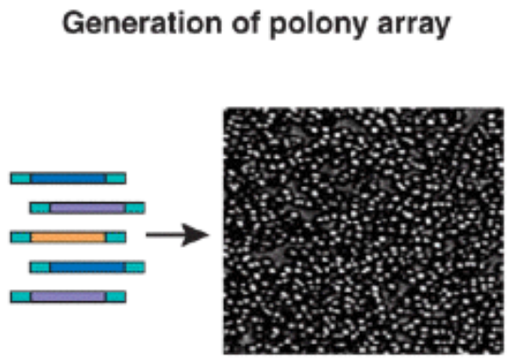


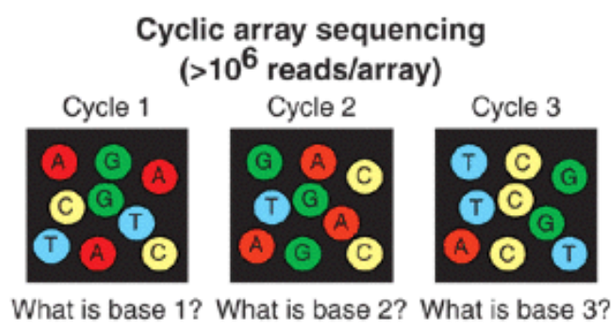




## NGS Sequencing

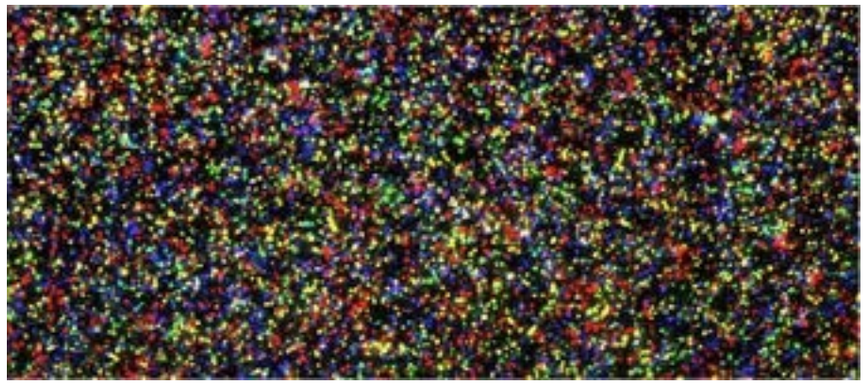



*Nature Reviews Genetics* **2**, 573-583 (August 2001)

# NGS Technologies



**Generate**:

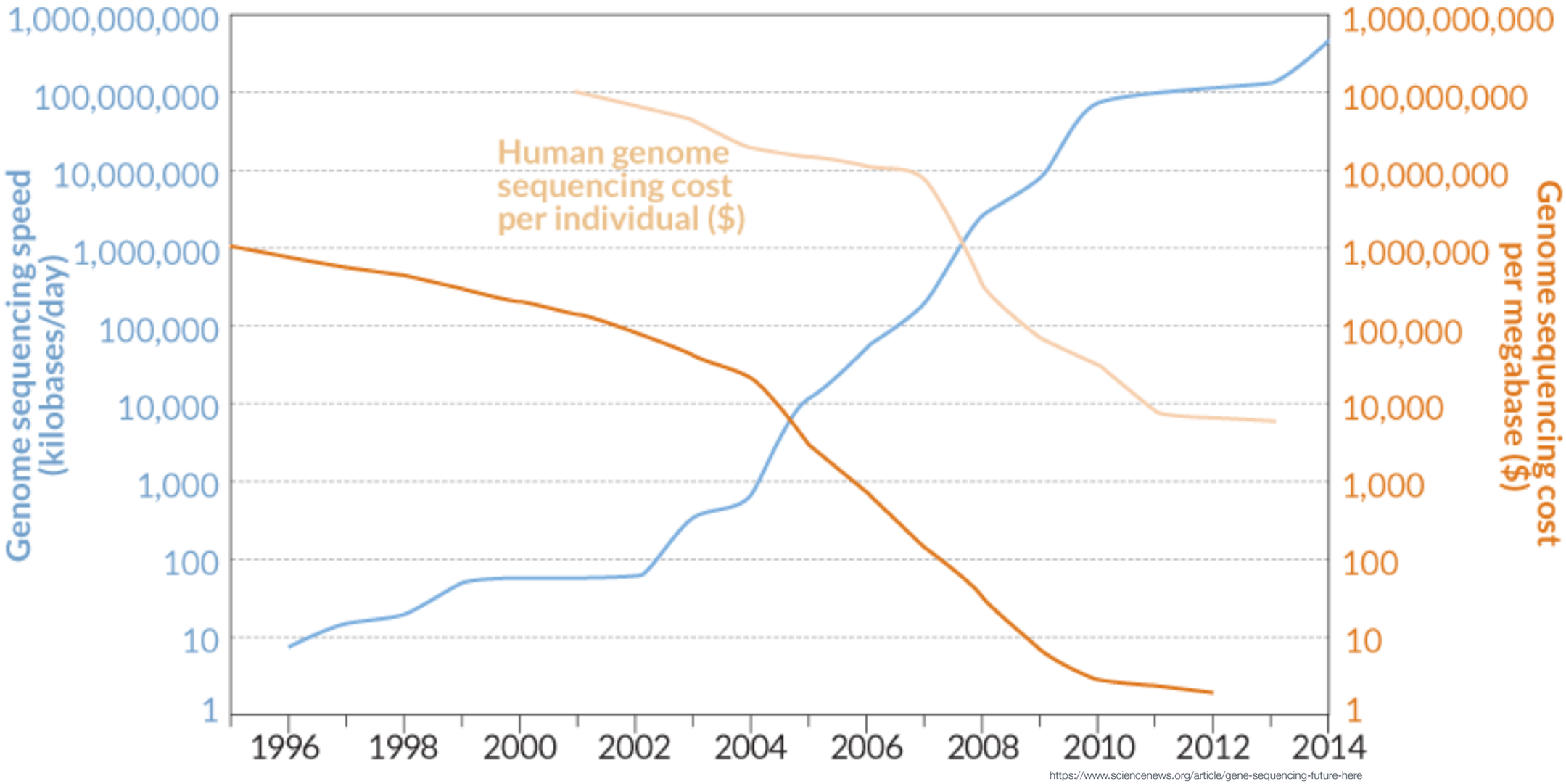
- Illumina HiSeq2500: 125 bp / read (250 bp / read Rapid-Run Mode)
- > **100 billion bp** / day

**Bioinformatics support required to handle:**

- Massive amount of data
- Shorter read lengths
- NGS technology-specific error profiles

**Sequence analysis rate limiting, not production**

# NGS Technology Accessibility




**Increased scale and lower cost increases access to sequencing technologies**

# The Genomic Era: Collaborative Projects



ENCODE

FANTOM
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

i5K

earth
microbiome project

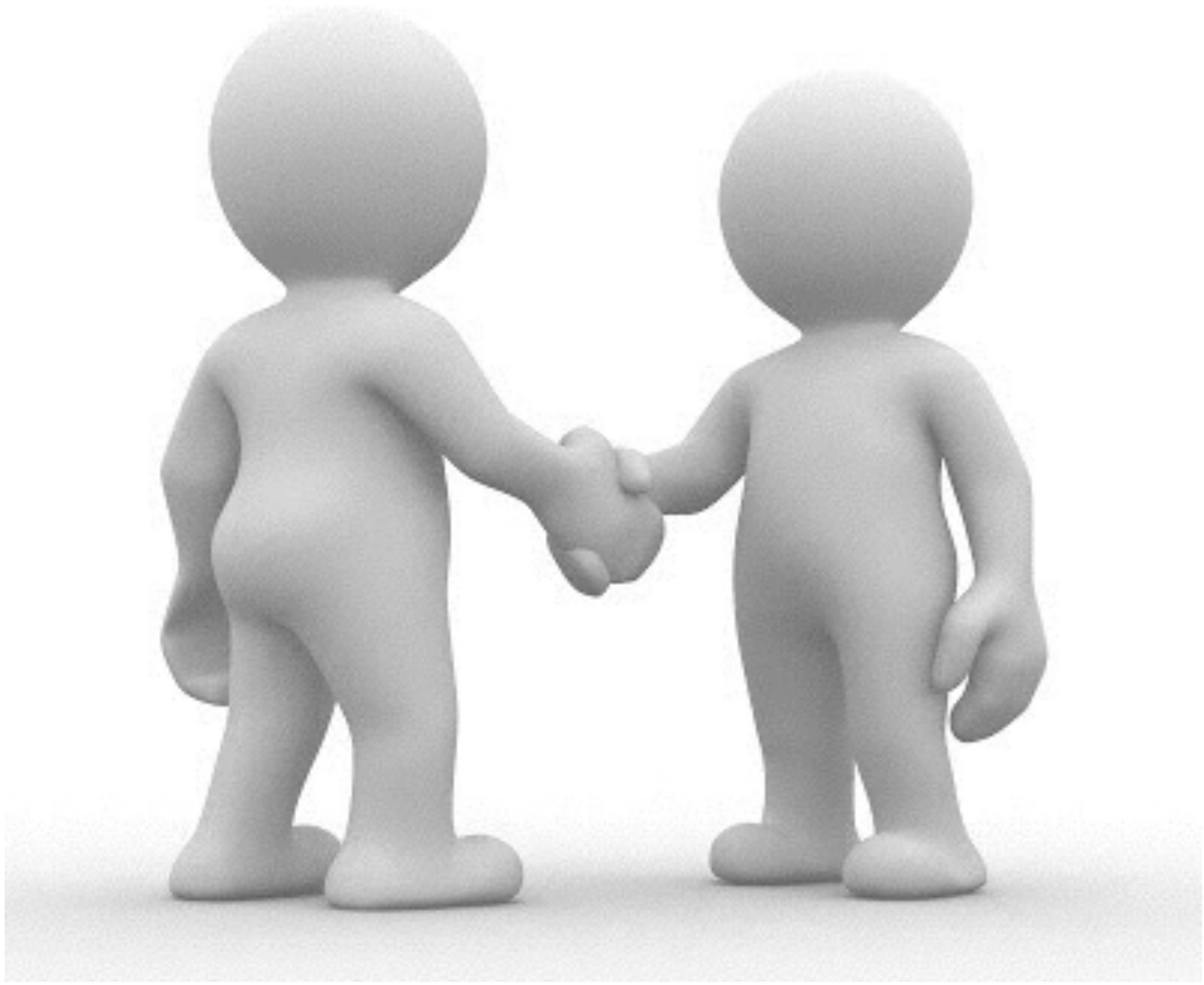
HMP

100K GENOME PROJECT

# The Genomic Era: Individual Projects


- Only a few experiments = vast amounts of data
- Data generation straightforward, but analysis requires bioinformatics expertise

# Bioinformatics in the Genomic Era



Alliances between experimentalists and computational biologists

# Bioinformatics Toolkit

# Programming languages

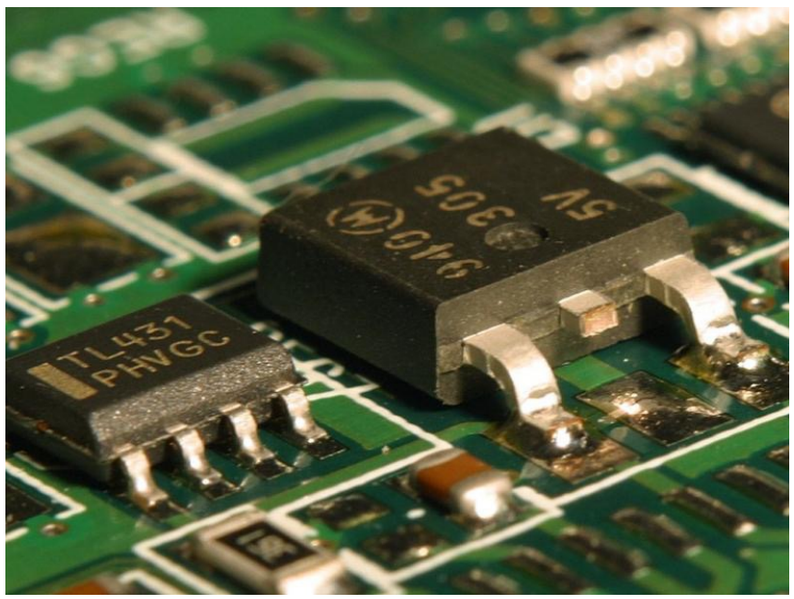
**Programming languages are critical in genomic analyses:**

- **Bash:** command line language used for interacting with *Linux / Unix-based* operating systems.
  - Attaining sequencing data
  - Accessing computing resources and analysis tools
  - Basic data manipulation, creating scripts for running tools

- **R:** a programming language and environment for statistical computing and graphics
  - Manipulating data, performing statistical analyses
  - Creating figures and plots
  - Sharing analyses and data

# Computing Resources

**Large genomic datasets require extensive computational resources:**




## Storage

- Large datasets: a single raw sequence file can be 5GB to 150GB
- All sample files + intermediate files for every project can easily exceed 500GB to 1TB

## Memory/RAM (Random Access Memory)

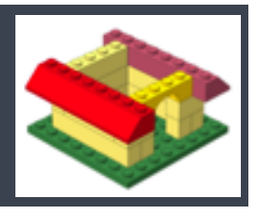
- Large datasets ≈ lots of RAM to perform analysis

## CPU (Central Processing Unit)

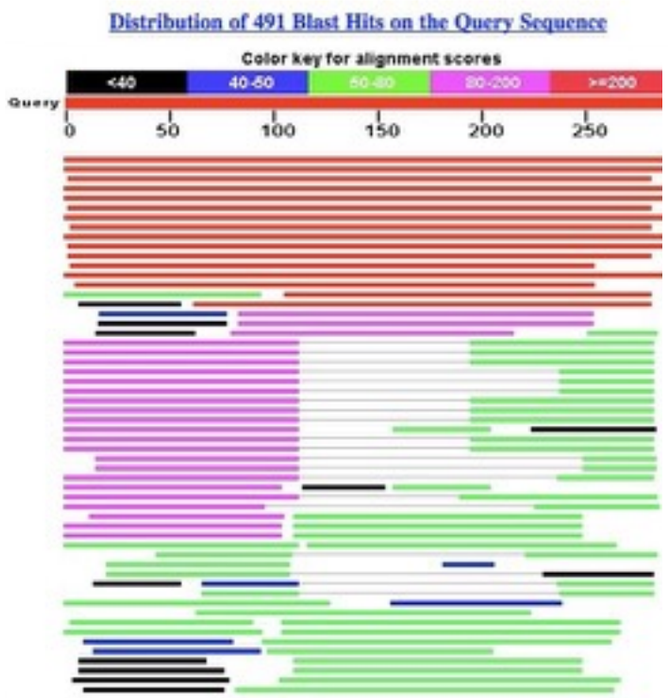
- Large datasets ≈ lots of time to perform analysis

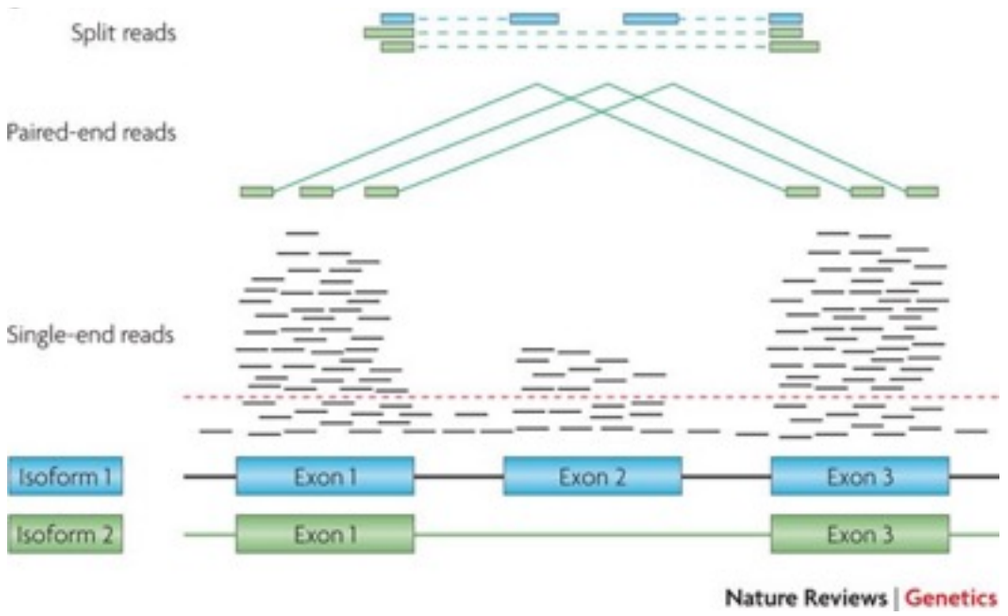
**Solution: Amazon Cloud or computing cluster**

# NGS Analysis Tools and Workflows

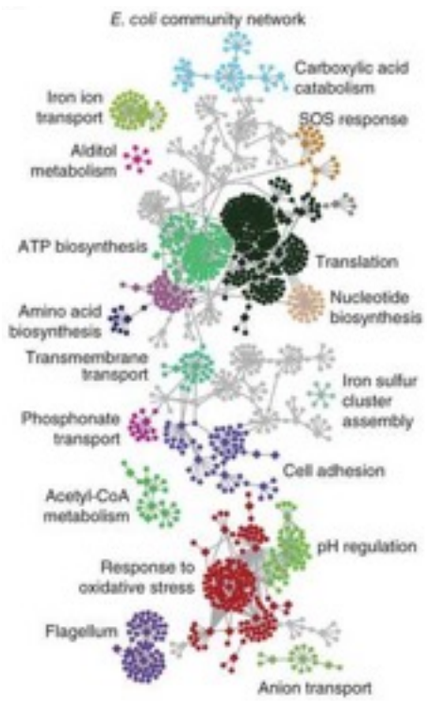


**Large genomics datasets require software (tools) to perform each of the steps in an NGS analysis workflow.**




*Nature Reviews Genetics* **11**, 559-571 (August 2010)



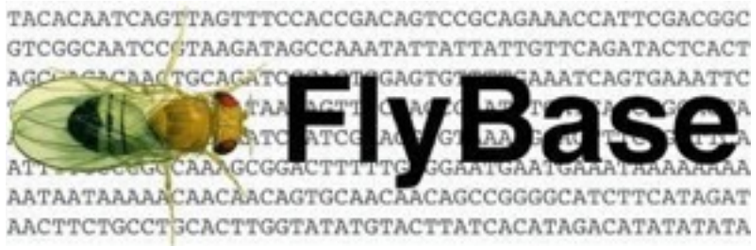

*Nature Methods* **9**, 796–804 (2012)

# Genome Databases

**Genome databases contain publicly available, searchable, and up-to-date genome data, including reference sequences and gene annotations**

MGI
CELEBRATING 25 YEARS

National Center for Biotechnology Information
NCBI




Ensembl

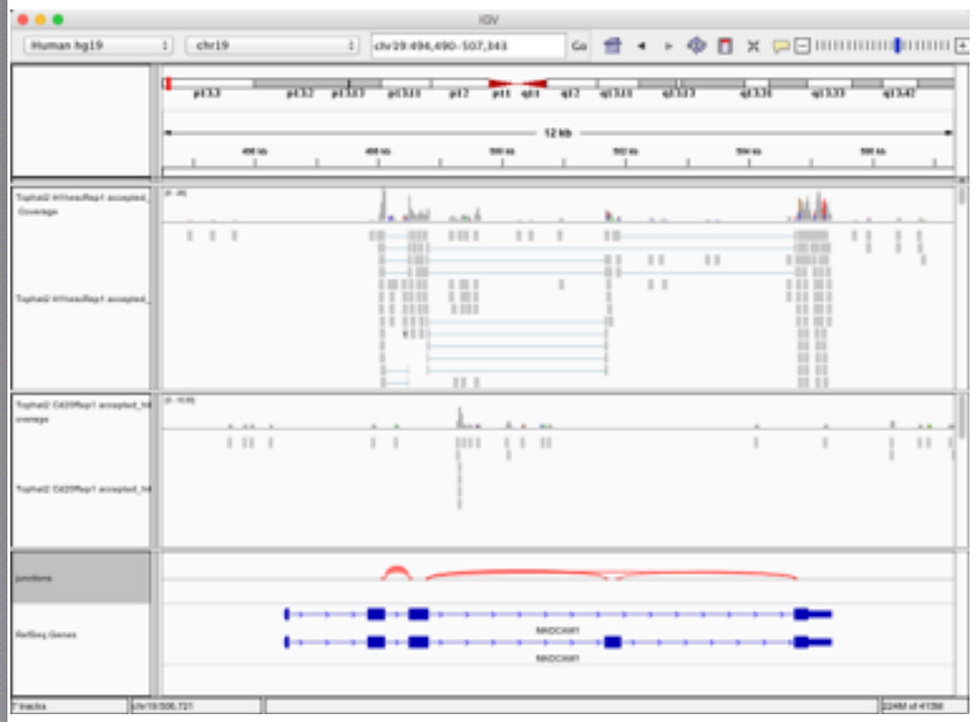
UCSC Genome Bioinformatics

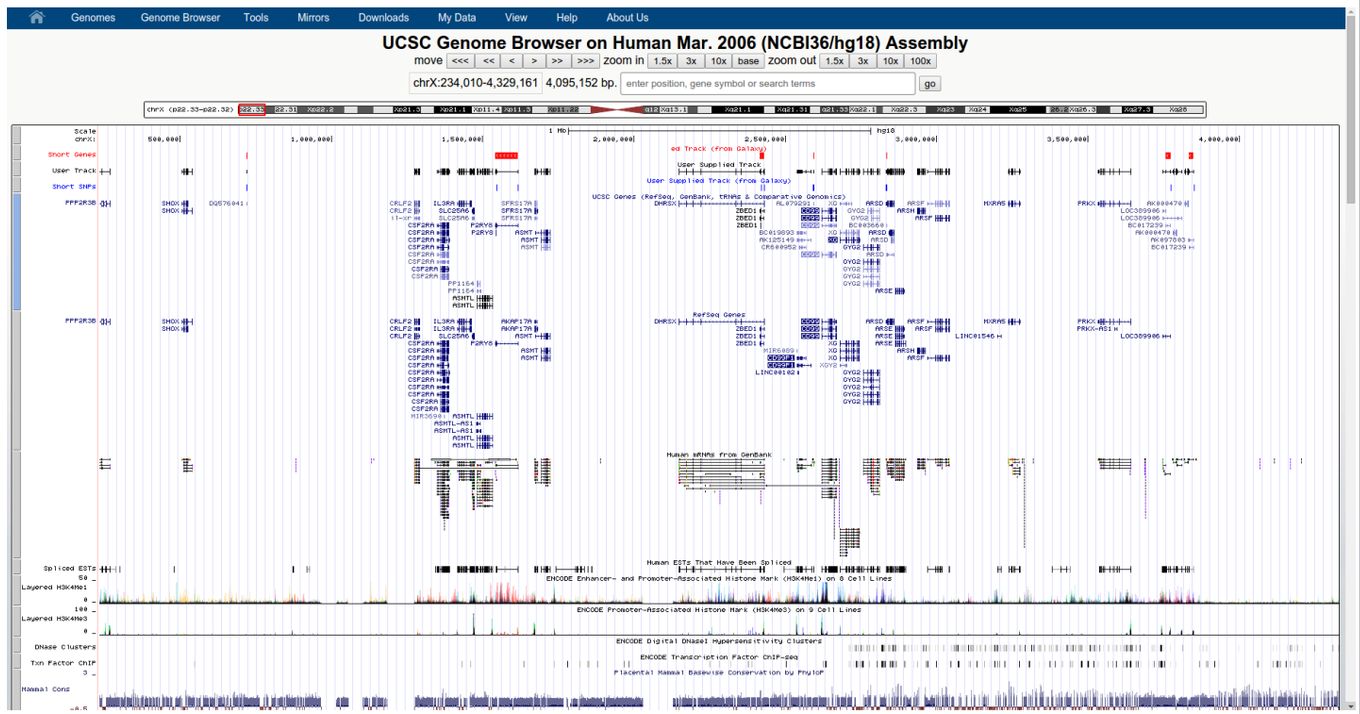
ViPR
Virus Pathogen Resource

PATRIC
Pathosystems Resource Integration Center

WormBase

# Genome Browsers




**Visualize genomic data from**:

- ***Genome databases:*** entire genomes, regulation sequences, gene predictions and structures, and data from comparative analyses.
- ***Your own analyses***

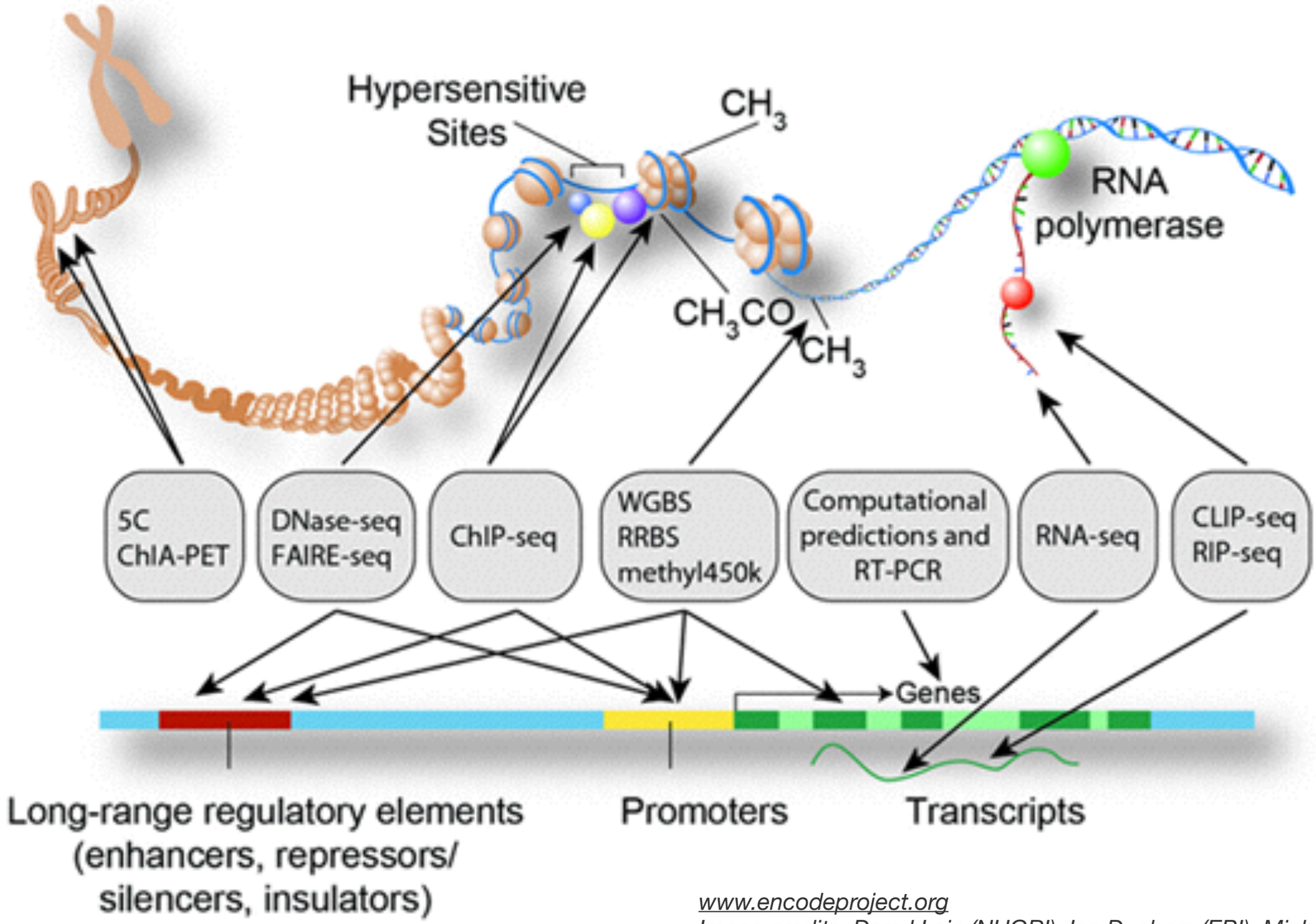
# Community

**Seek and respond to questions regarding genomics analyses and use of tools**



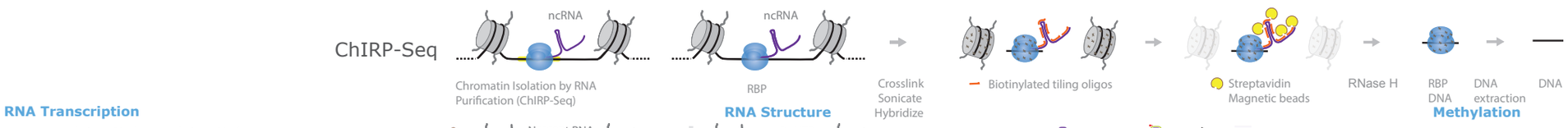
# NGS Applications

# NGS Applications

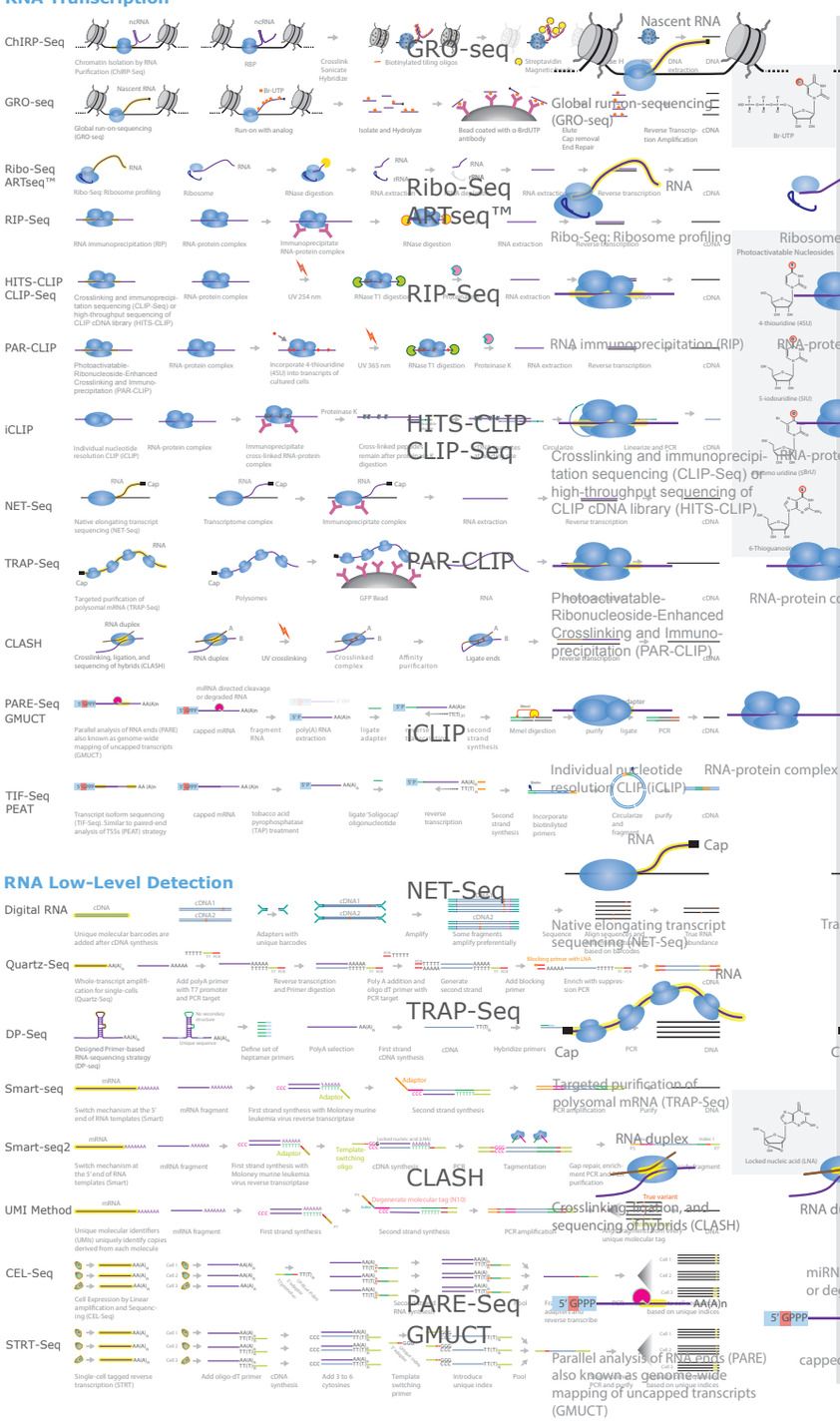



www.encodeproject.org
*Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)*

# RNA Transcription



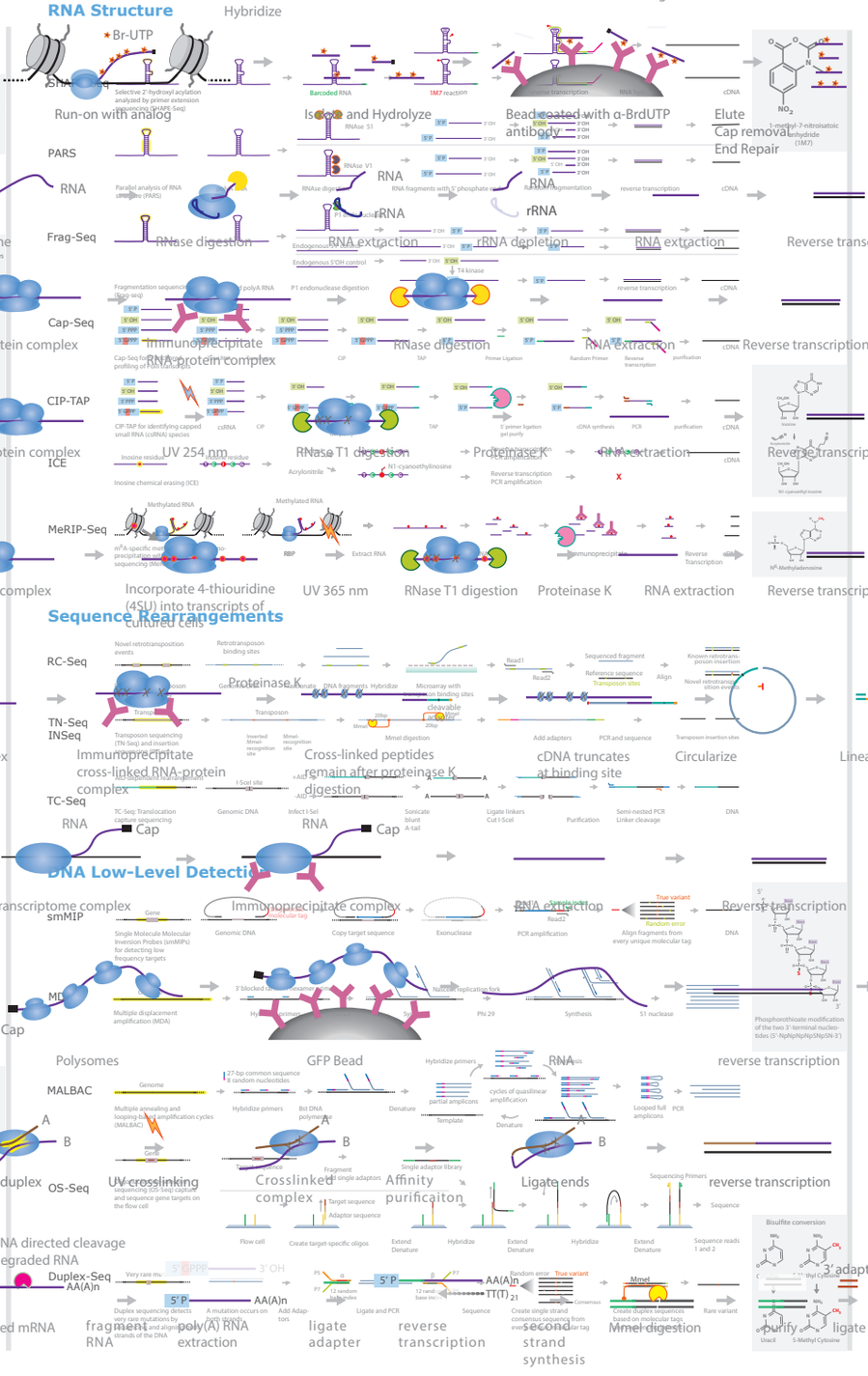

## RNA Transcription




## RNA Low-Level Detection




## RNA Structure



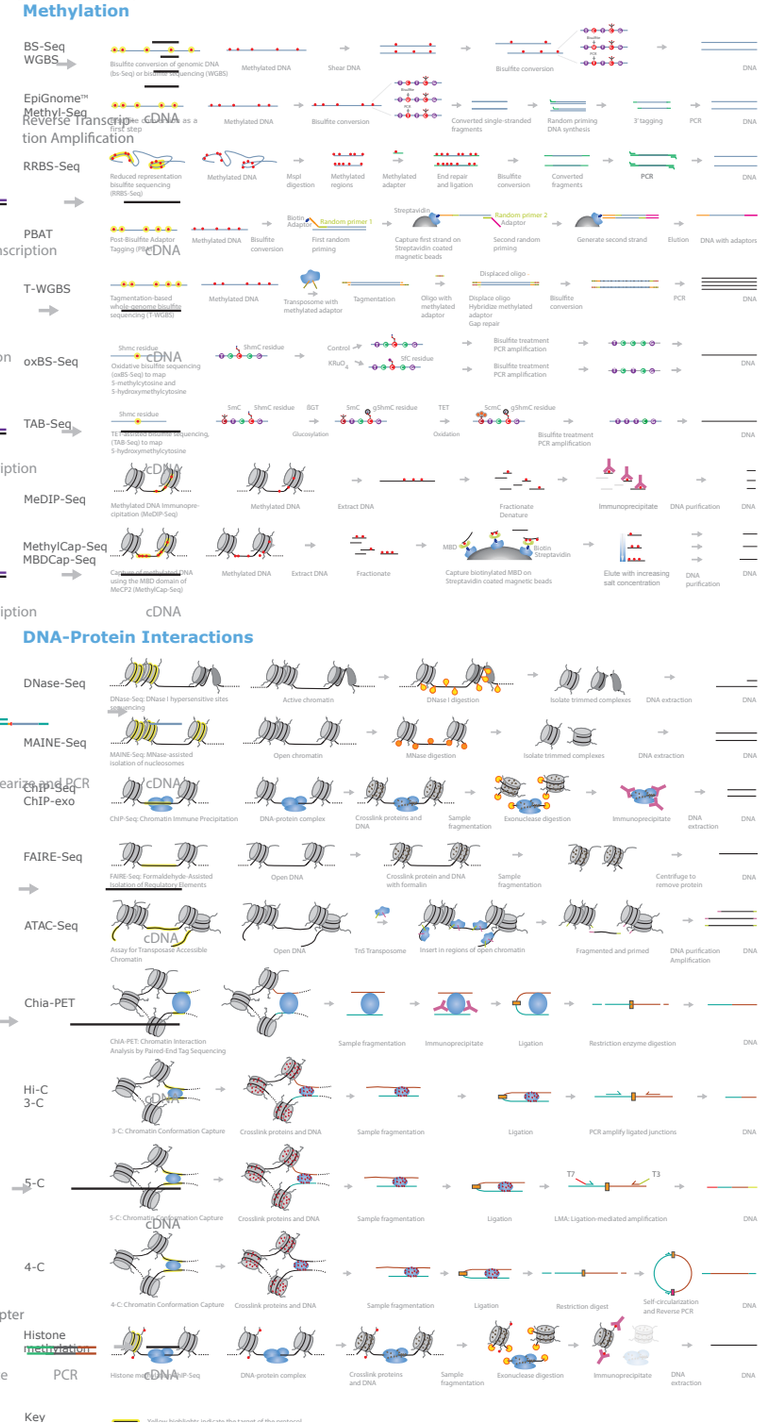

## Sequence Rearrangements




## DNA Low-Level Detection




## Methylation




## DNA-Protein Interactions

# Expectations


- **Unix / Orchestra**
- **R**
- **Genome databases / browsers**
- **Analysis tools and workflows**
- **Best practices**