



Overview of Bioinformatics

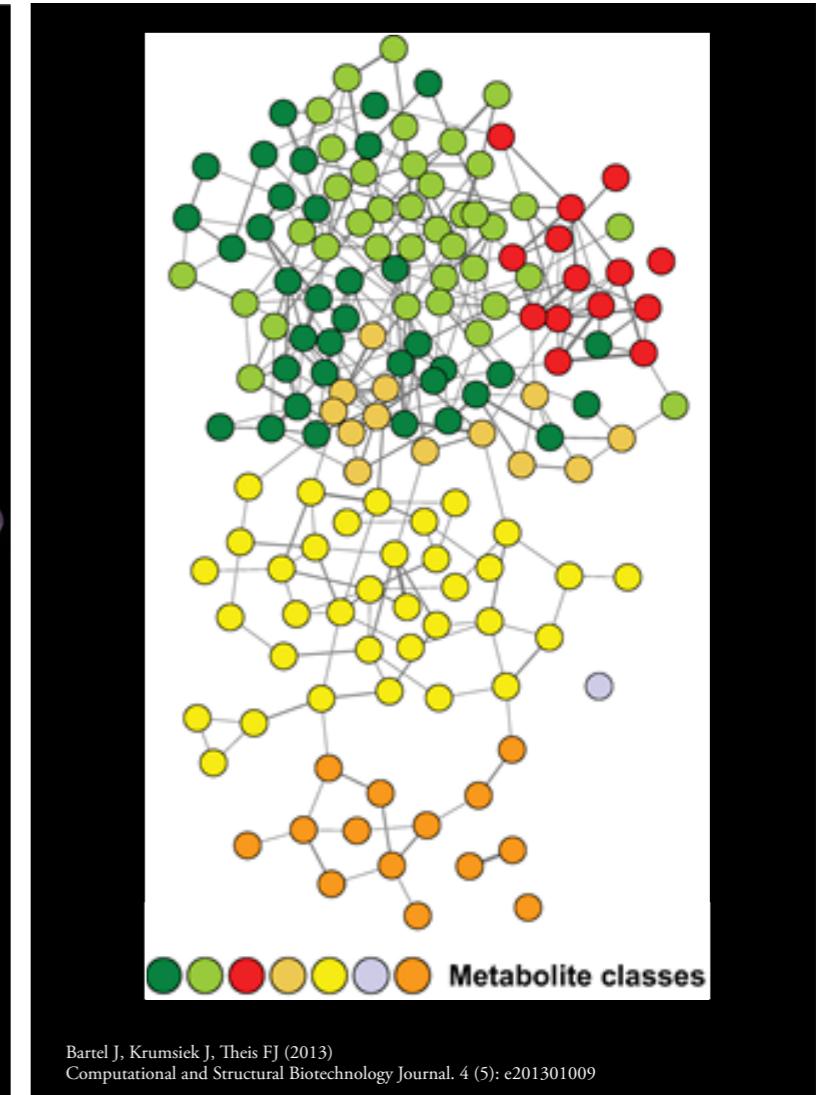
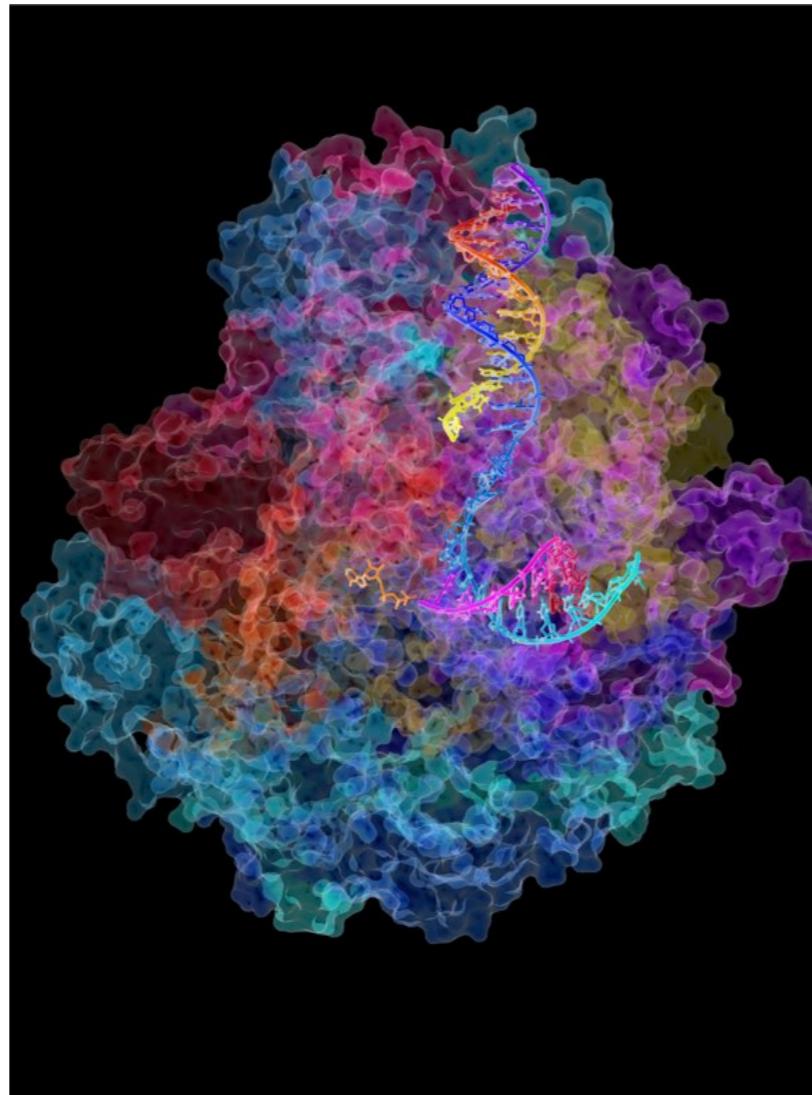
Harvard Chan Bioinformatics Core

NGS Data Analysis Course 2018

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?



shutterstock_97071 Copyright: Sergey

Why Genomics?

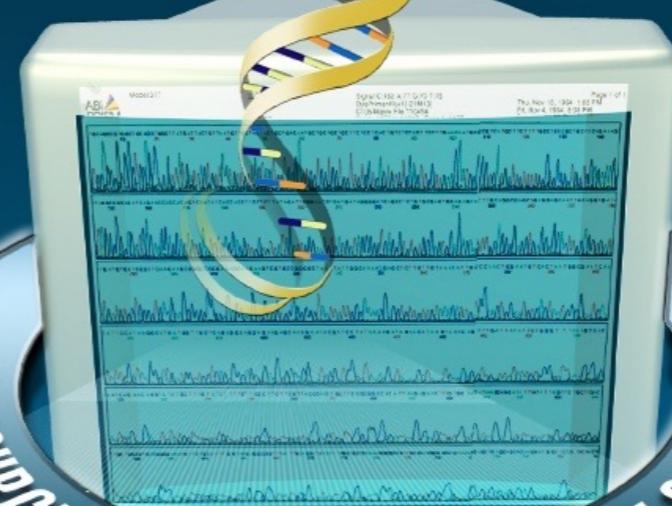
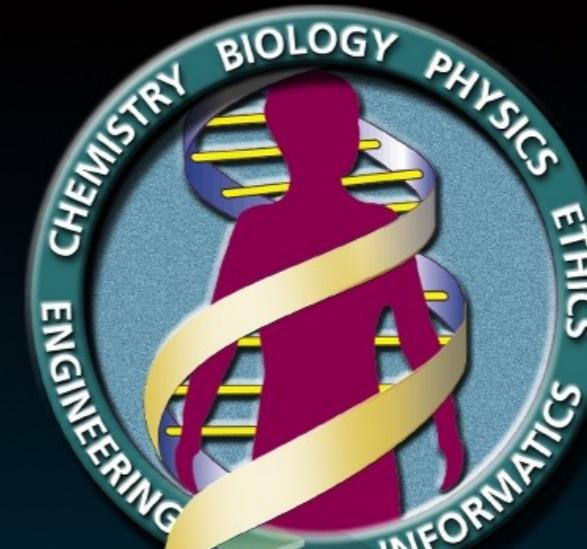


shutterstock_97071 Copyright: Sergey

**Comprehensive
High Throughput
Exploratory**

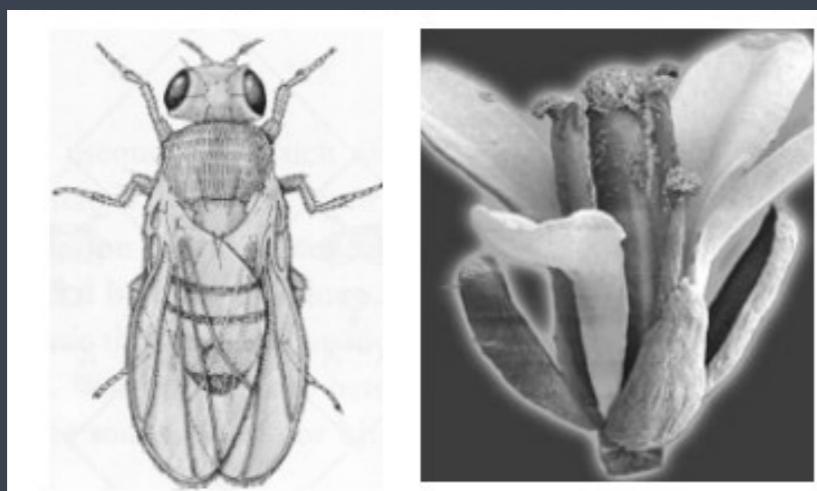
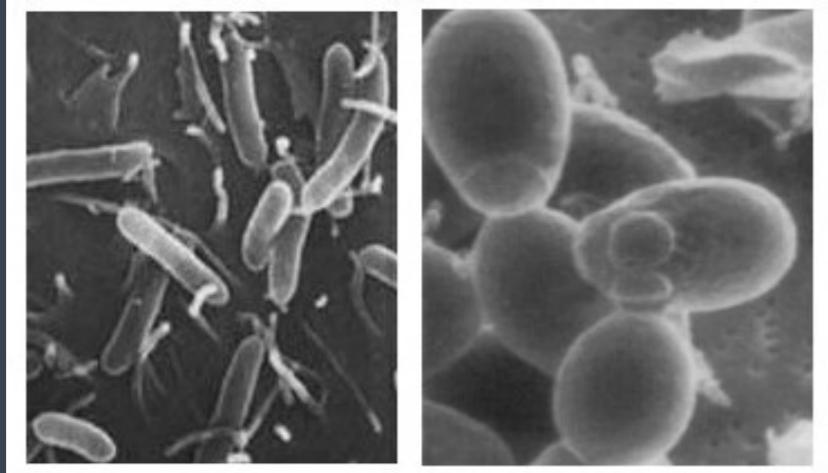
Human Genome Project

1990 - 2003



RESOURCES FOR THE BIOLOGY CENTURY

Pilot Projects



E. coli

S. cerevisiae

C. elegans

D. melanogaster

A. thaliana

Sequencing of the Human Genome

Human Genome Project



(1990)



image credit: [401\(K\) 2012](#)

Mandate to submit all DNA data to public data bank within 24 hours

Celera Genomics



(1998)



image credit: [401\(K\) 2012](#)

No free redistribution or scientific use of the data

Sequencing of the Human Genome

Human Genome Project



(1990)

Celera Genomics



(1998)

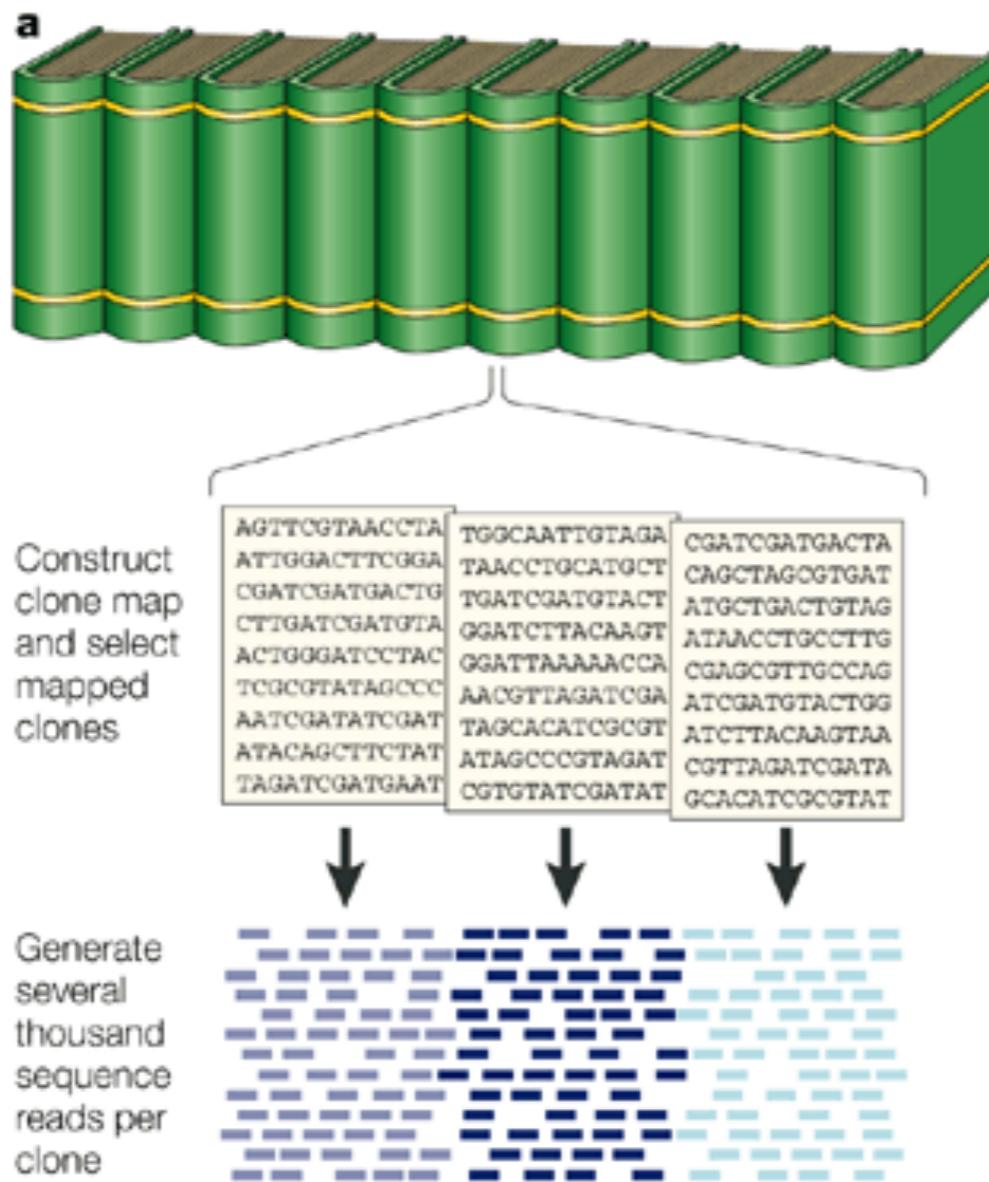


In 2000, President Clinton announced that the human genome could not be patentable, resulting in greater collaboration between the two efforts.

Sequencing of the Human Genome

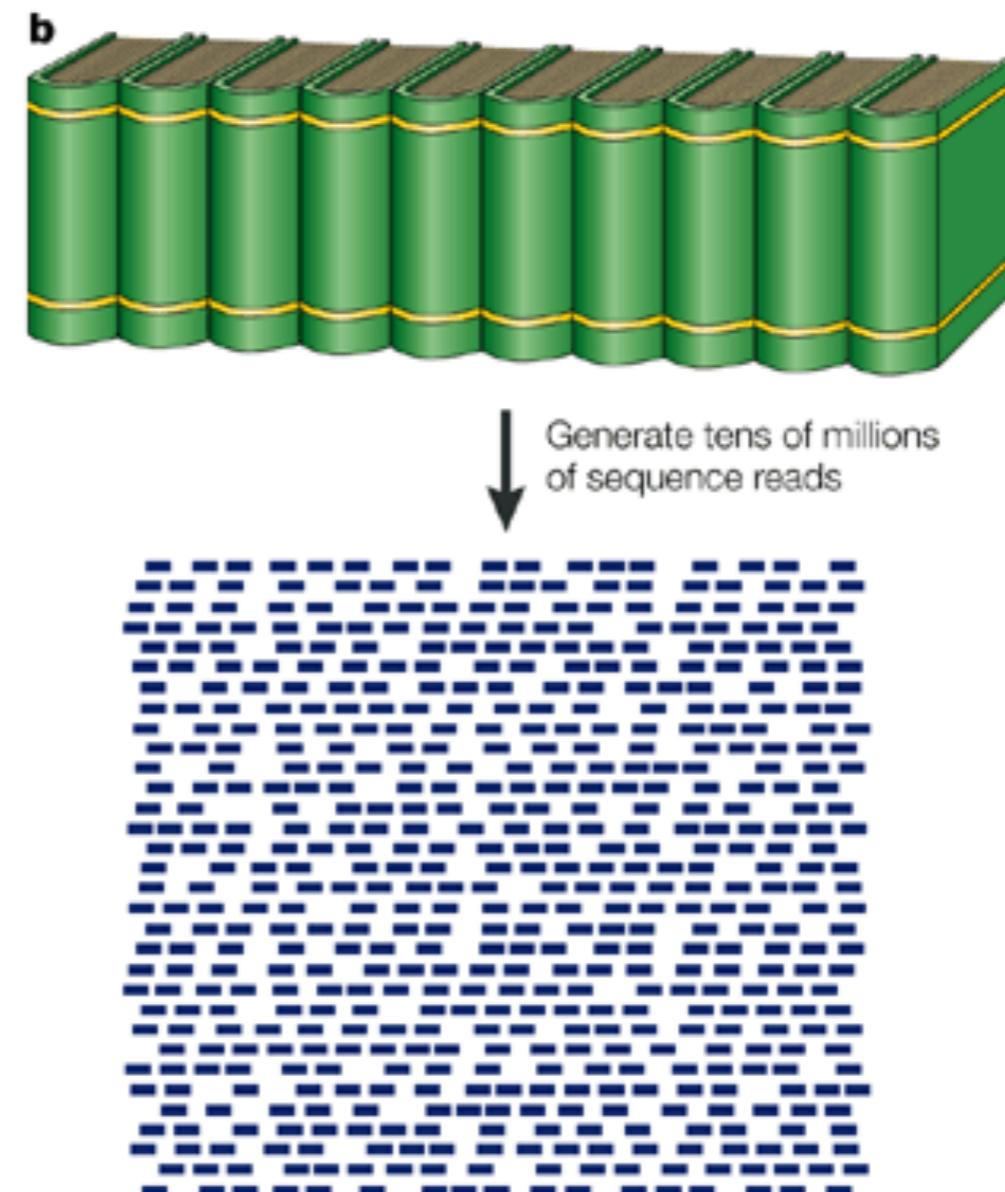
Human Genome Project

Heirarchical Shotgun Sequencing

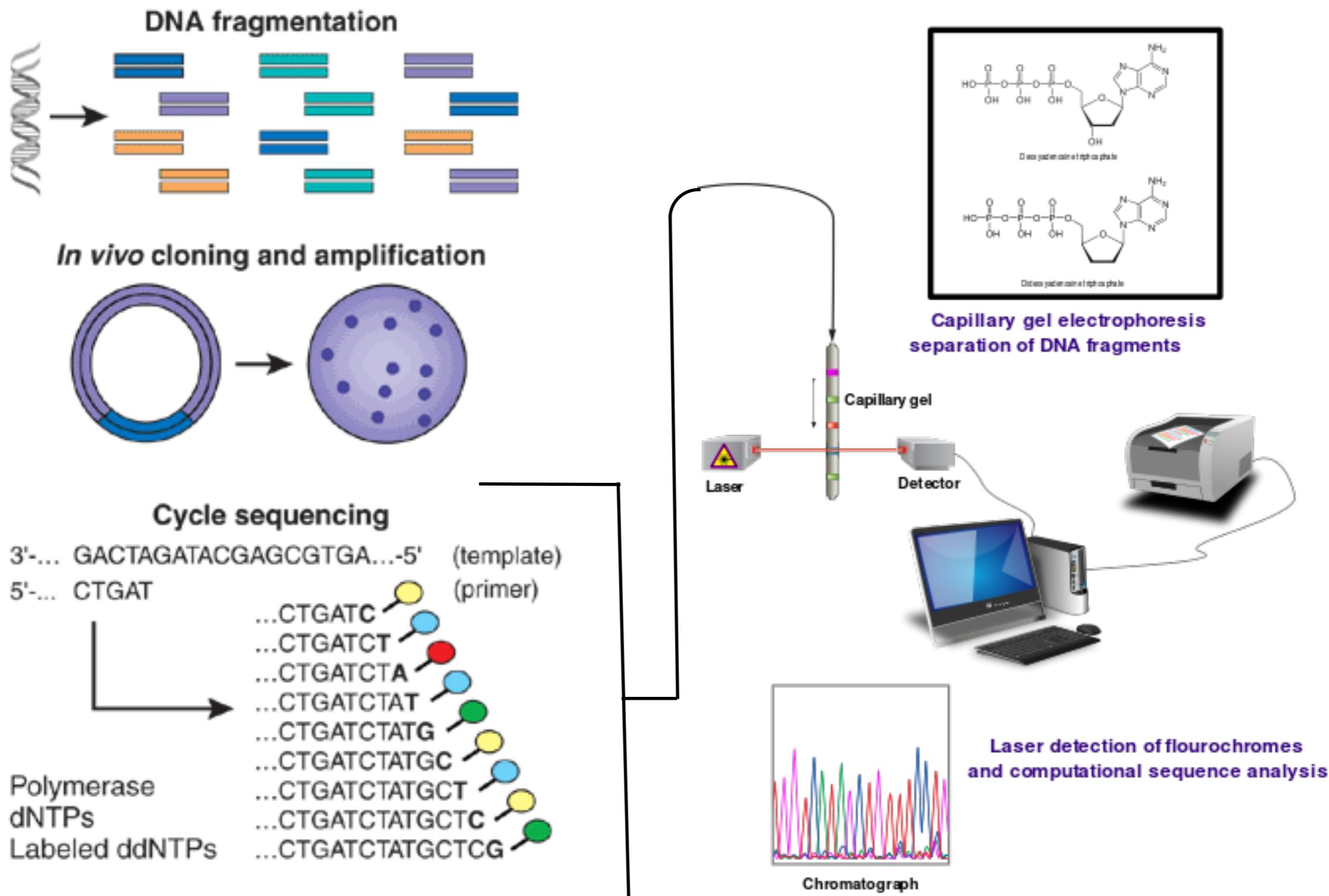


Celera Genomics

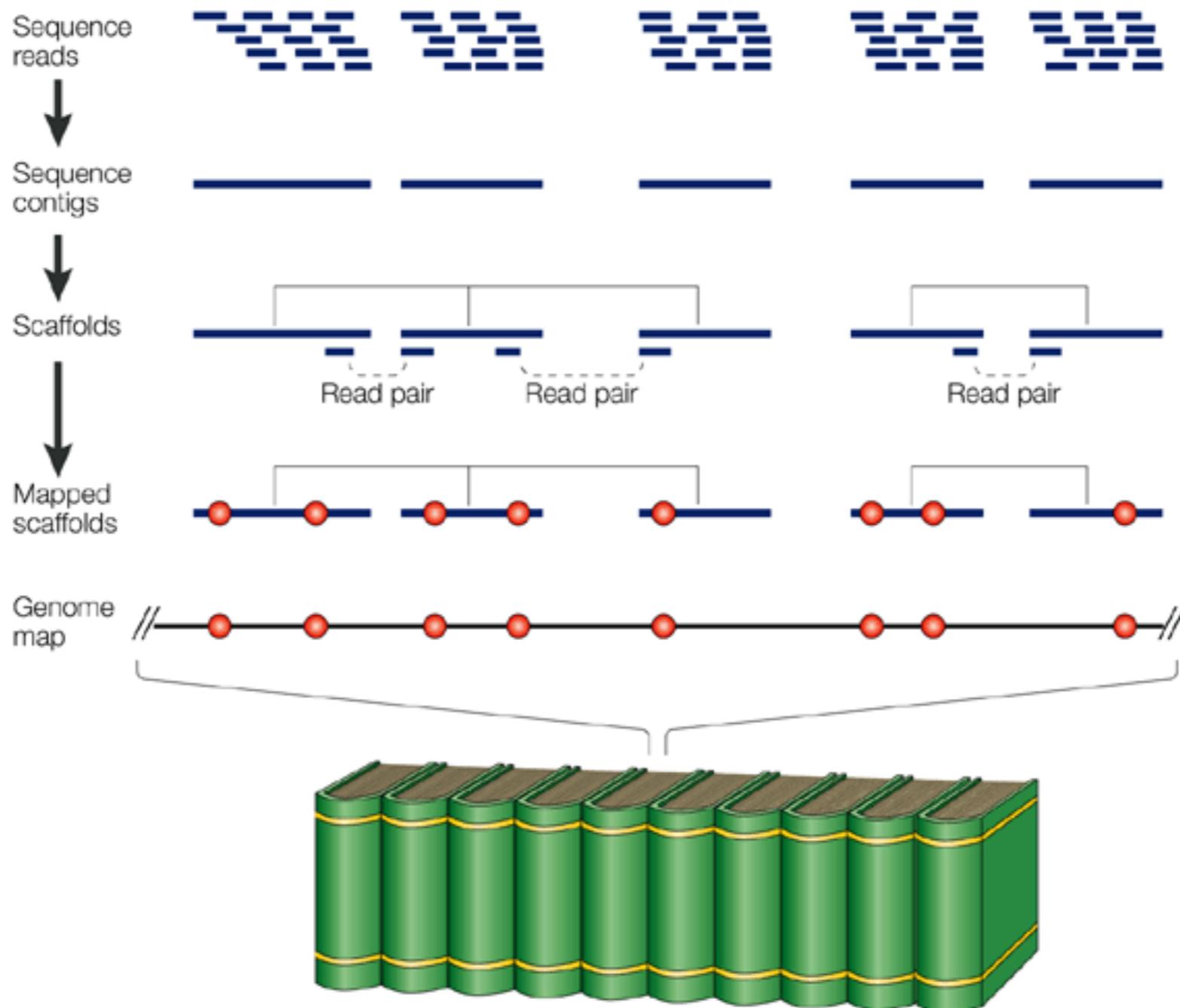
Whole Genome Shotgun Sequencing



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>

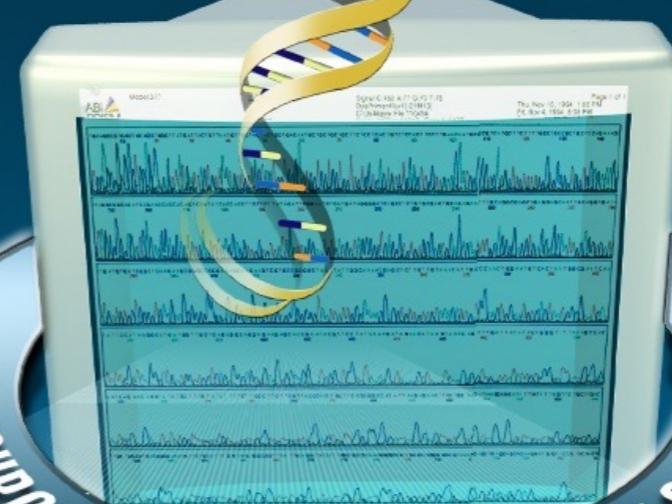
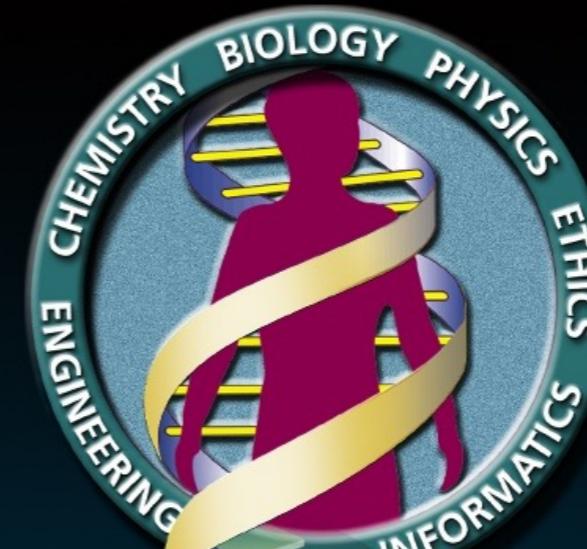
Cycle (Sanger) sequencing generated:

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

Sequence production was rate limiting, not analysis

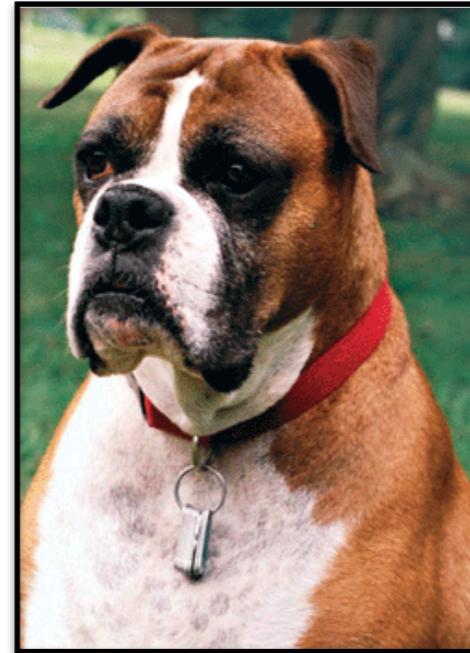
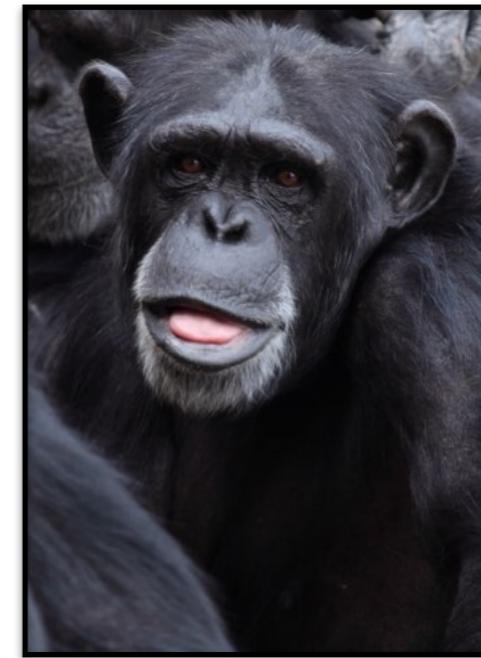
Human Genome Project

1990 - 2003



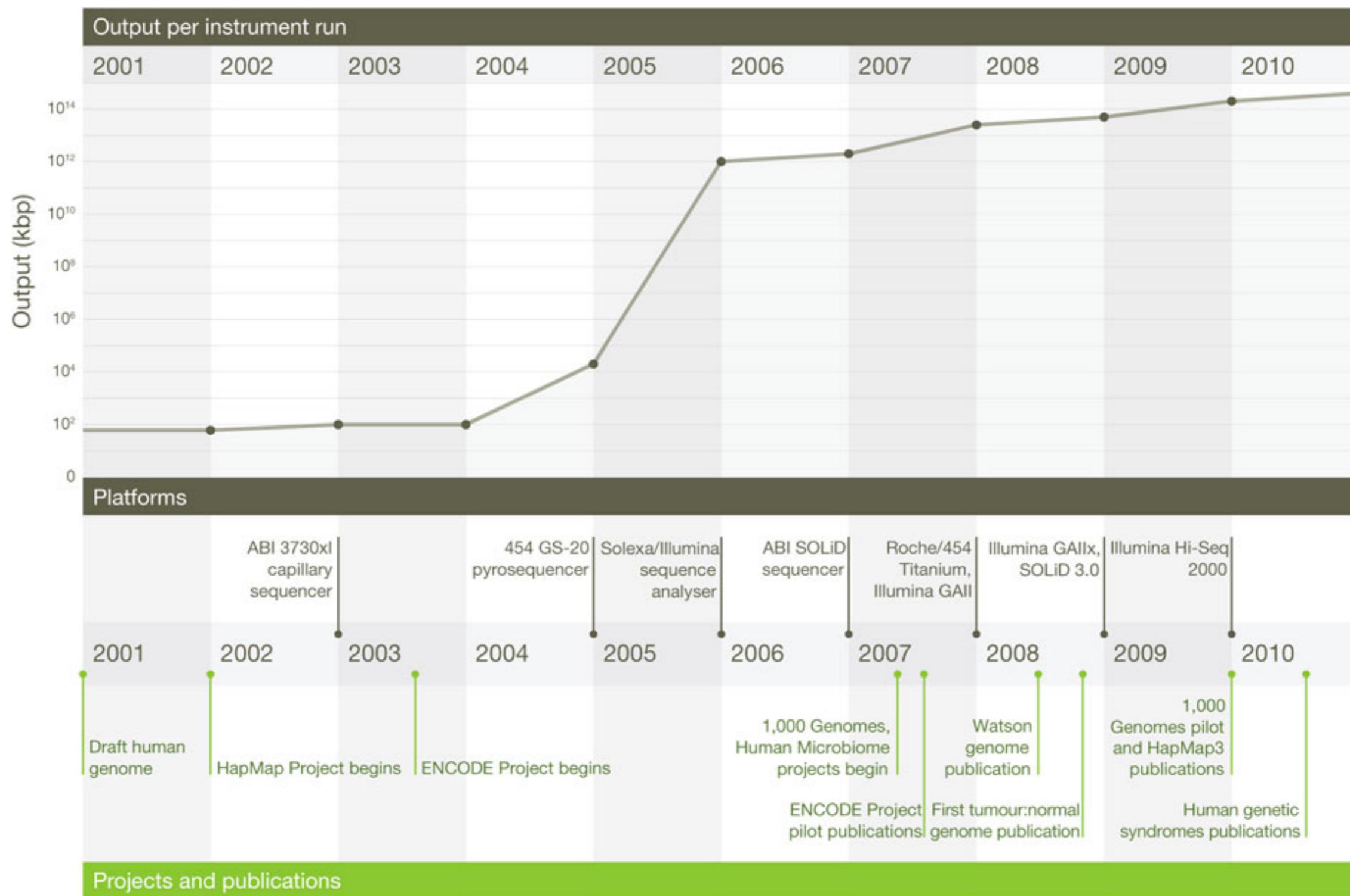
RESOURCES FOR THE BIOLOGY CENTURY

Comparative Genomics Research



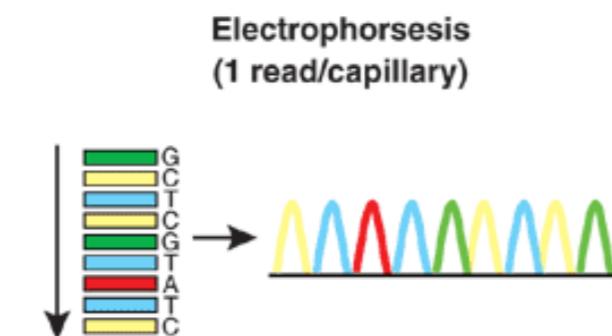
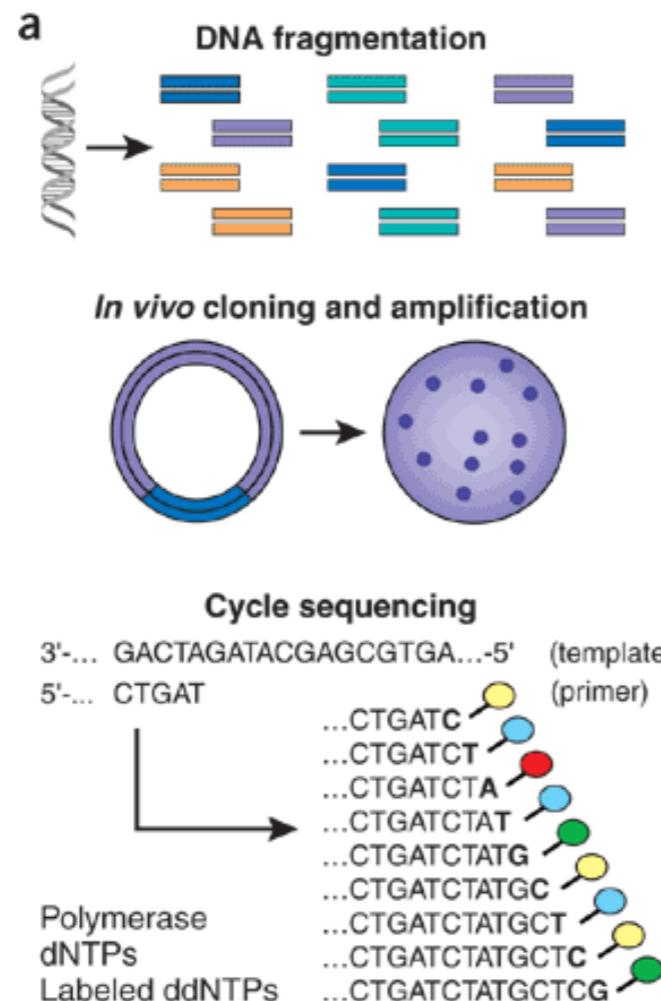
Highly conserved regions of DNA likely to be functional

Advancements in Sequencing Technology

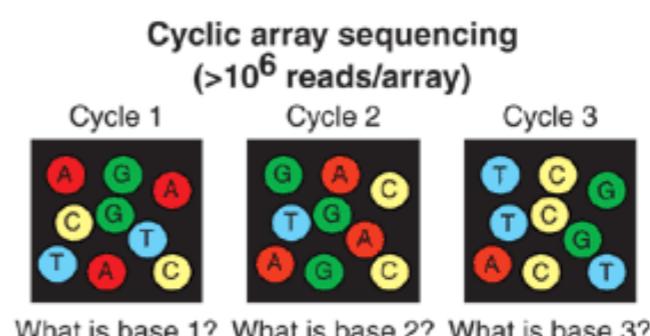
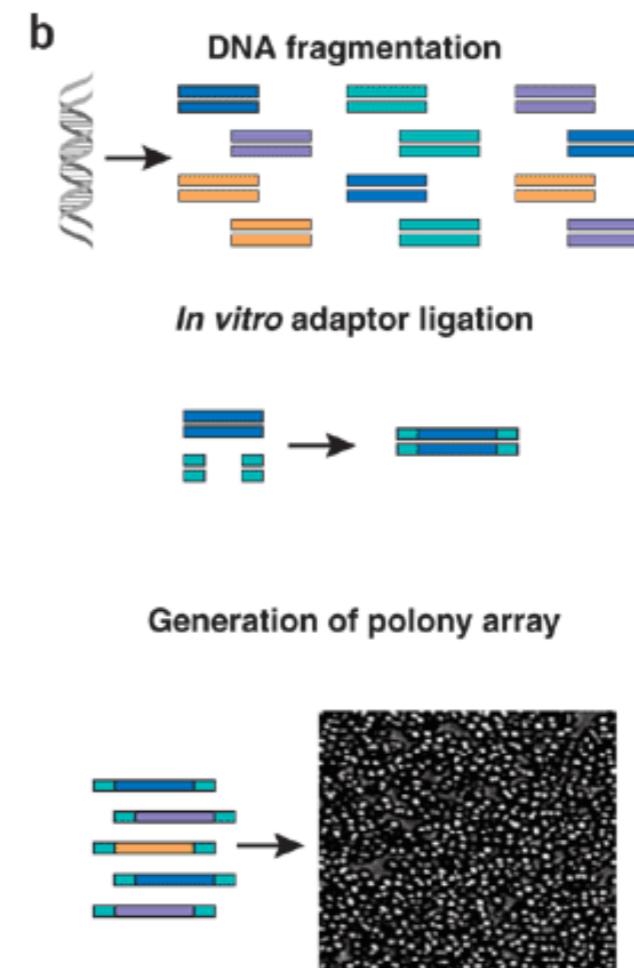


NGS Technologies

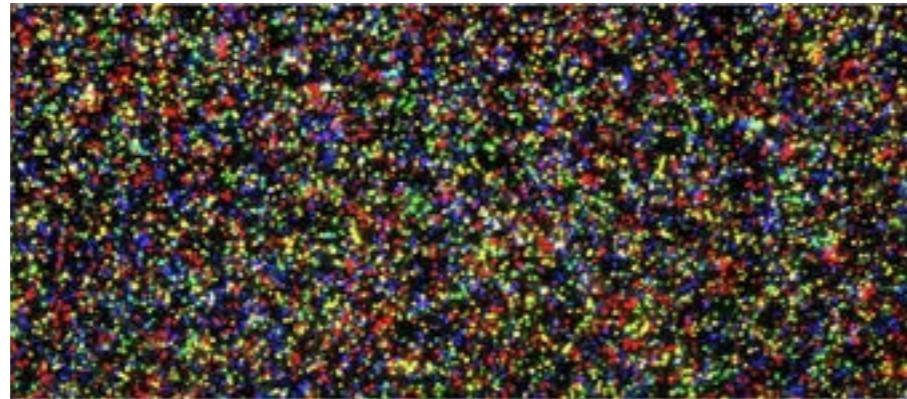
Cycle Sequencing



NGS Sequencing



NGS Technologies



Generate:

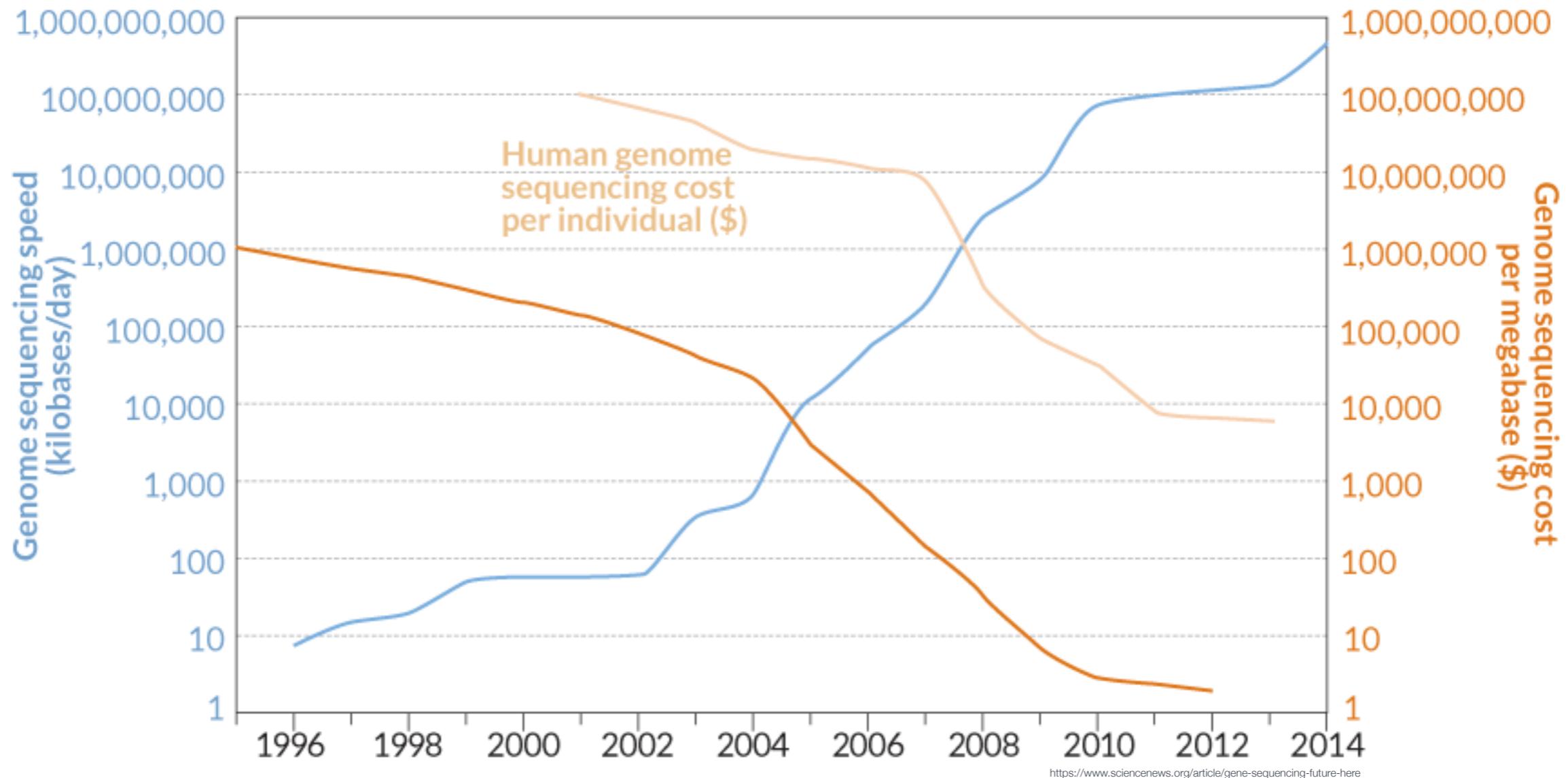
- Illumina HiSeq2500: 150 bp / read (250 bp / read Rapid-Run Mode)
- > **100 billion bp** / day

Bioinformatics support required to handle:

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

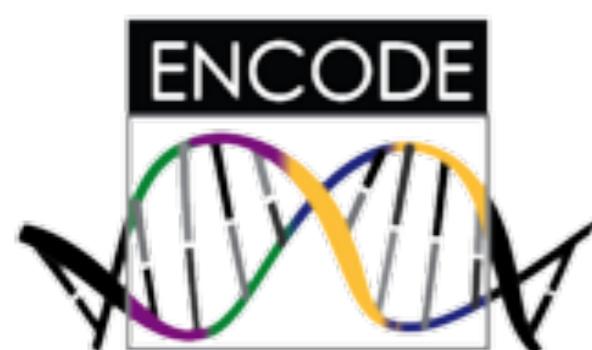
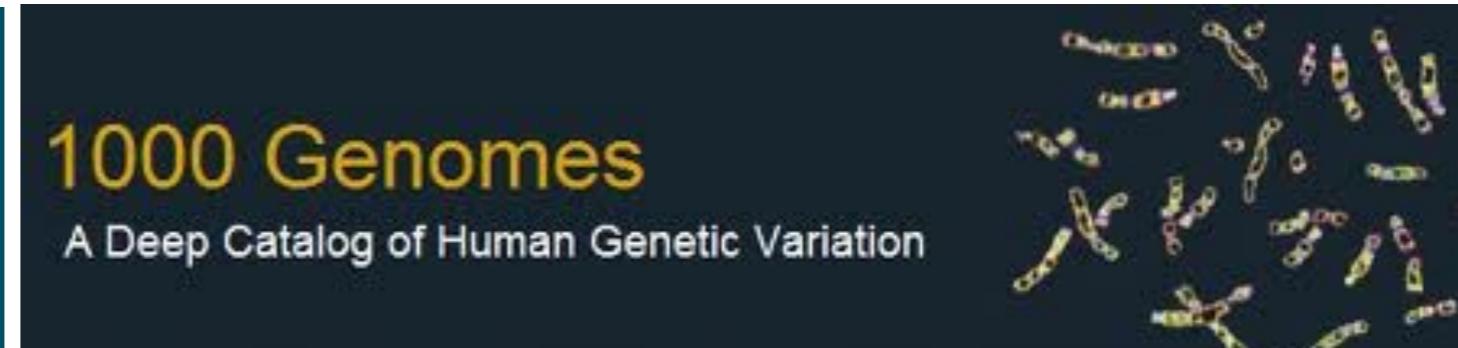
Sequence analysis rate limiting, not production

NGS Technology Accessibility



Increased scale and lower cost increases access to sequencing technologies

The Genomic Era: Collaborative Projects



FANTOM
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME



The Genomic Era: Individual Projects



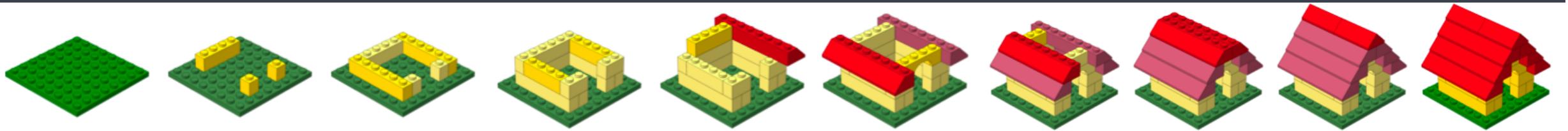
- Only a few experiments = vast amounts of data
- Data generation straightforward, but analysis requires bioinformatics expertise

Bioinformatics in the Genomic Era

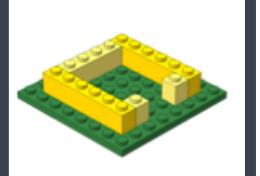


Alliances between experimentalists and computational biologists

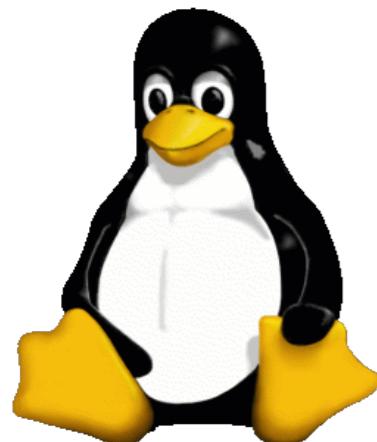
Bioinformatics Toolkit



Programming languages



Programming languages are critical in genomic analyses:

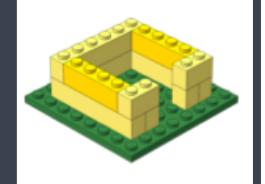


- ▶ **Bash:** command line language used for interacting with *Linux / Unix-based* operating systems.
 - Attaining sequencing data
 - Accessing computing resources and analysis tools
 - Basic data manipulation, creating scripts for running tools

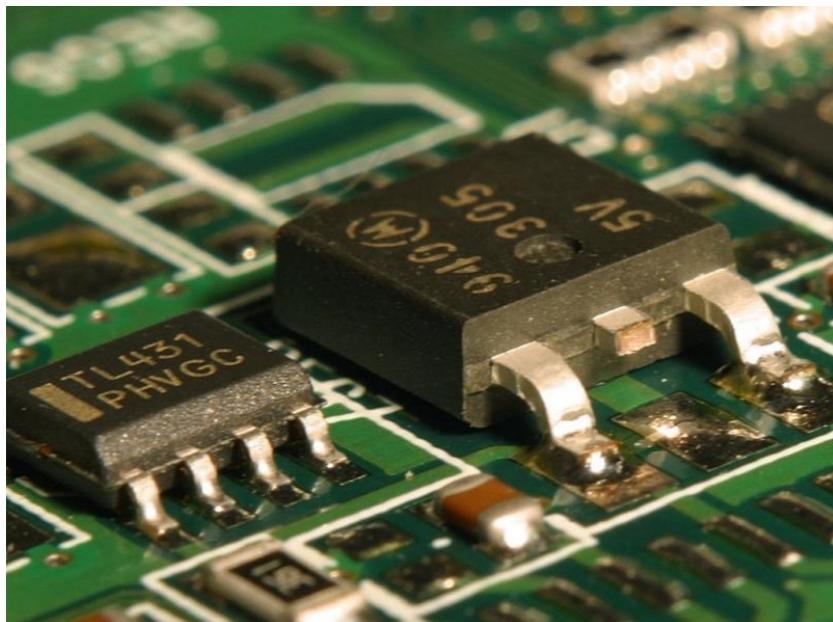
- ▶ **R:** a programming language and environment for statistical computing and graphics
 - Wrangling data, performing statistical analyses
 - Creating figures and plots



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

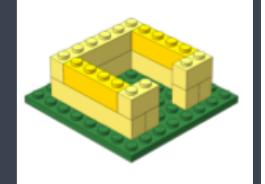
Memory/RAM (Random Access Memory)

- Large datasets ≈ lots of RAM to perform analysis

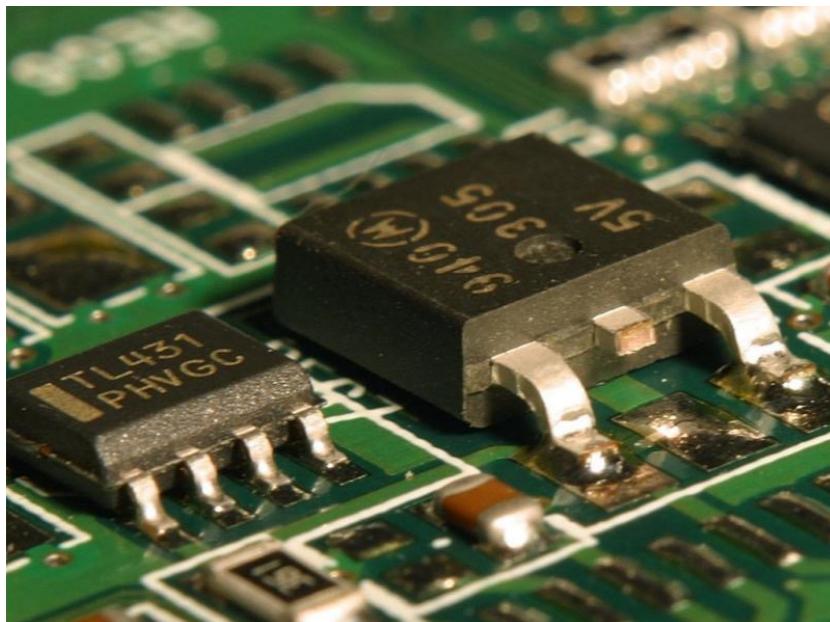
CPU (Central Processing Unit)

- 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)

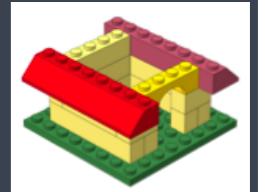
- Large datasets ≈ lots of RAM to perform analysis

CPU (Central Processing Unit)

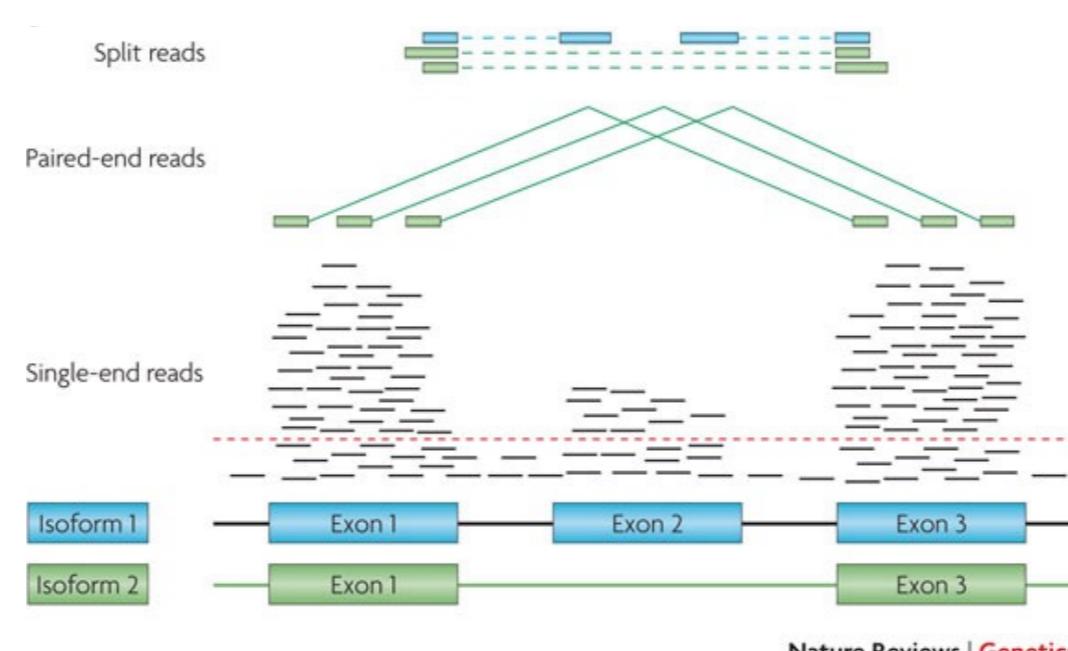
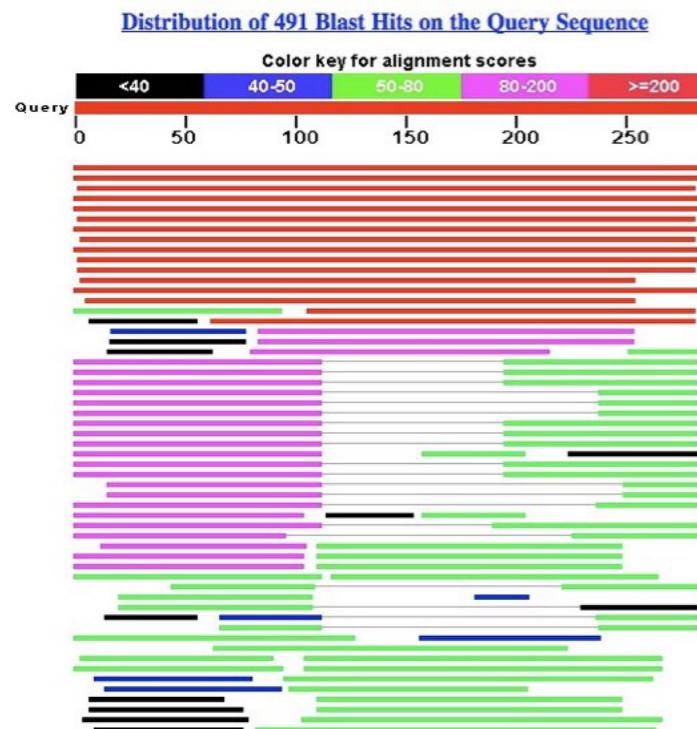
- Large datasets ≈ lots of time to perform analysis

Solution: Amazon Cloud or computing cluster

NGS Analysis Tools and Workflows

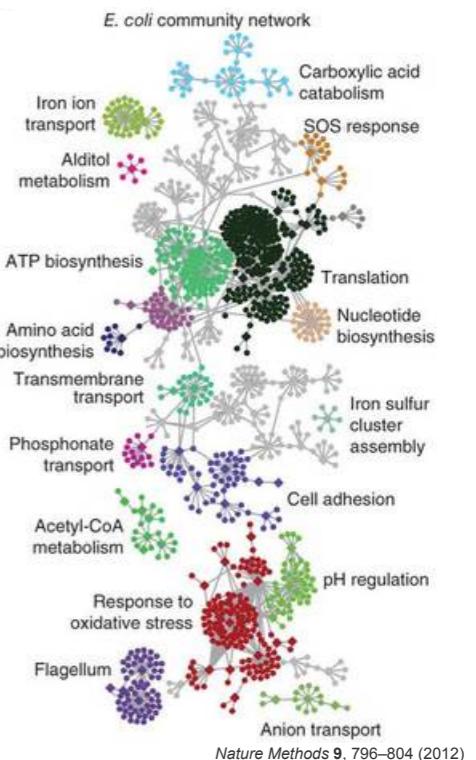


Large genomics datasets require software (tools) to perform each of the steps in an NGS analysis workflow.



Nature Reviews | Genetics

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



Genome Databases



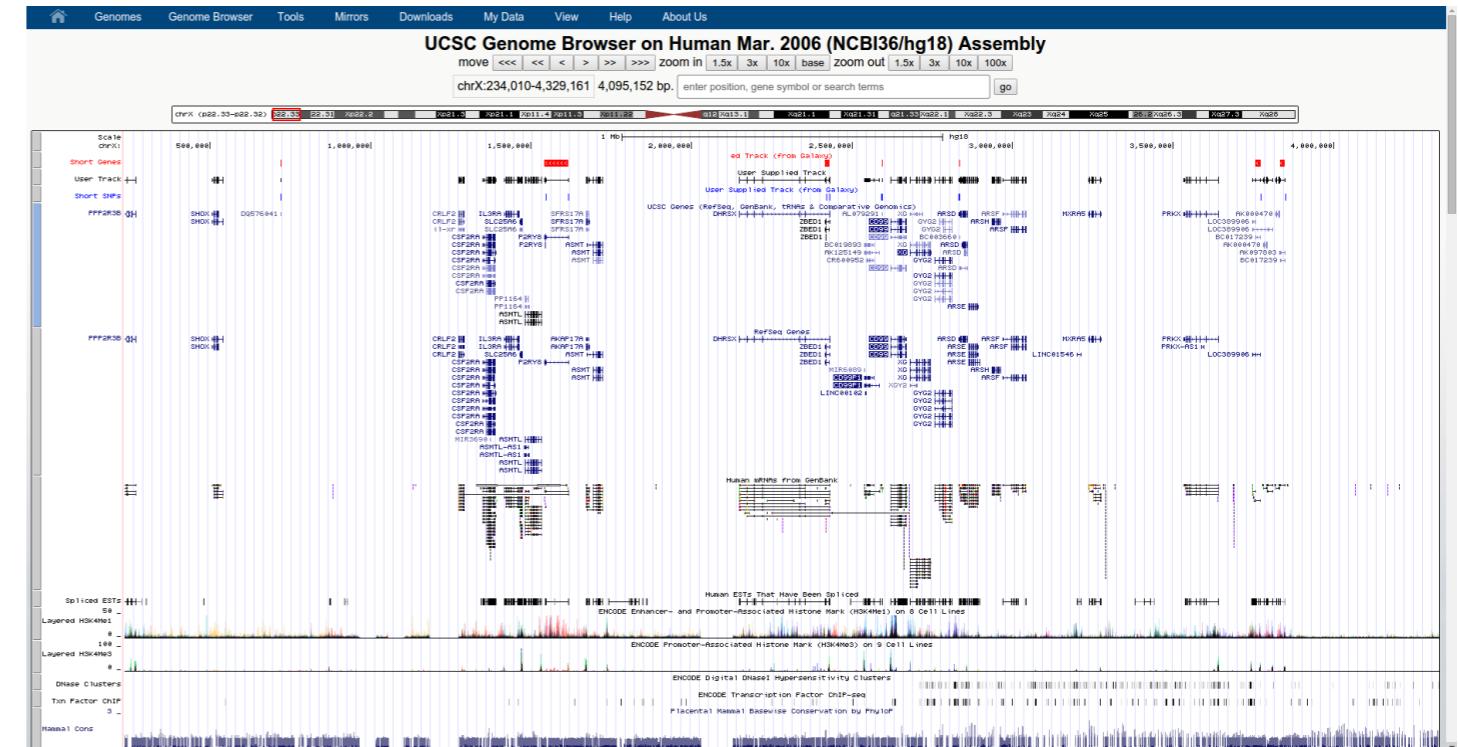
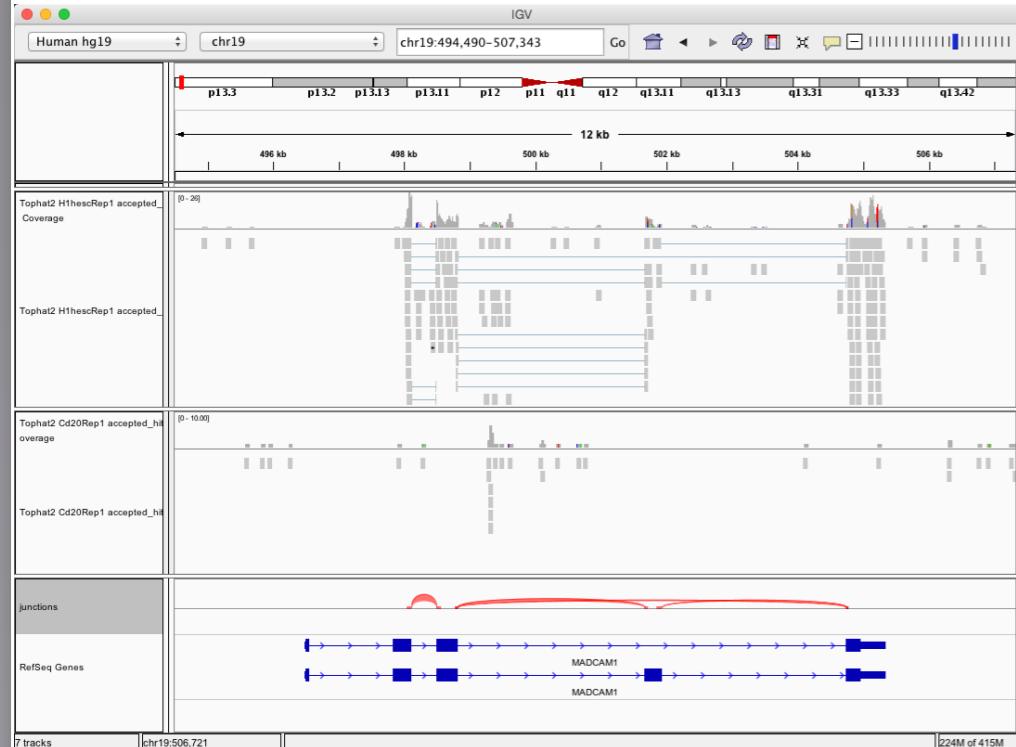
Genome databases contain publicly available, searchable, and up-to-date genome data, including reference sequences and gene annotations



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

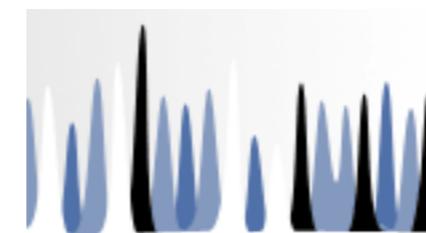


Visualize genomic data from:

- **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

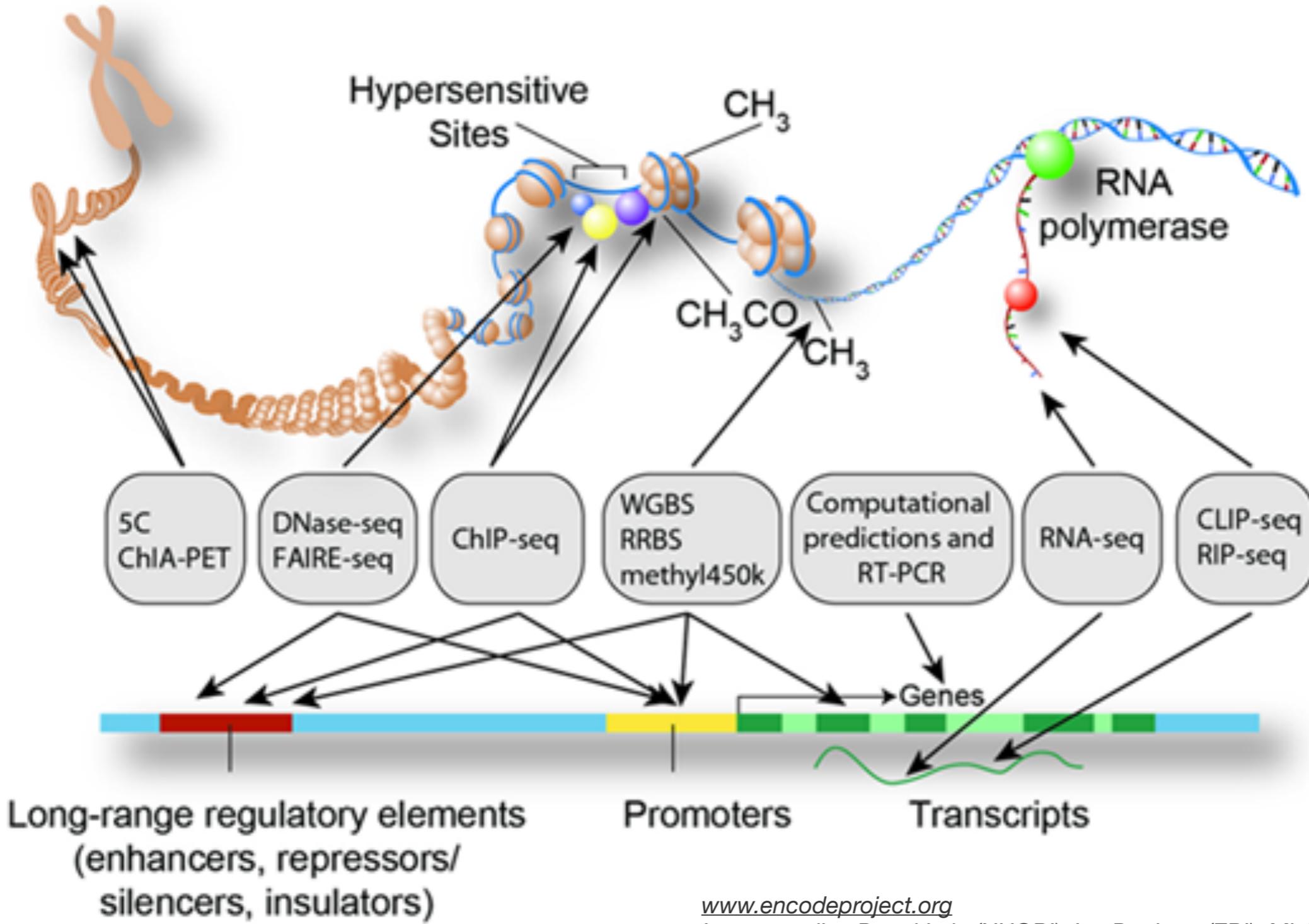


SEQanswers
the next generation sequencing community



NGS Applications

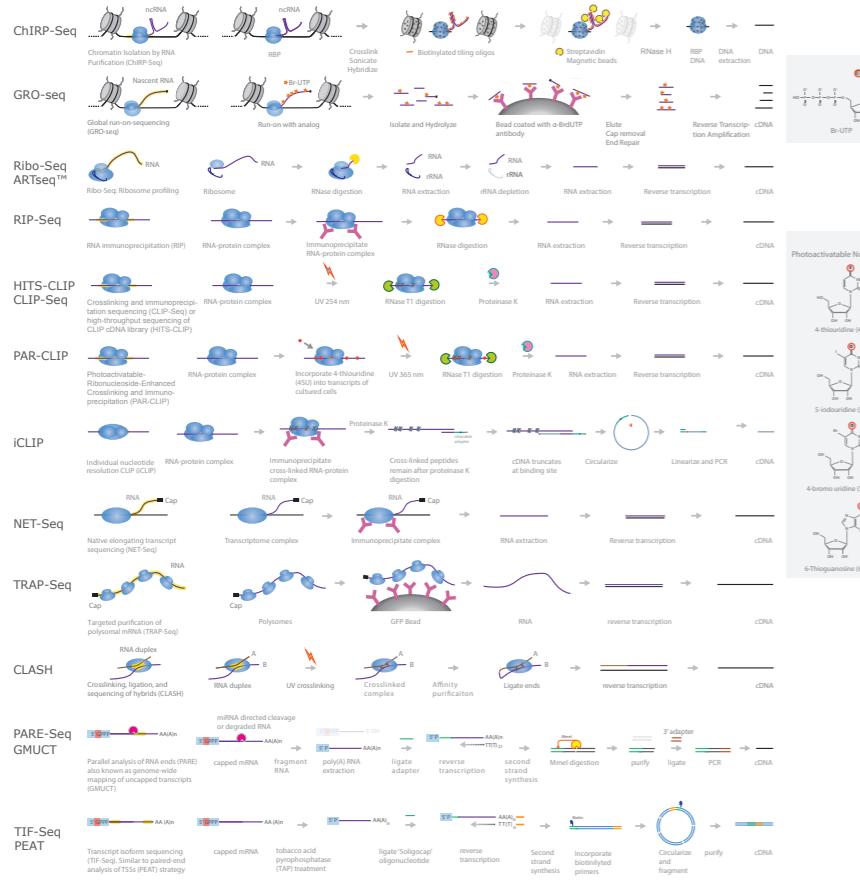
NGS Applications



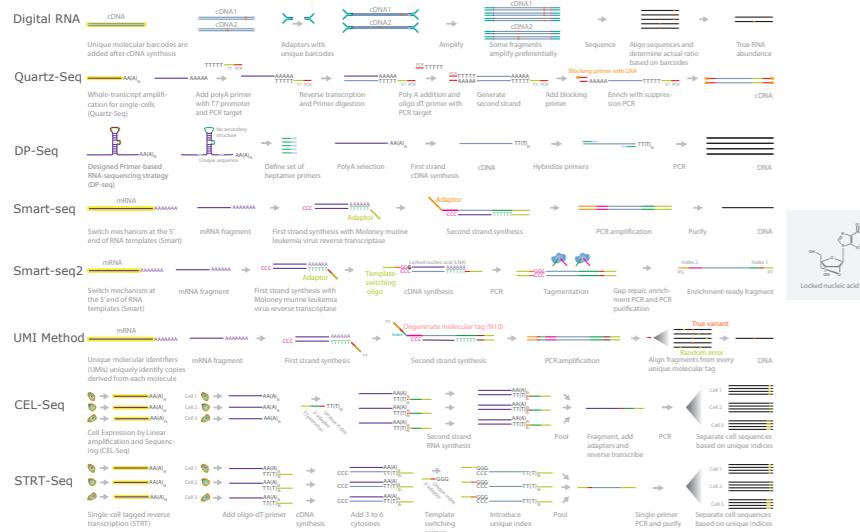
www.encodeproject.org

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

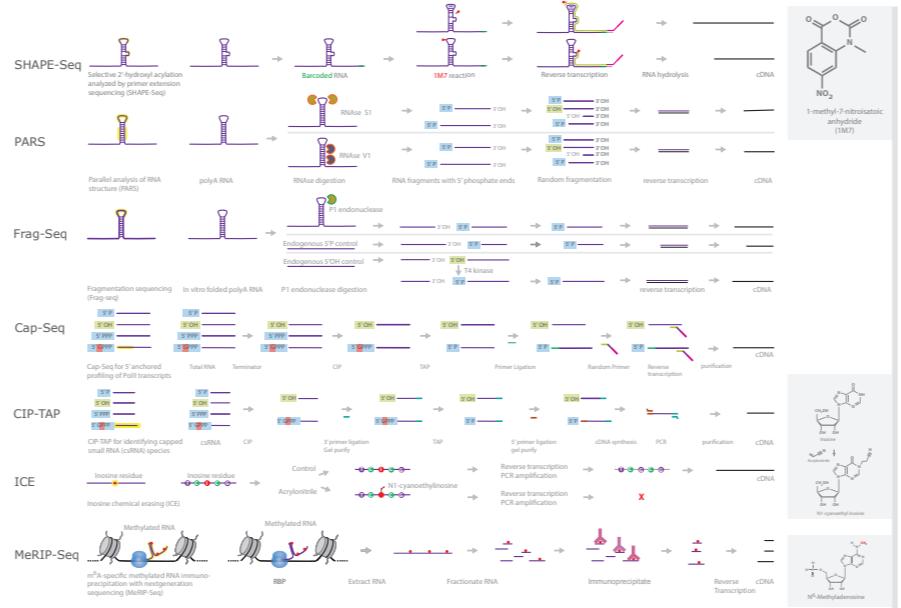
RNA Transcription



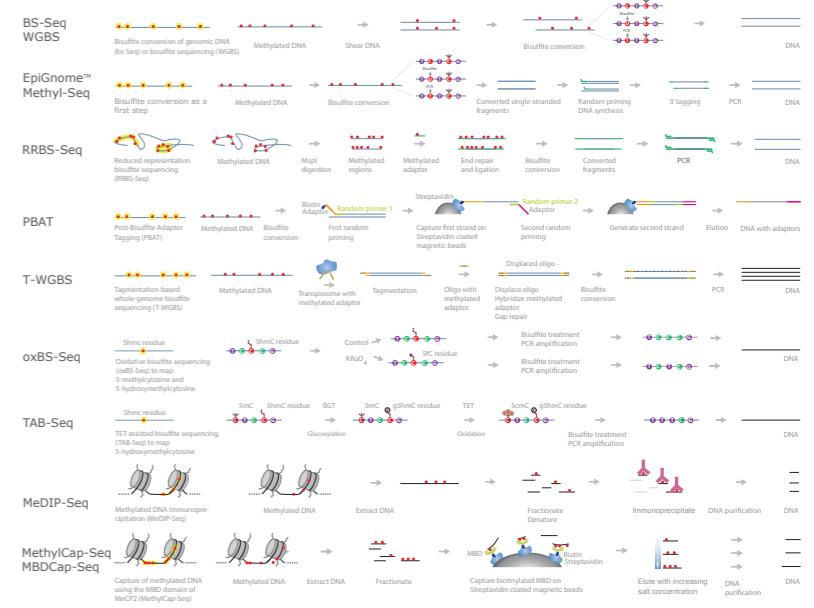
RNA Low-Level Detection



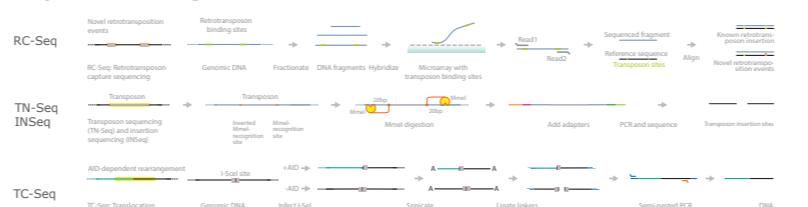
RNA Structure



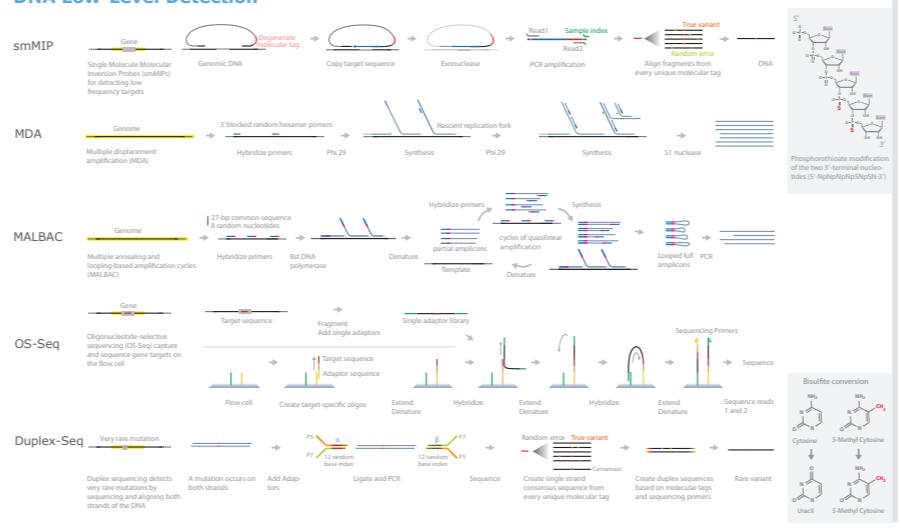
Methylation



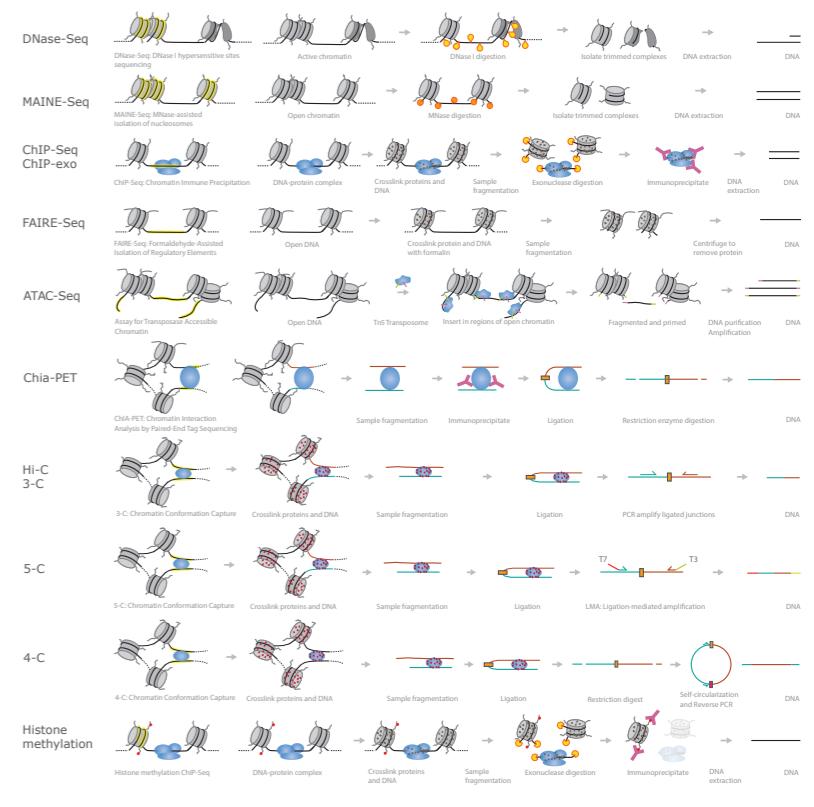
Sequence Rearrangements



DNA Low-Level Detection



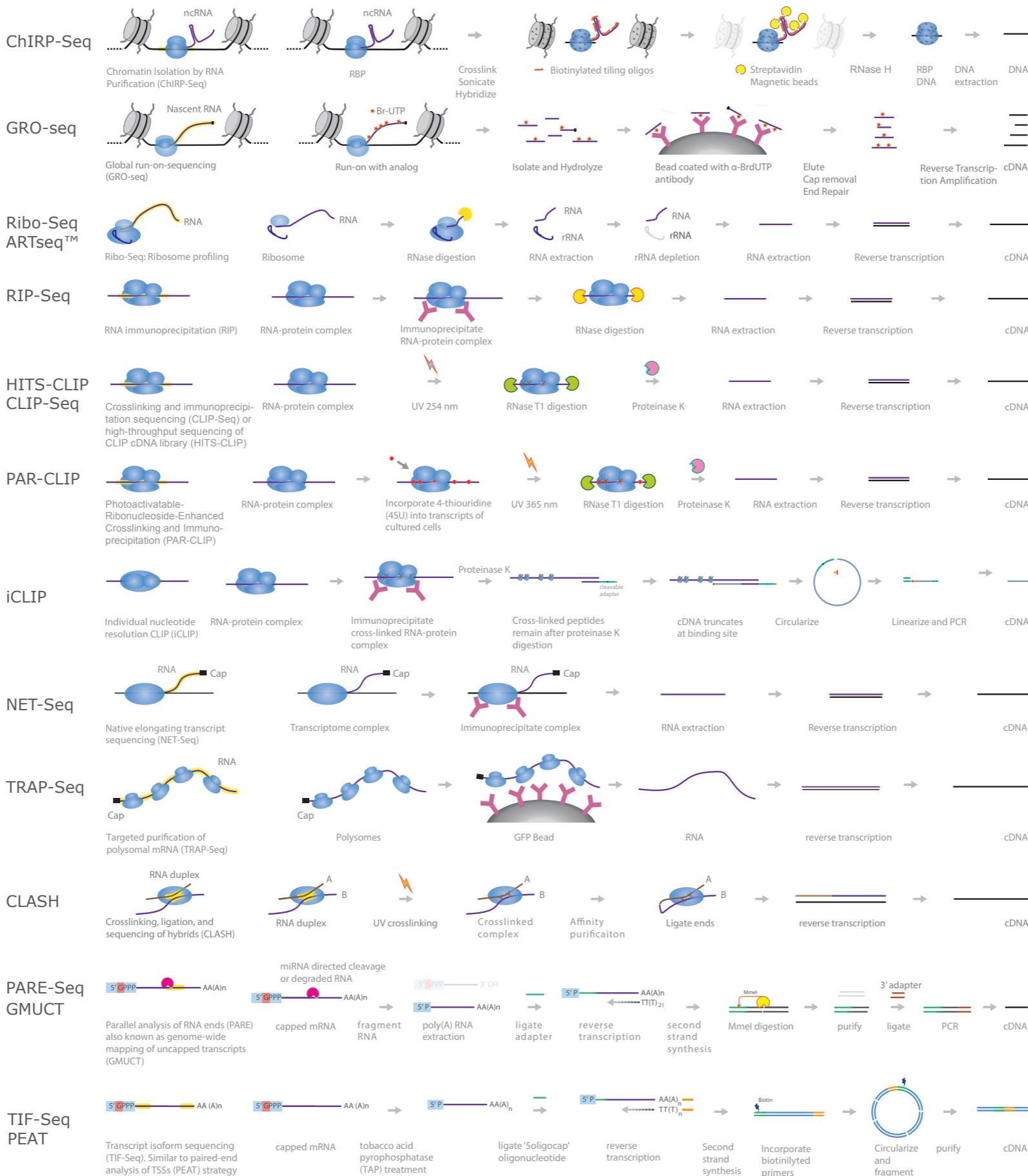
DNA-Protein Interactions



Key

Yellow double-headed arrow indicates that two steps are optional.

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.

