



Overview of Bioinformatics

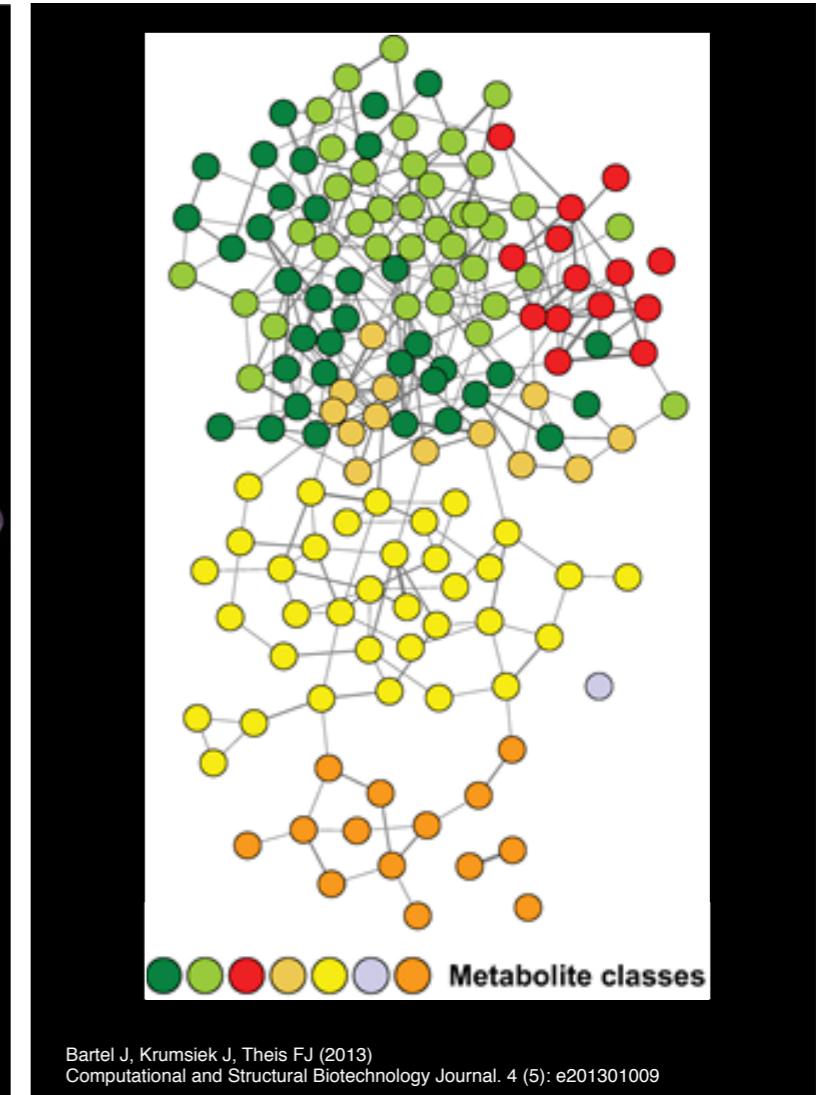
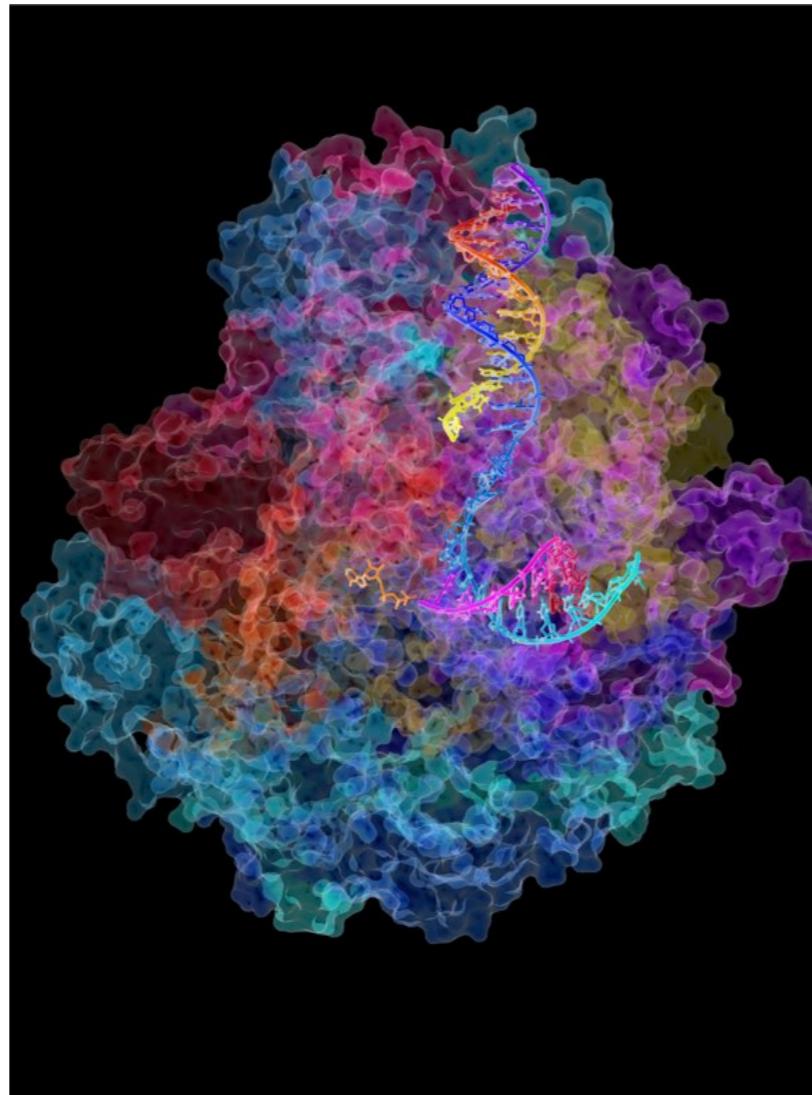
Harvard Chan Bioinformatics Core

NGS Data Analysis Course 2016

What is Bioinformatics?



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



Bioinformatics in the Omics Era

Why Genomics?



shutterstock_97071 Copyright: Sergey

Why Genomics?

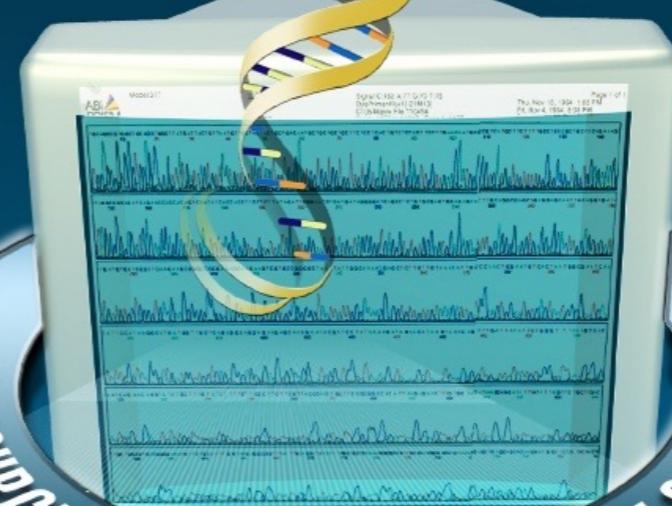
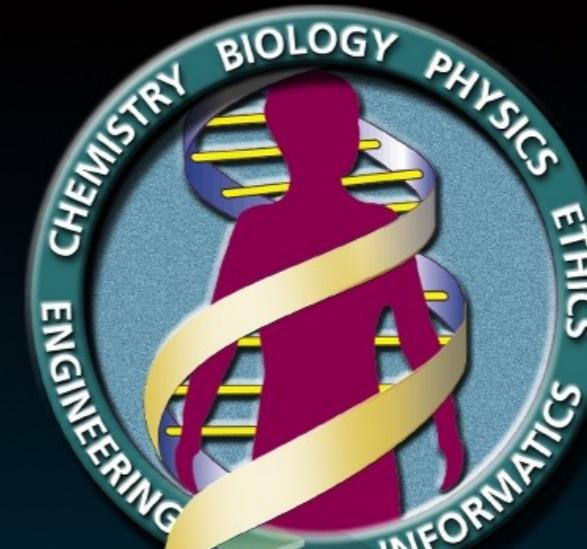


shutterstock_97071 Copyright: Sergey

**High Throughput
Comprehensive
Exploratory**

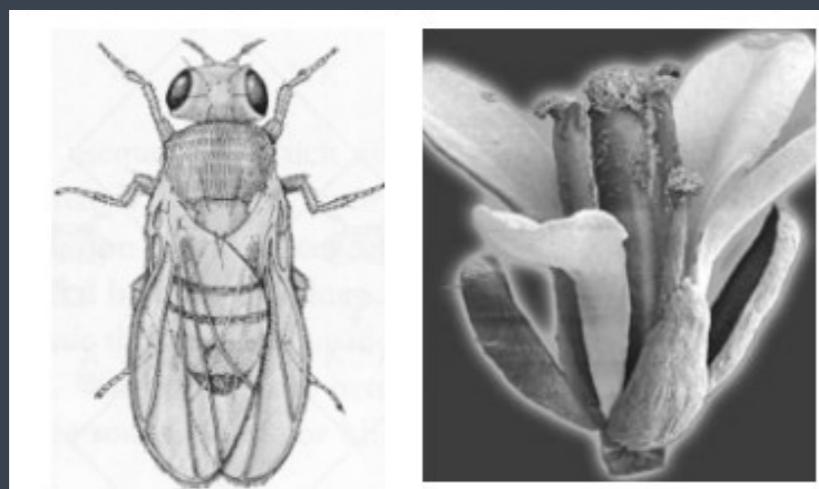
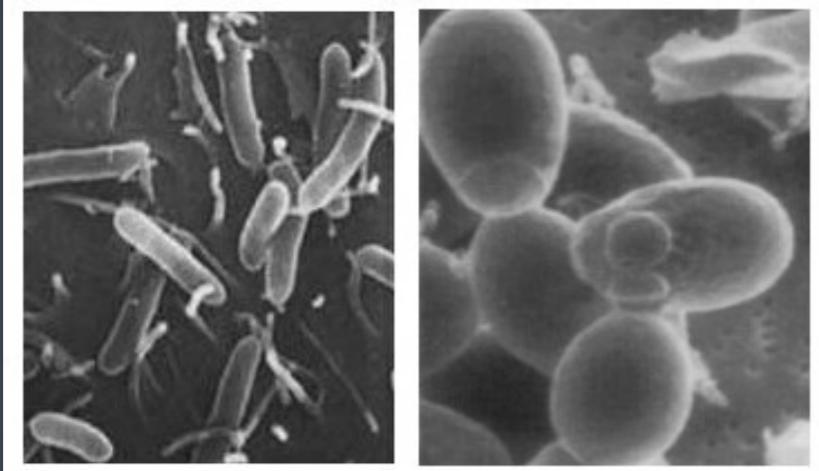
Human Genome Project

1990 - 2003



RESOURCES FOR THE BIOLOGY CENTURY

Sequenced organisms (2000)



38 bacteria

S. cerevisiae

C. elegans

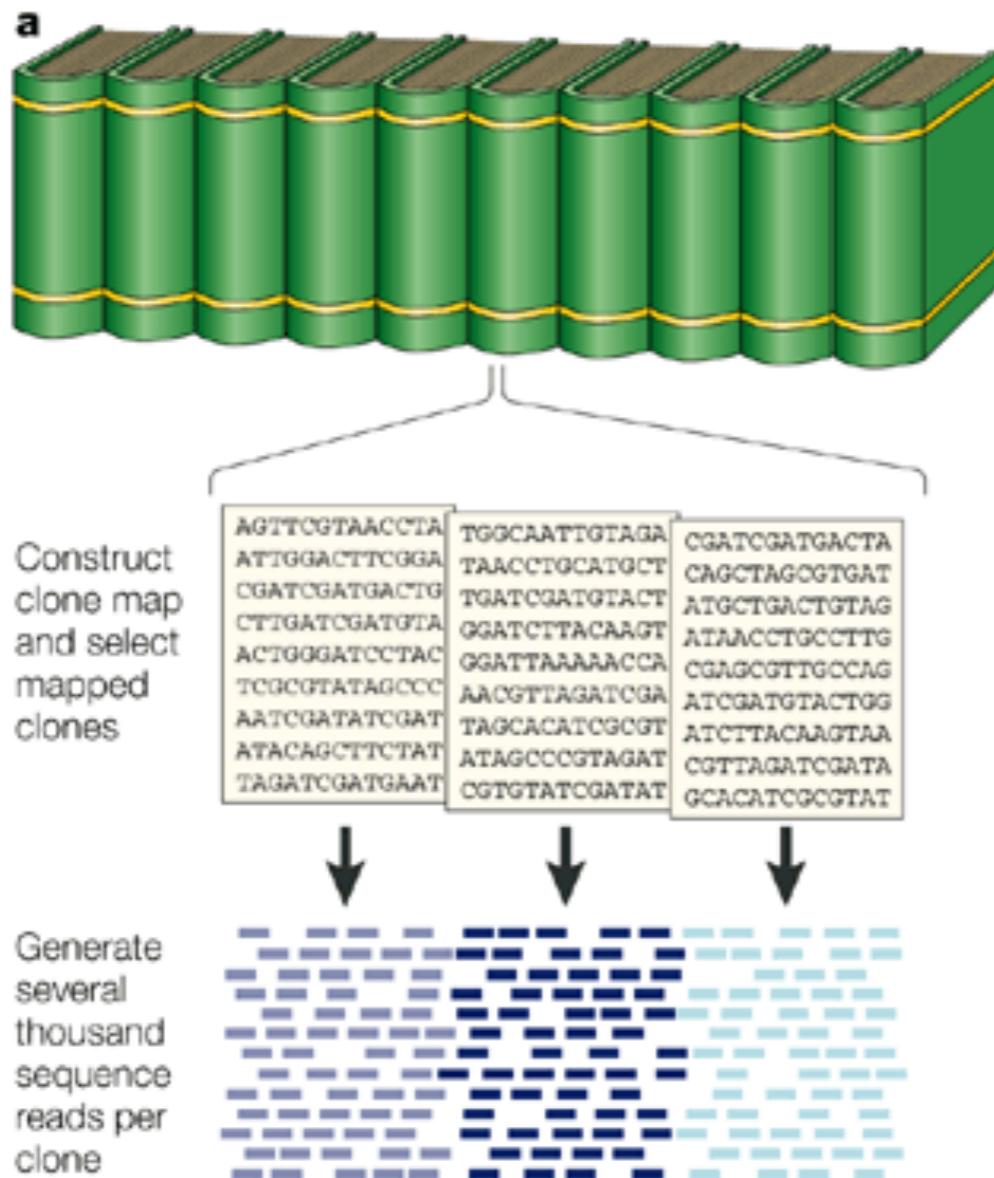
D. melanogaster

A. thaliana

Sequencing of the Human Genome

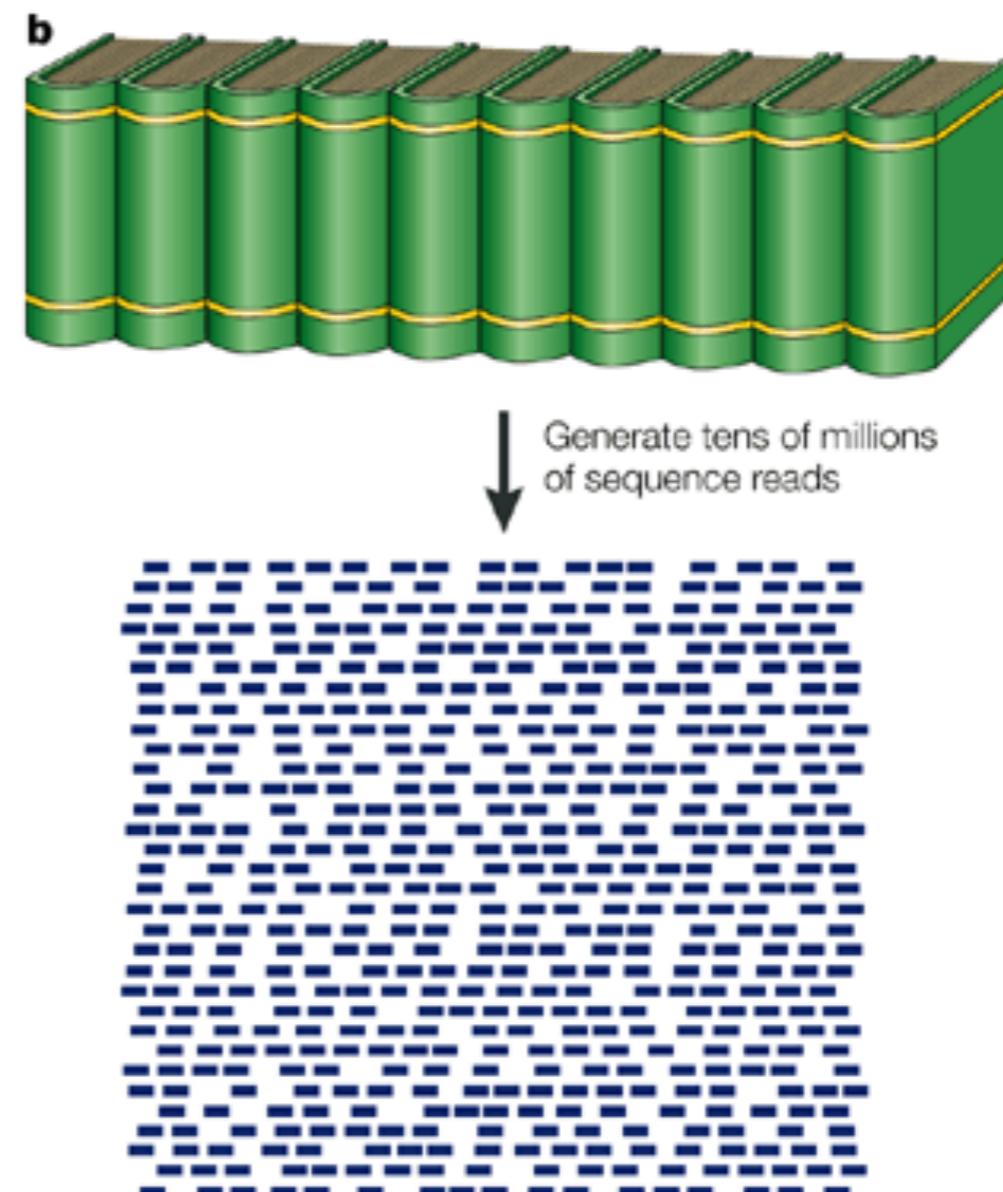
Human Genome Project

Heirarchical Shotgun Sequencing

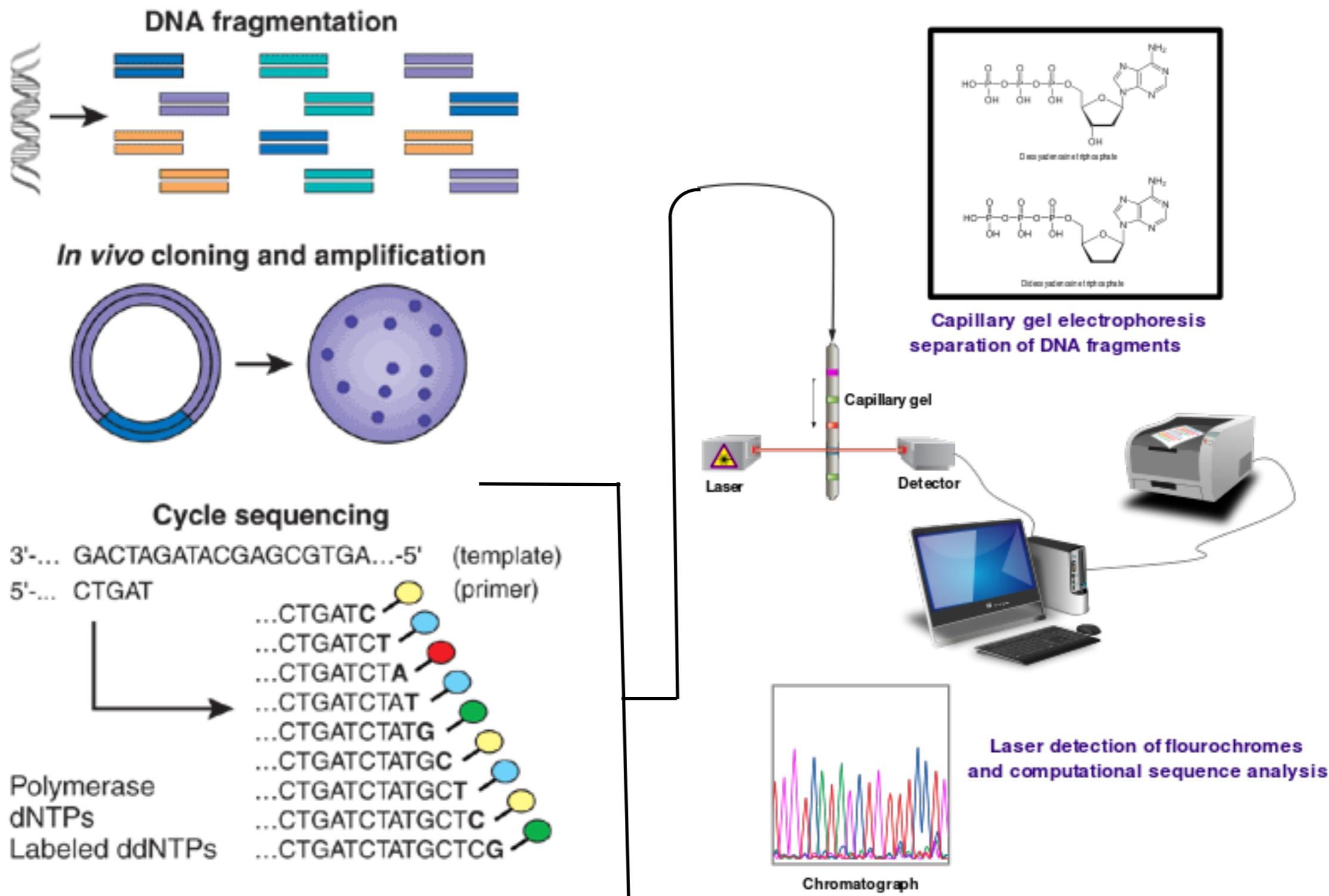


Celera Genomics

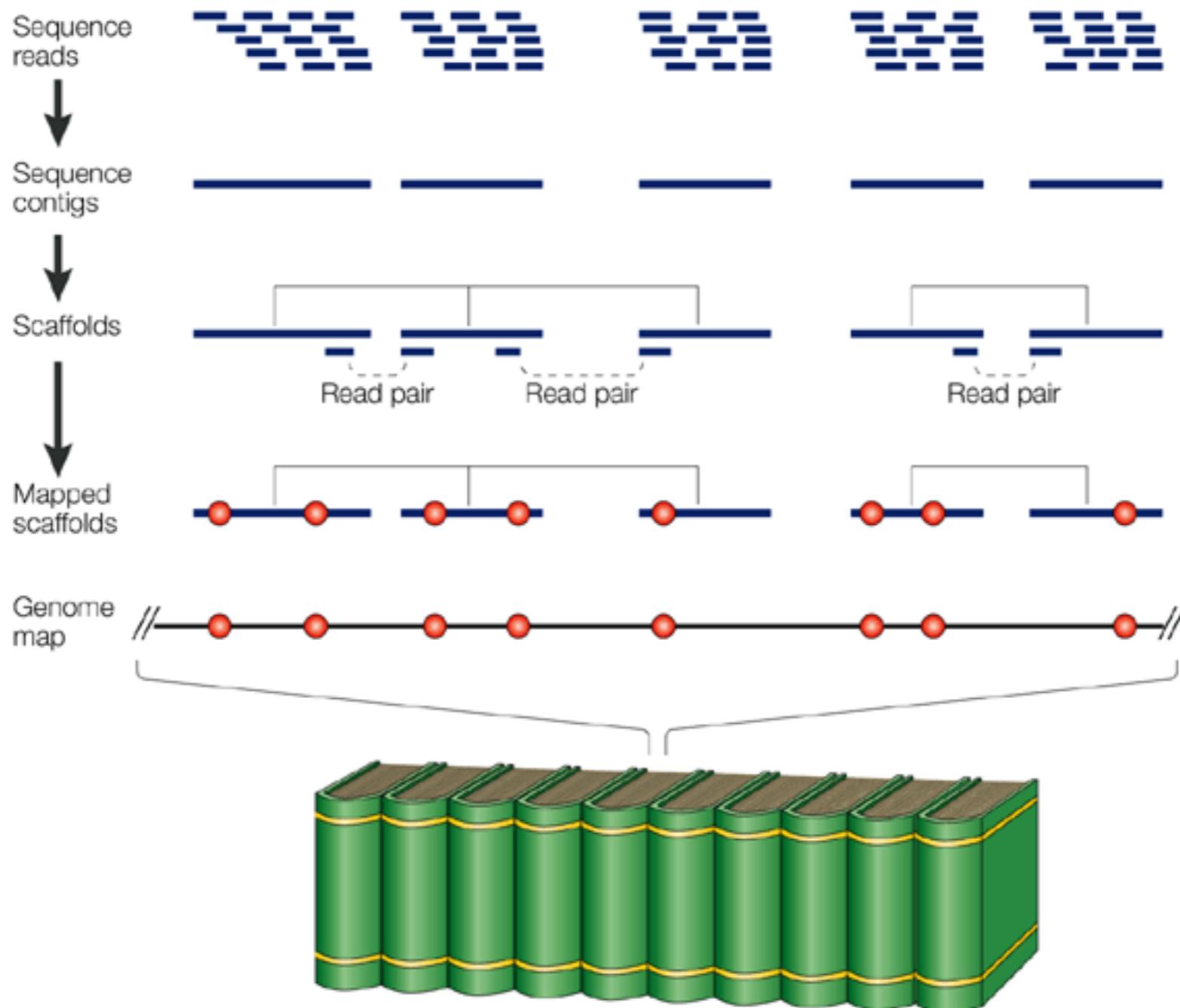
Whole Genome Shotgun Sequencing



Sequencing of the Human Genome



Sequencing of the Human Genome



Sequencing of the Human Genome



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

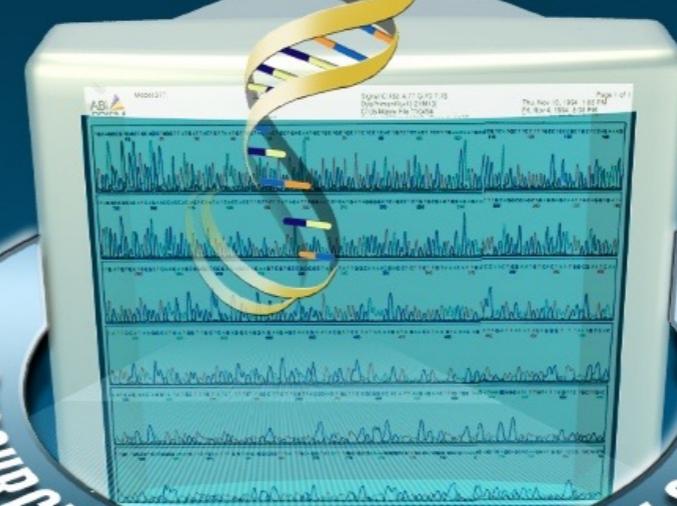
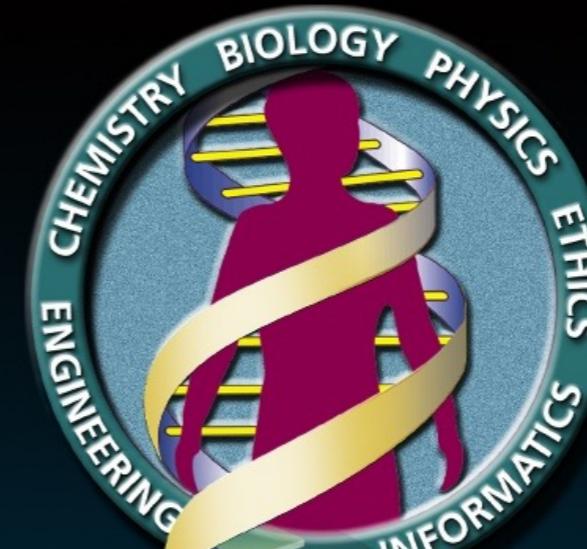
**Capillary (cycle)
sequencing generated:**

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

Sequence production was rate limiting, not analysis

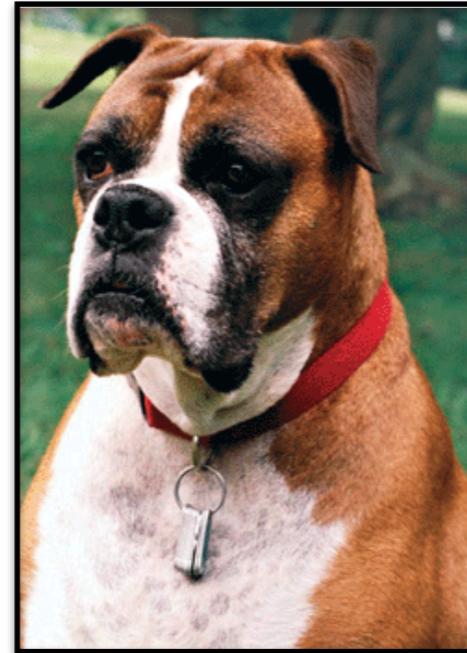
Human Genome Project

1990 - 2003



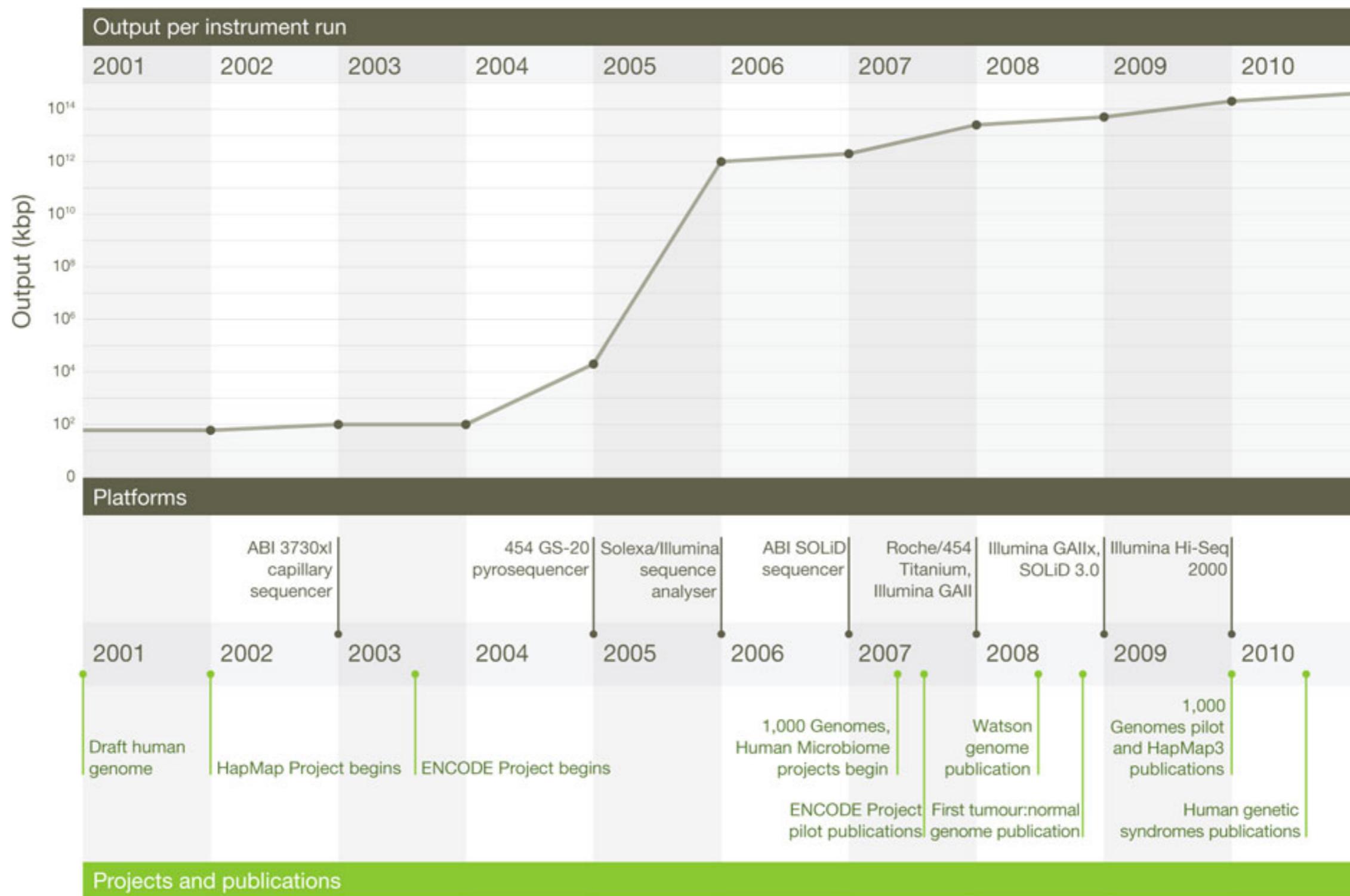
RESOURCES FOR THE BIOLOGY CENTURY

Comparative Genomics Research



Highly conserved regions of DNA likely to be functional

Advancements in Sequencing Technology



NGS Technologies



HiSeq2000 / 2500

Ion torrent



Ion Proton



MiSeq

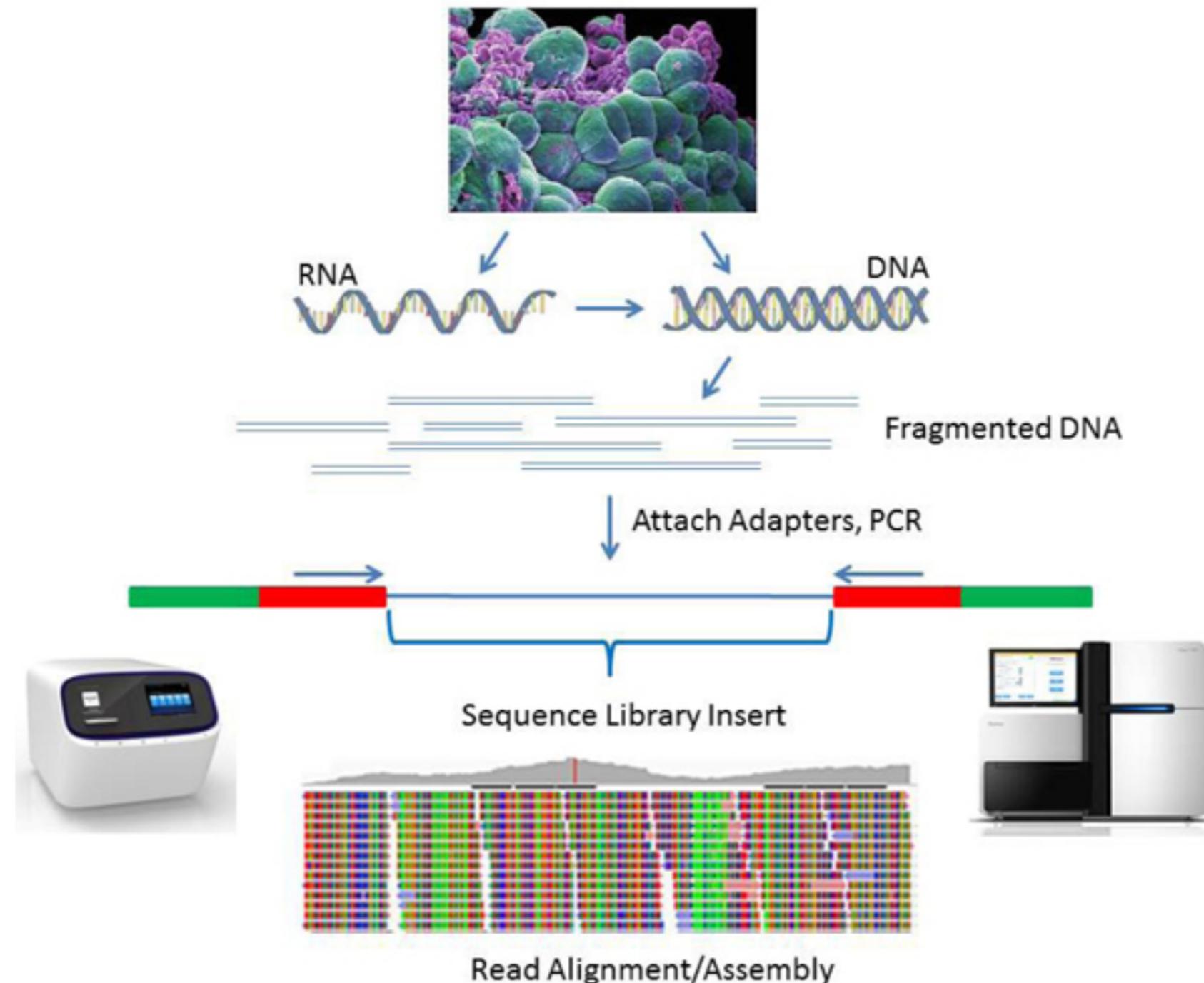
GS-FLX



PacBio RS

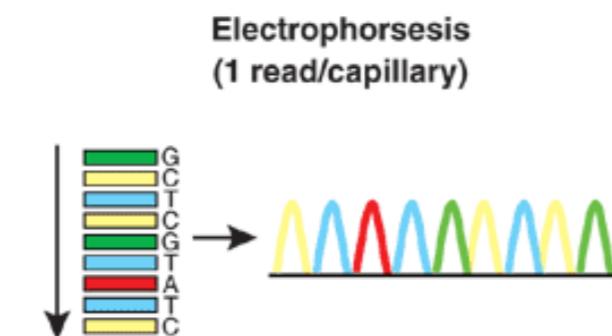
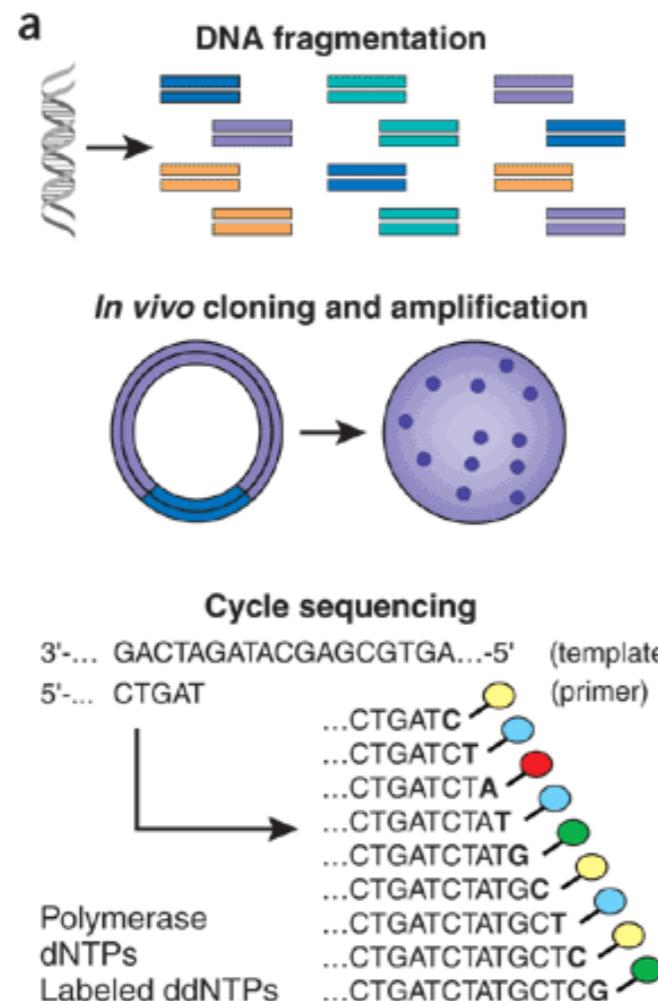
Platform	Chemistry	Read Length	Run Time	Gb/Run	Advantage	Disadvantage
454 GS FLX+ (Roche)	Pyro-sequencing	700	23 hrs.	0.7	Long Read Length	High error rate in homopolymer
HiSeq (Illumina)	Reversible Terminator	2*100	2 days (rapid mode)	120 (rapid mode)	High-throughput / cost	Short reads Long run time (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

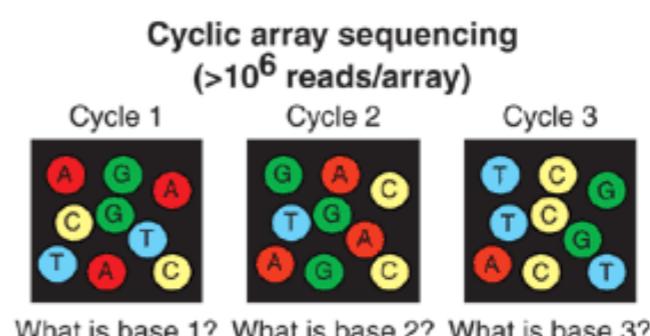
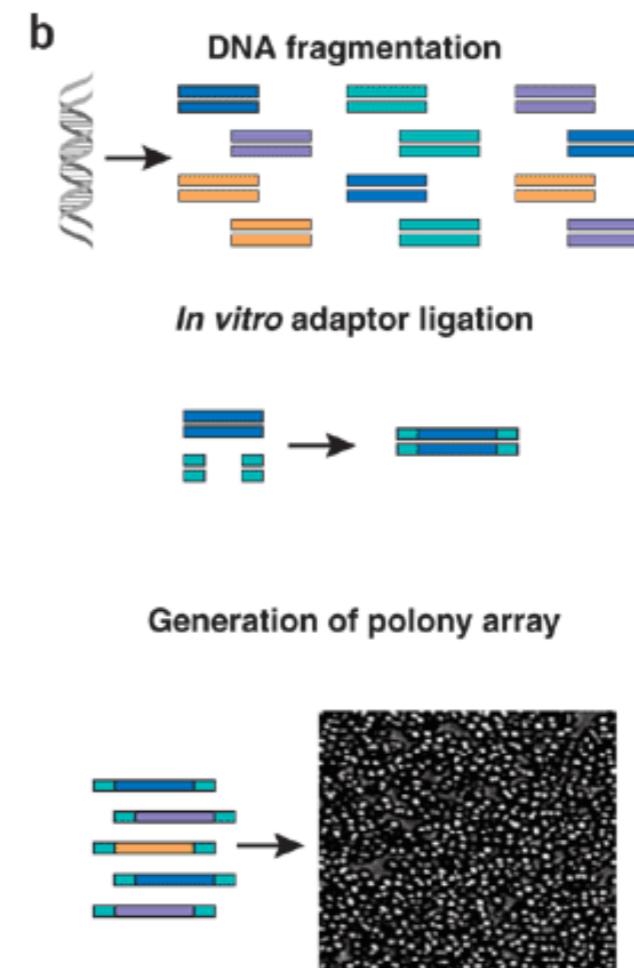


NGS Technologies

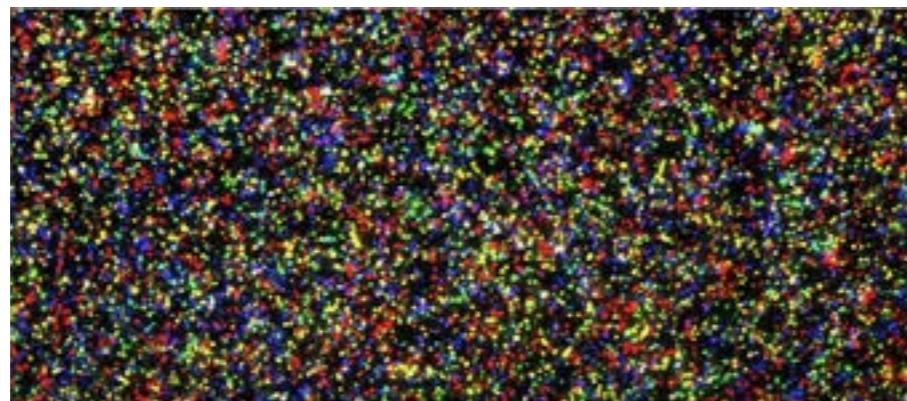
Cycle Sequencing



NGS Sequencing



NGS Technologies



Generate:

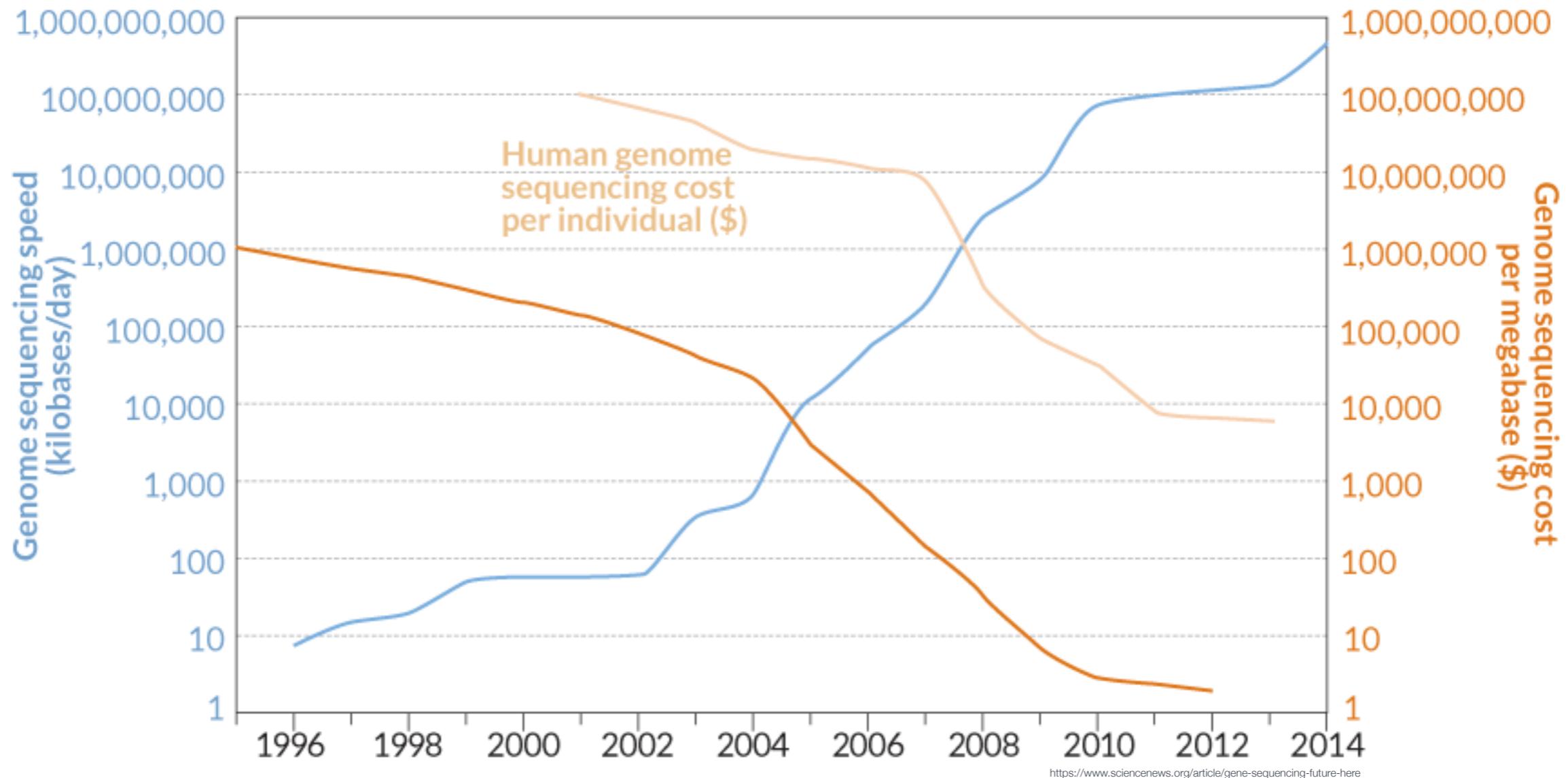
- Illumina HiSeq2500: 125 bp / read (250 bp / read Rapid-Run Mode)
- > **100 billion bp** / day

Bioinformatics support required to handle:

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

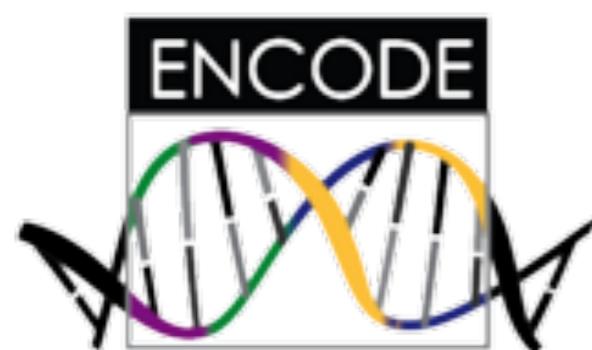
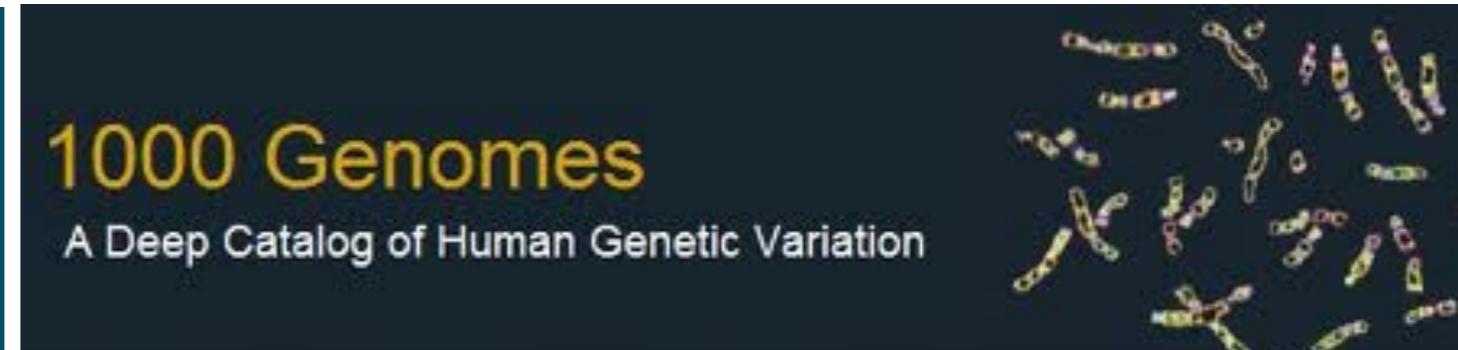
Sequence analysis rate limiting, not production

NGS Technology Accessibility



Increased scale and lower cost increases access to sequencing technologies

The Genomic Era: Collaborative Projects



The Genomic Era: Individual Projects



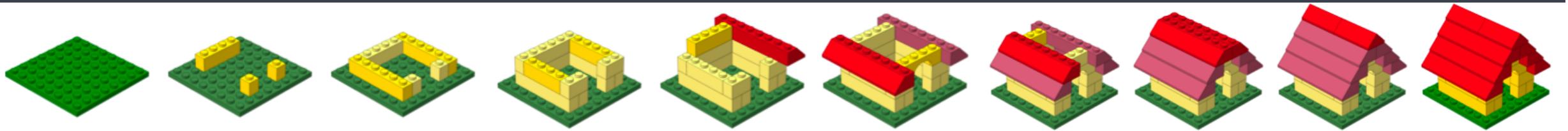
- Only a few experiments = vast amounts of data
- Data generation straightforward, but analysis requires bioinformatics expertise

Bioinformatics in the Genomic Era



Alliances between experimentalists and computational biologists

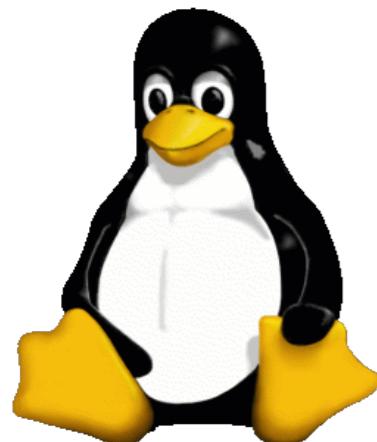
Bioinformatics Toolkit



Programming languages



Programming languages are critical in genomic analyses:

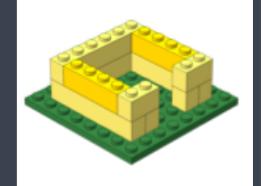


- ▶ **Bash:** command line language used for interacting with *Linux / Unix-based* operating systems.
 - Attaining sequencing data
 - Accessing computing resources and analysis tools
 - Basic data manipulation, creating scripts for running tools

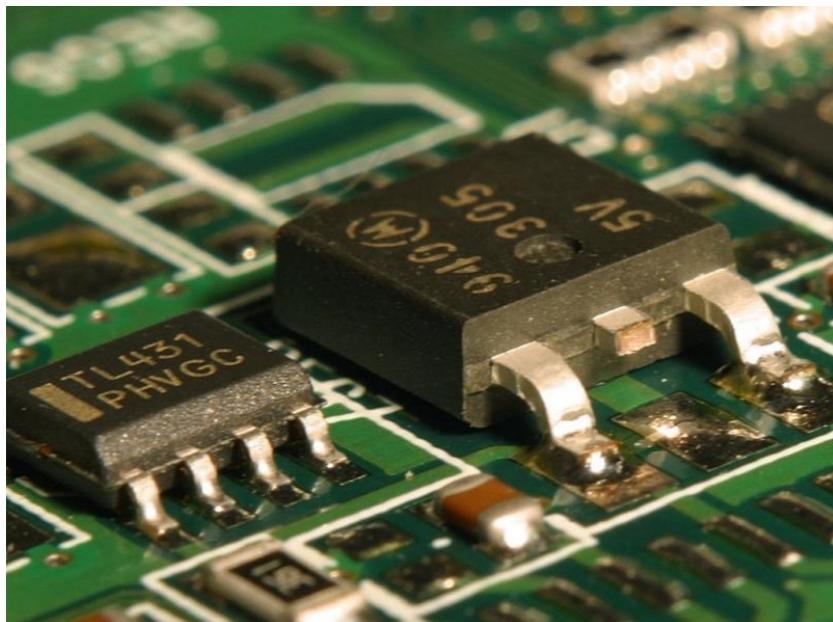
- ▶ **R:** a programming language and environment for statistical computing and graphics
 - Manipulating data, performing statistical analyses
 - Creating figures and plots
 - Sharing analyses and data



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

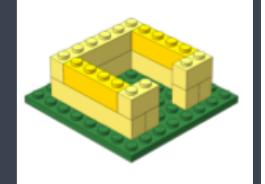
Memory/RAM (Random Access Memory)

- Large datasets ≈ lots of RAM to perform analysis

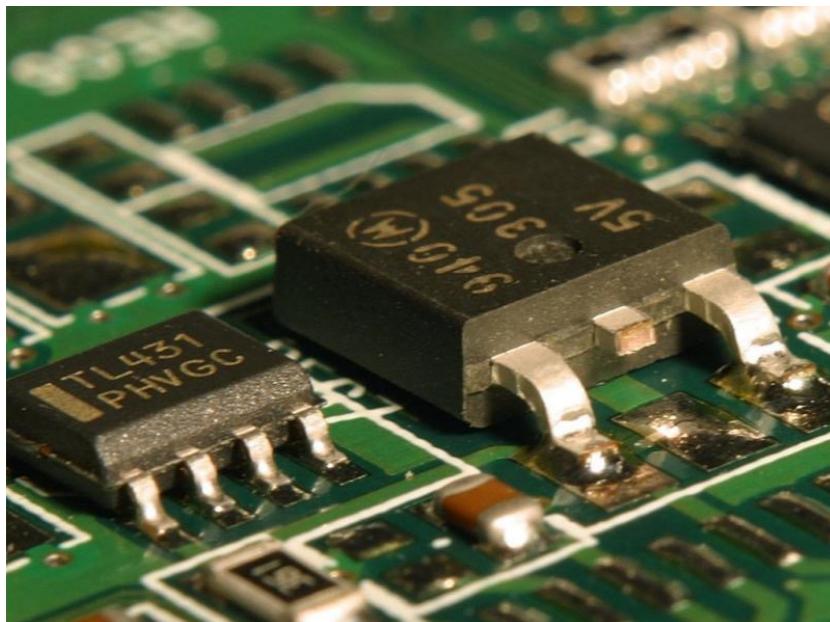
CPU (Central Processing Unit)

- Large datasets ≈ lots of time to perform analysis

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)

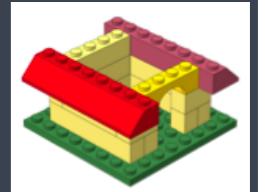
- Large datasets ≈ lots of RAM to perform analysis

CPU (Central Processing Unit)

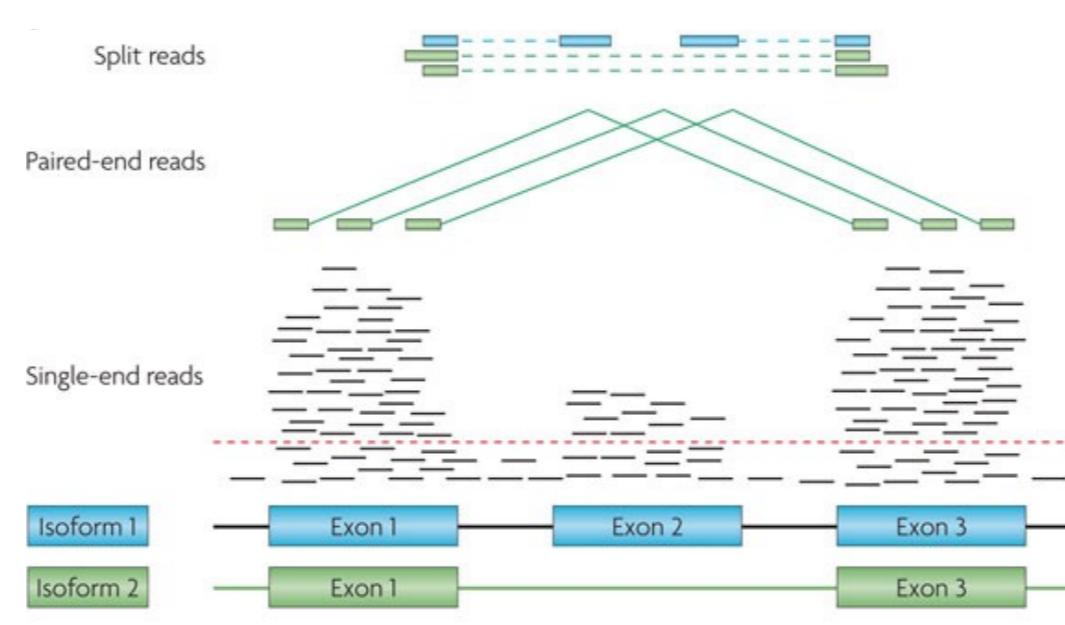
- Large datasets ≈ lots of time to perform analysis

Solution: Amazon Cloud or computing cluster

NGS Analysis Tools and Workflows

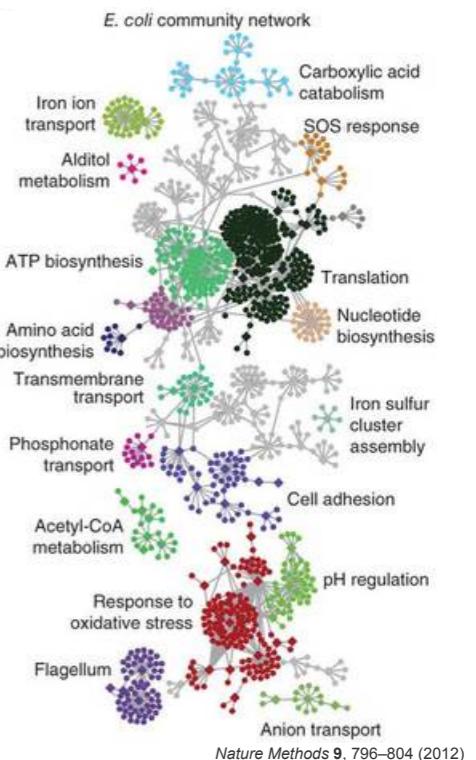


Large genomics datasets require software (tools) to perform each of the steps in an NGS analysis workflow.



Nature Reviews | Genetics

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



Genome Databases



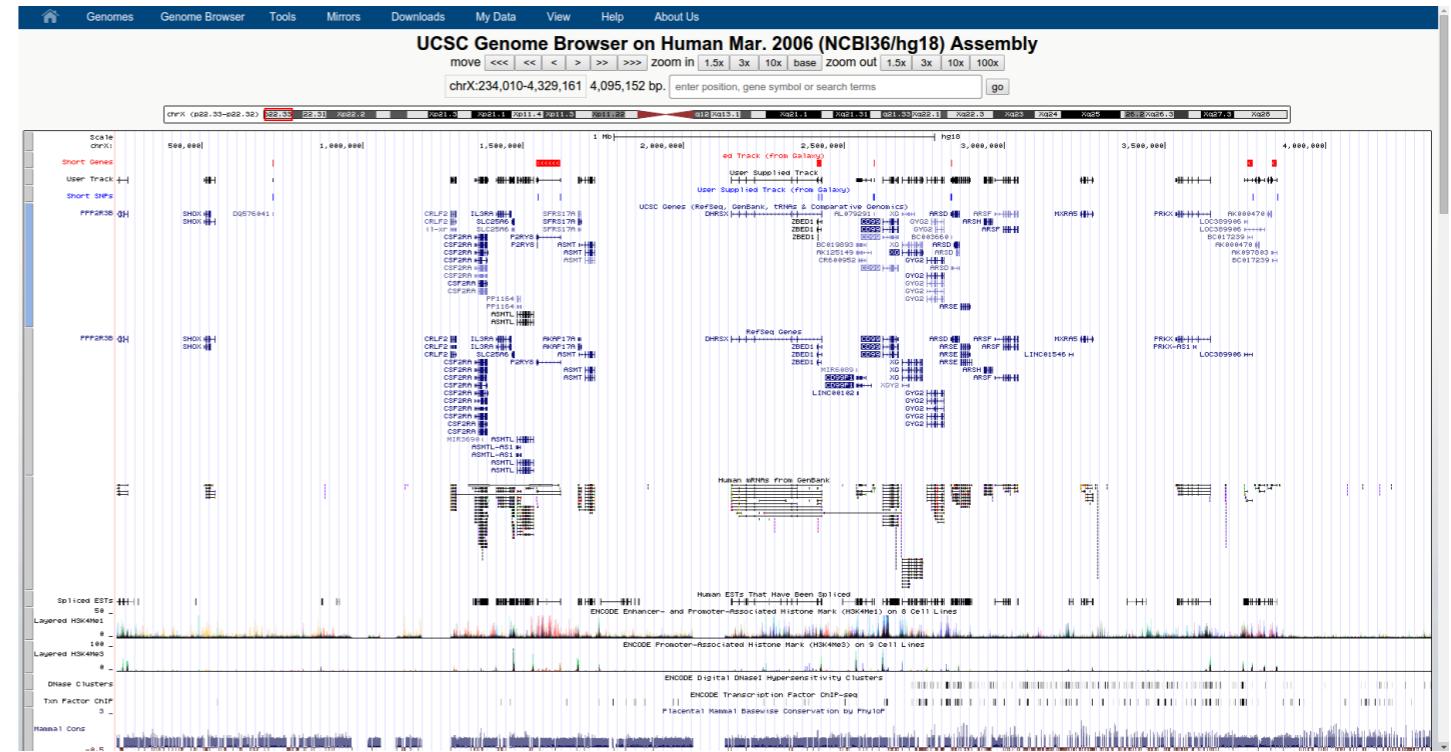
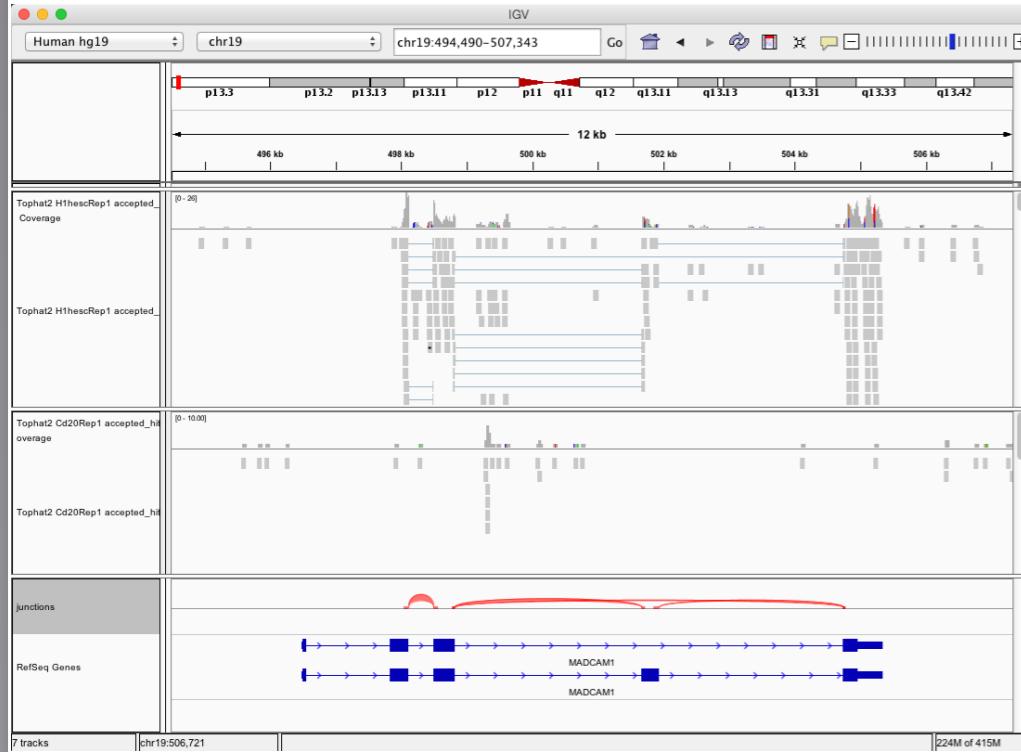
Genome databases contain publicly available, searchable, and up-to-date genome data, including reference sequences and gene annotations



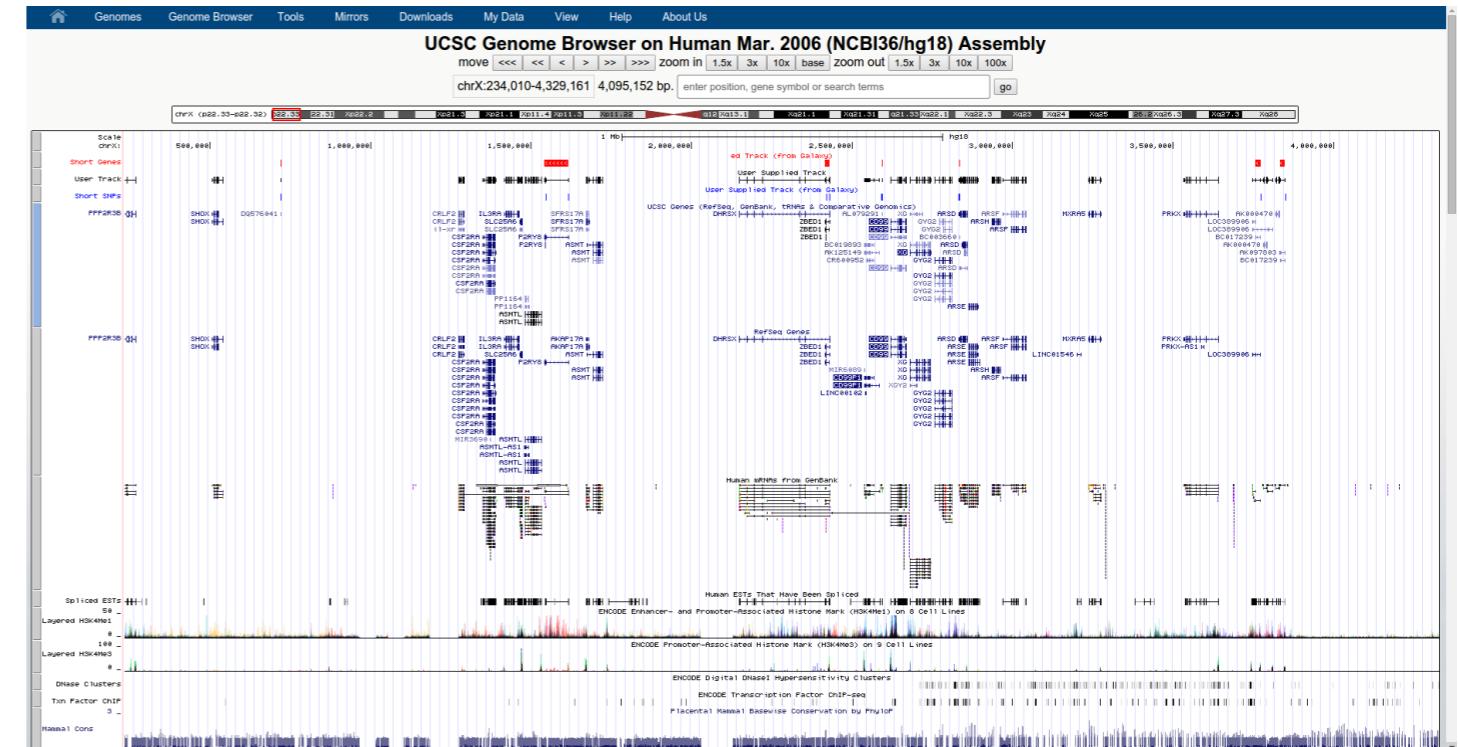
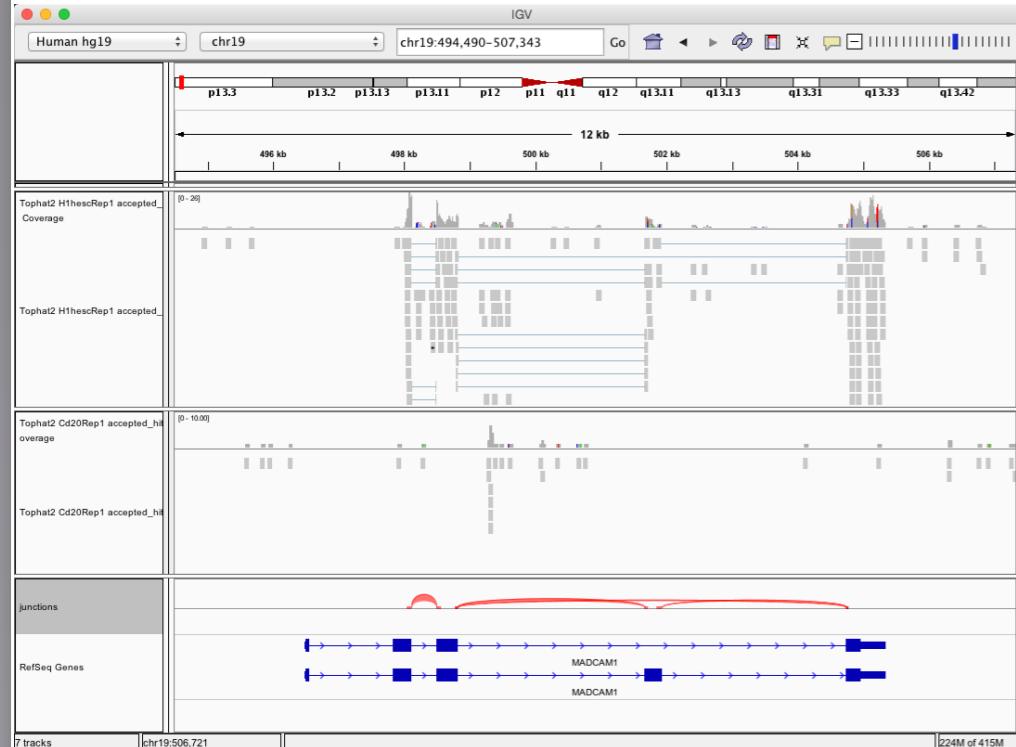
TACACAAATCAGTTAGTTCCACCGACAGTCCGCAGAAACCATTGACGGC
GTCGGCAATCCGTAAAGATGCCAAATATTATTGTTCAGATACTCACT
AGCCGCGCAACTGCAGATGCCAACTGAGTGTTCAAATCAGTGAATTC
TAAACTTCAACCGGATTCGATGAACTGAACTTCGATTA
ATC ATCG ACT GAA GT AA G AG TT C G T T A
ATT CCGG CAAAGCGGACTTTTG GGAATGAATGAAATAAAAAAA
AATAATAAAAACAAACAACAGTGCAACACAGCCGGGCATCTTCATAGAT
AACTTCTGCCTGCACTGGTATATGTACTTATCACATAGACATATATA



Genome Browsers



Genome Browsers

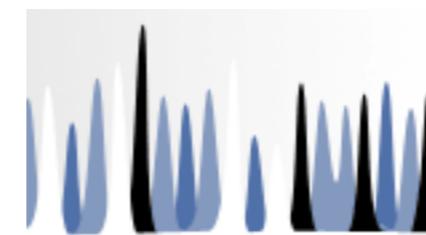


Visualize genomic data from:

- **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

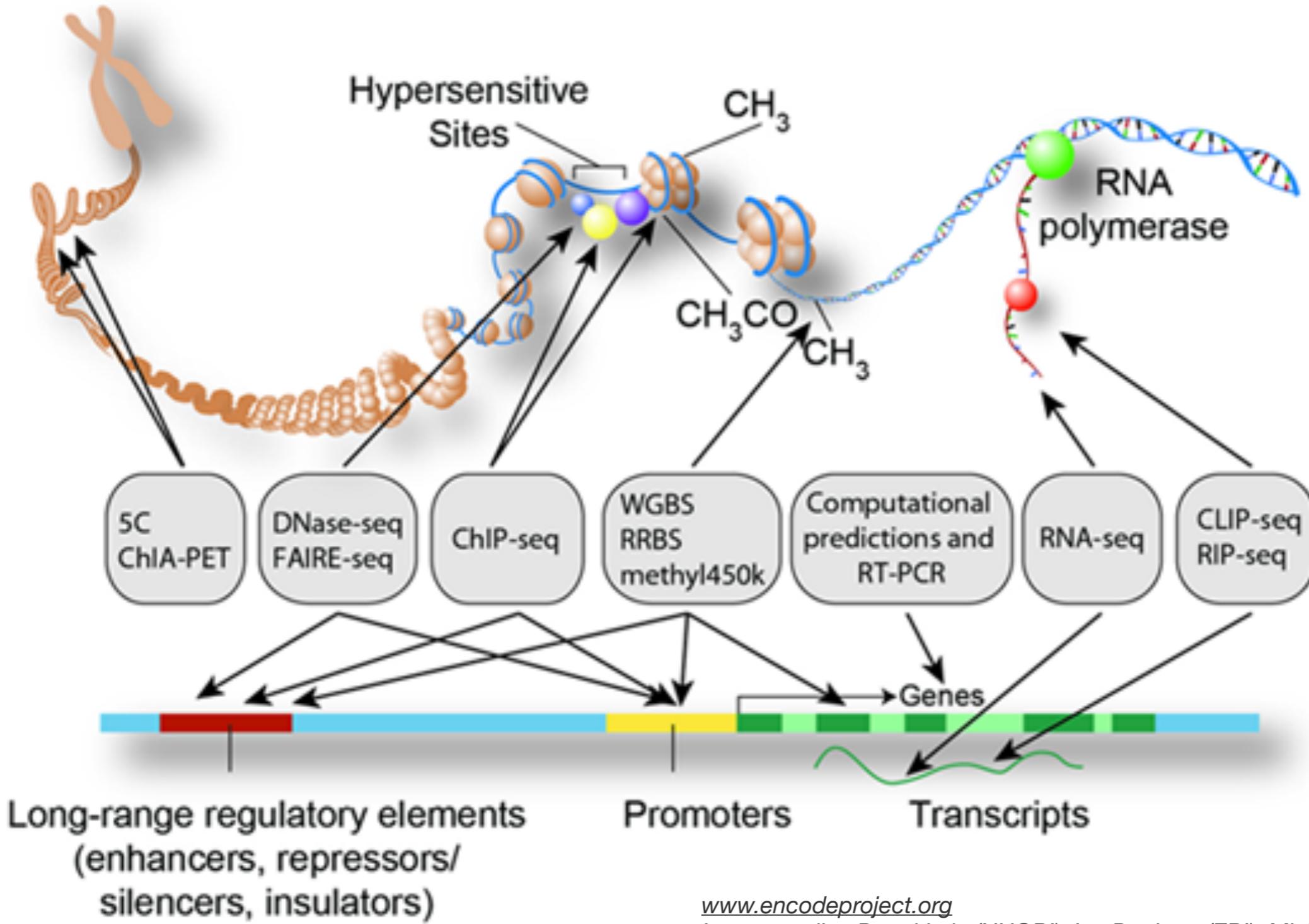


SEQanswers
the next generation sequencing community



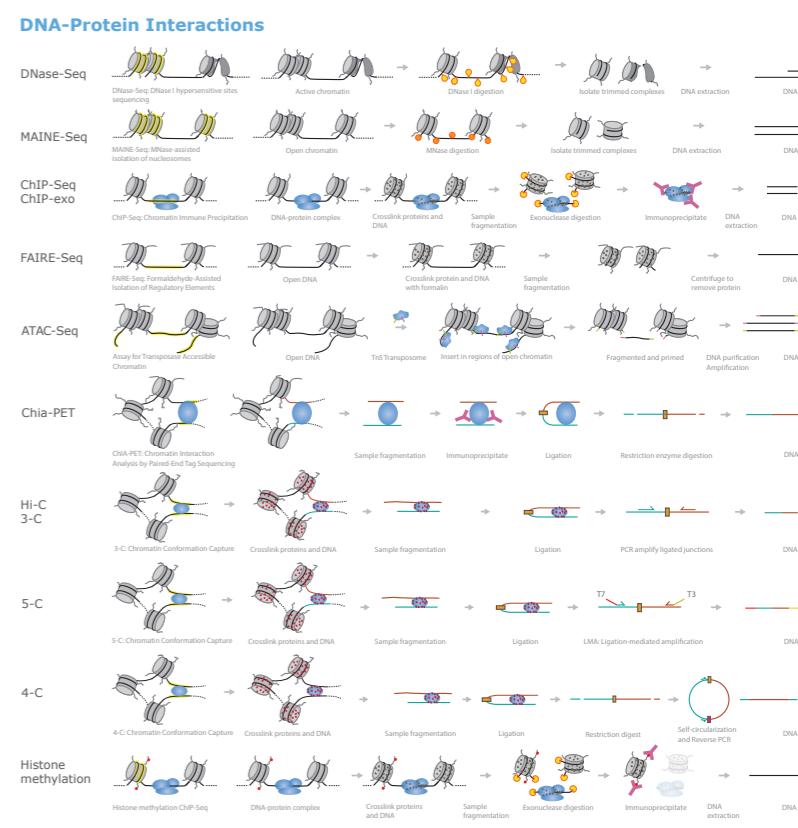
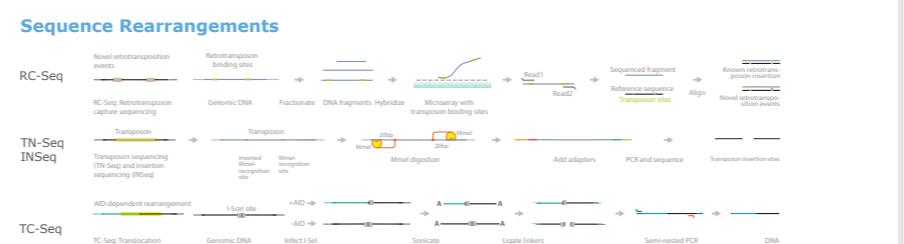
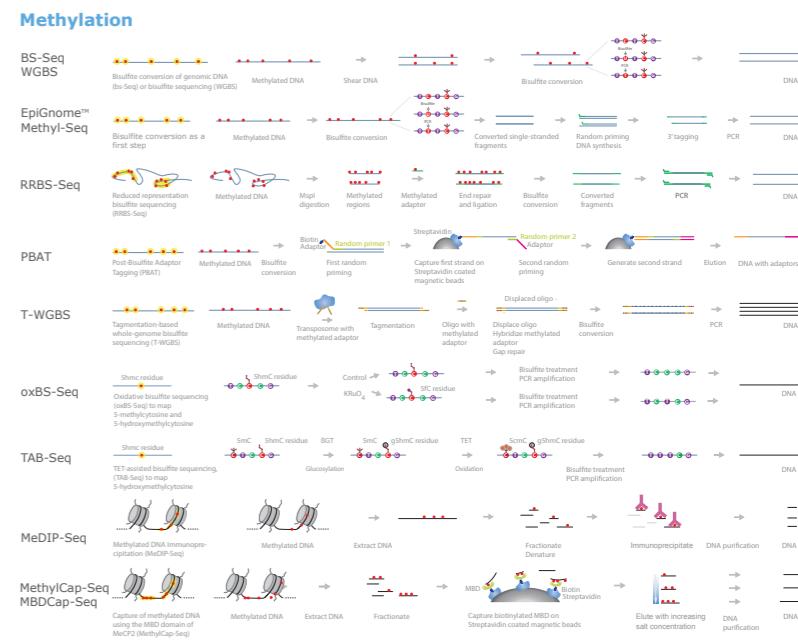
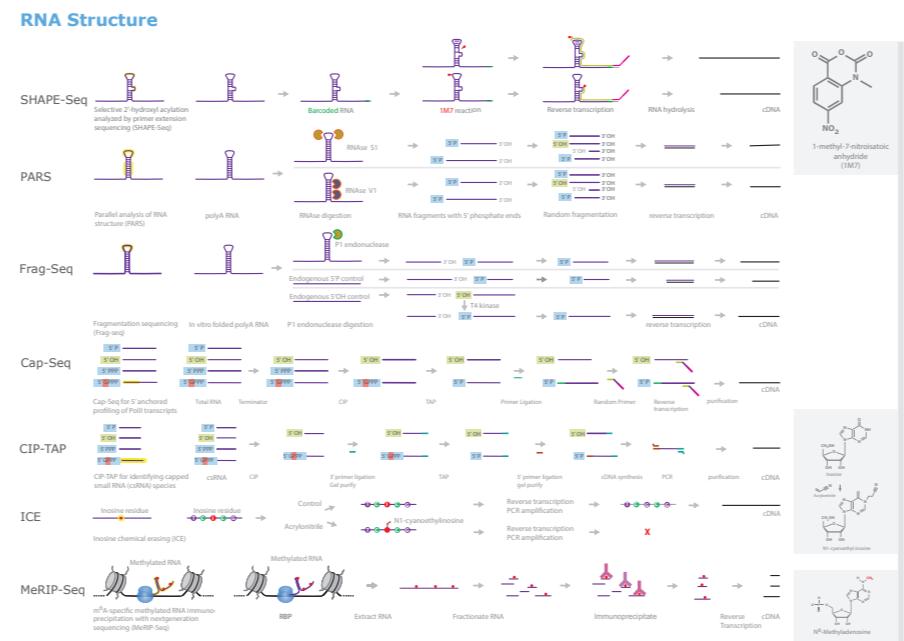
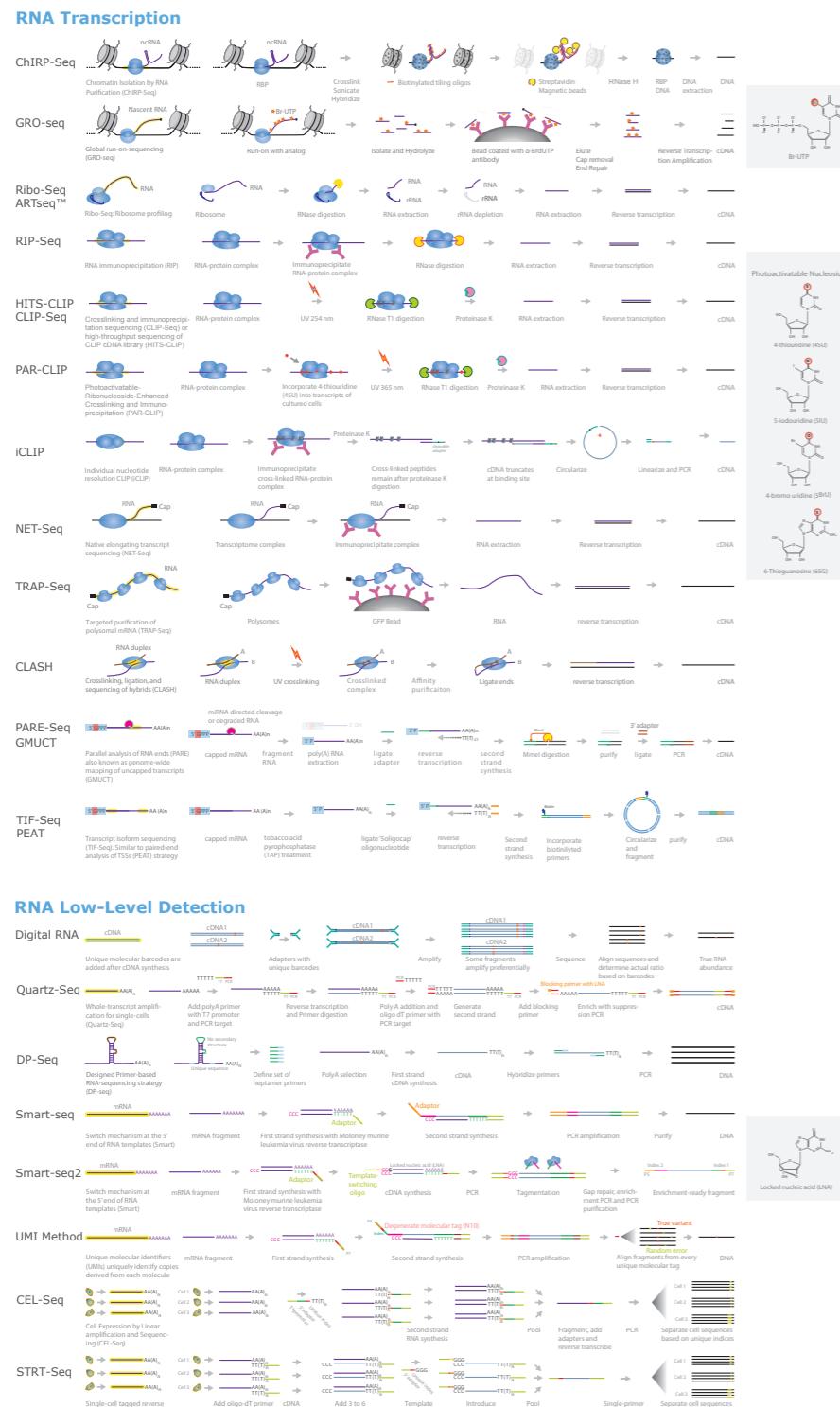
NGS Applications

NGS Applications

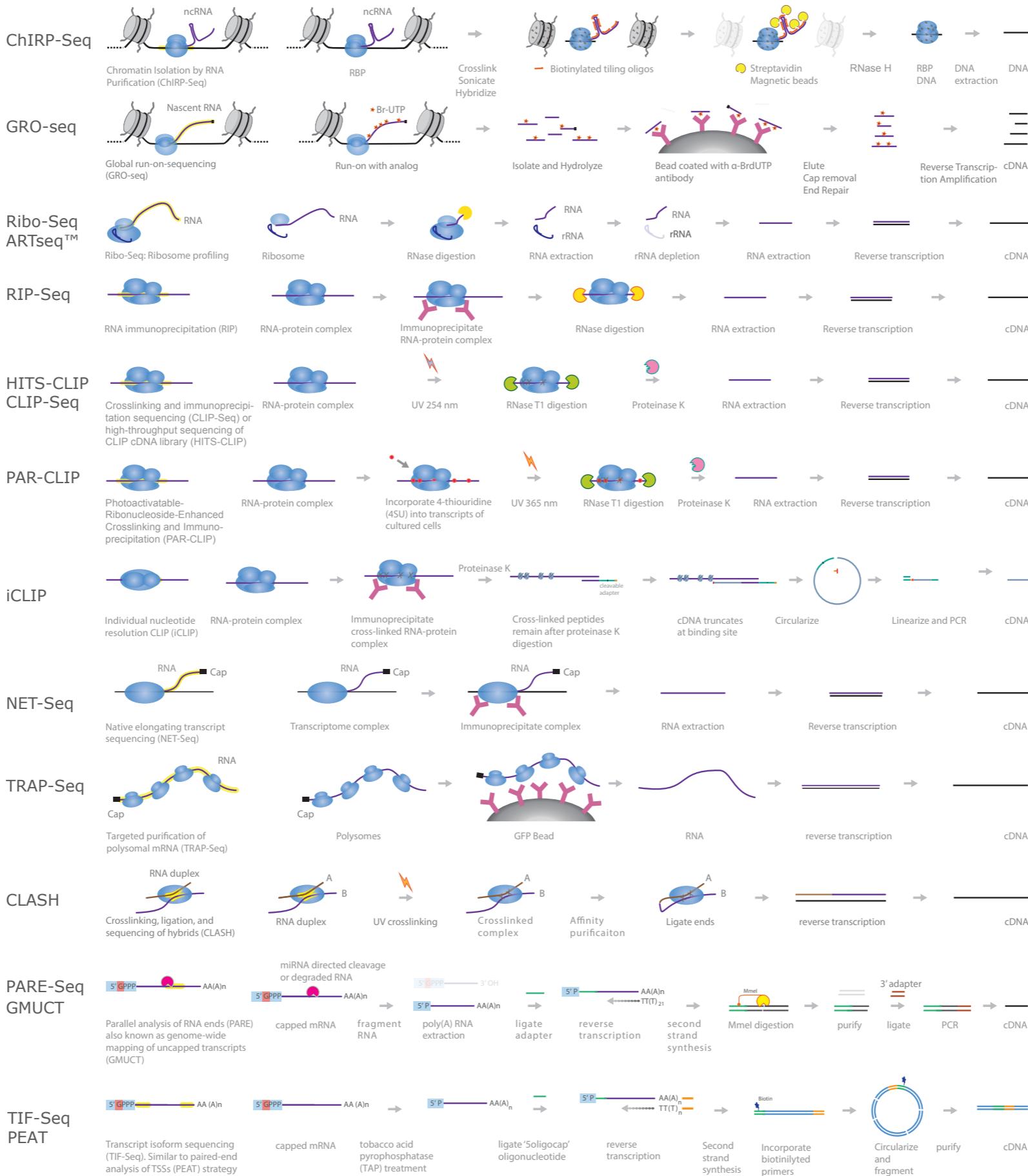


www.encodeproject.org

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)



RNA Transcription



Expectations



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

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.