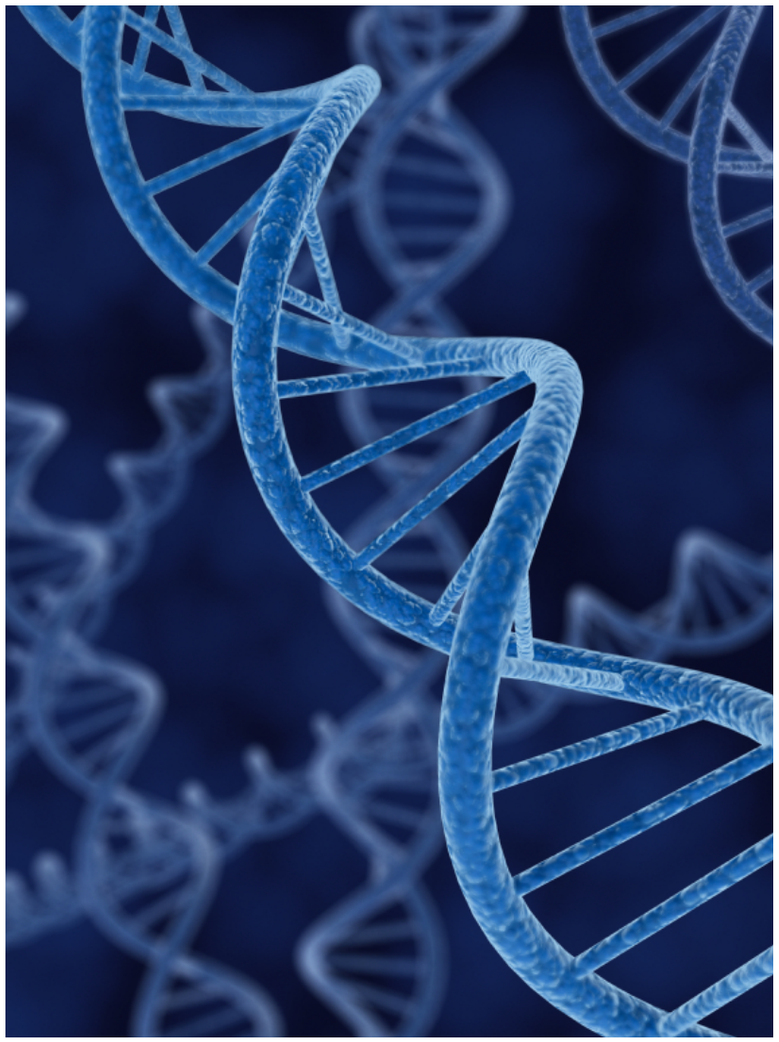
Overview of Bioinformatics

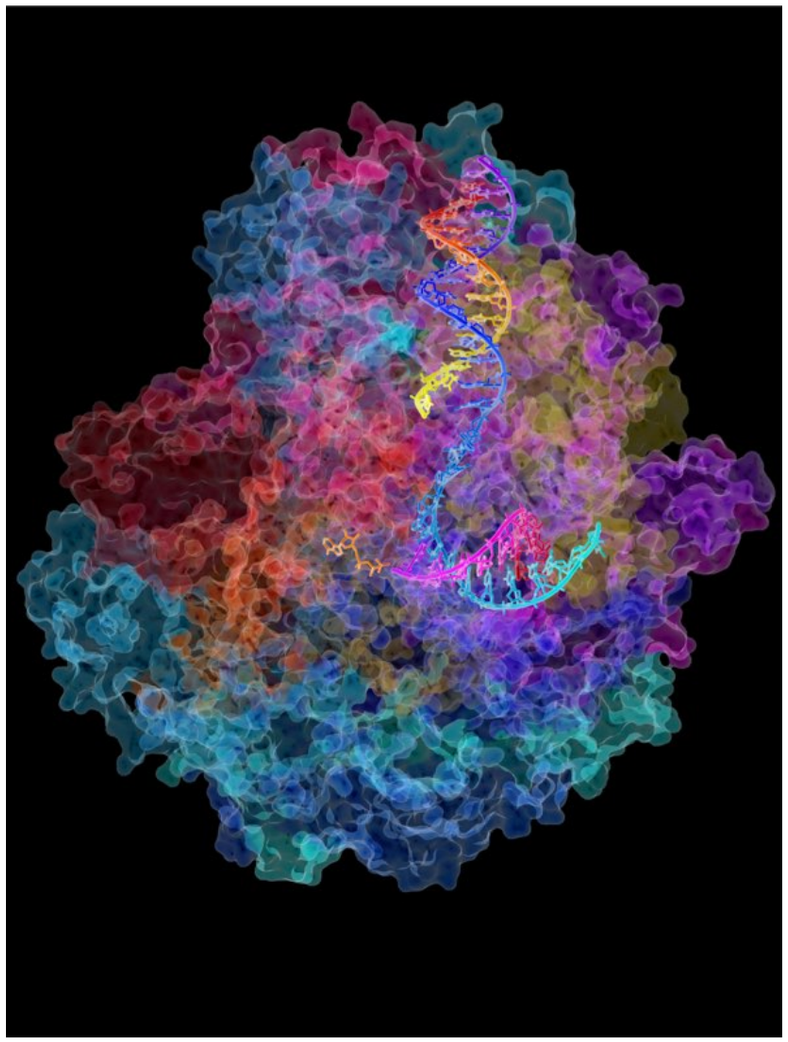
Harvard Chan Bioinformatics Core

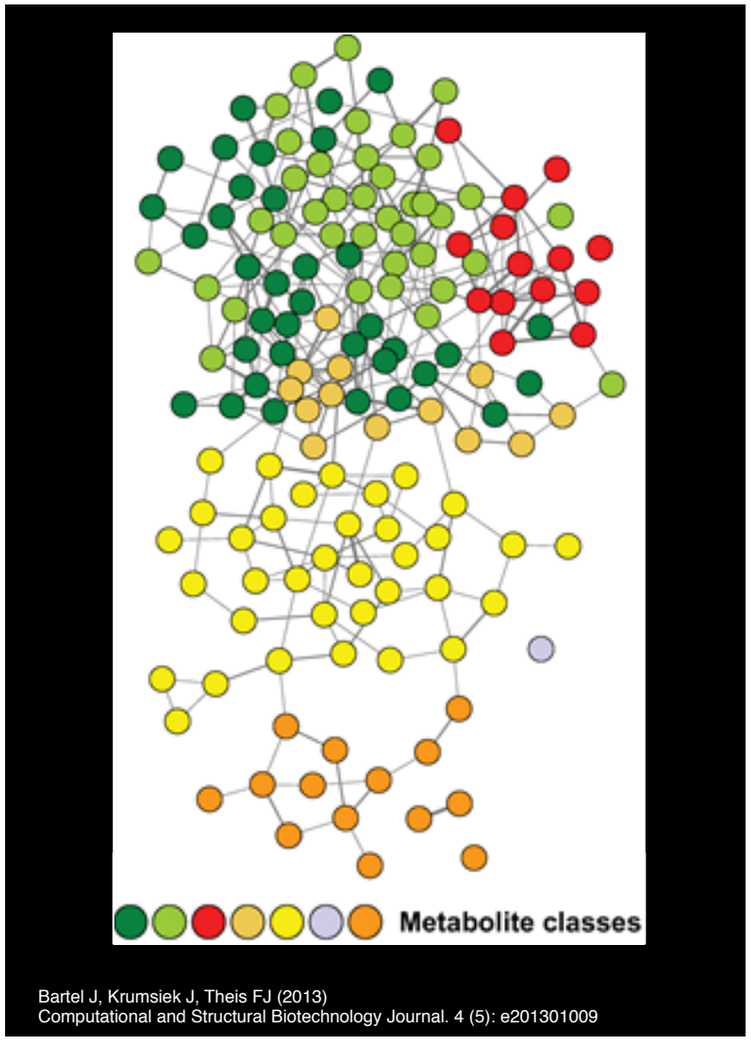
NGS Data Analysis Course 2016

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

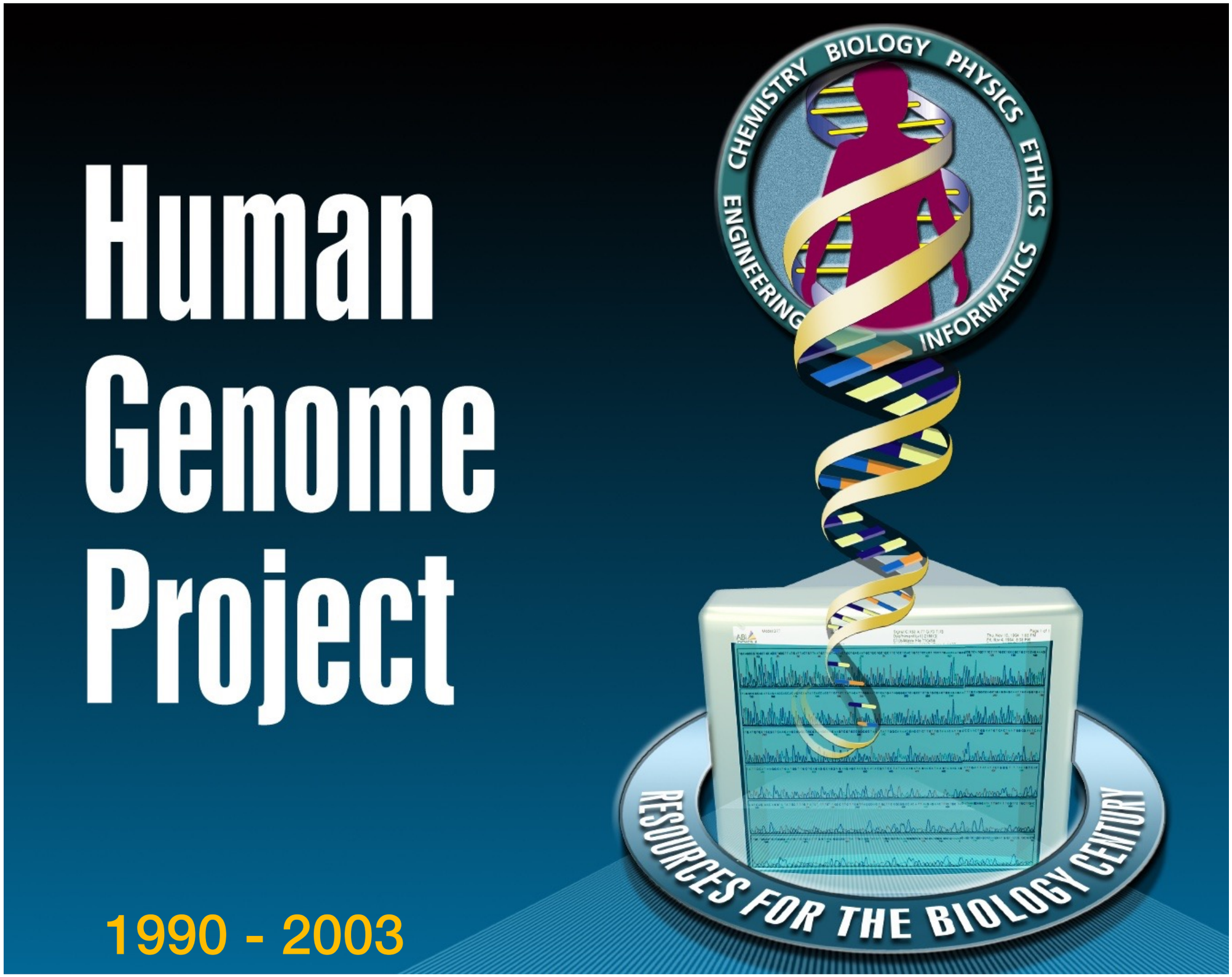
Why Genomics?

Why Genomics?

High Throughput

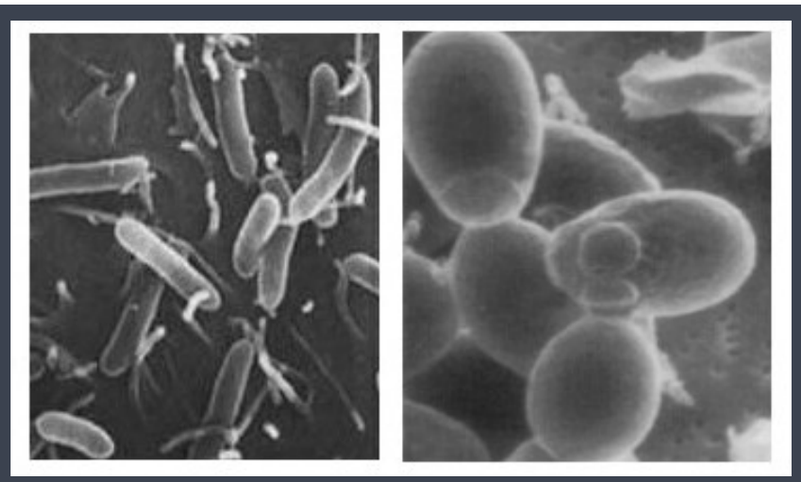
Comprehensive

Exploratory



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

Pilot Projects



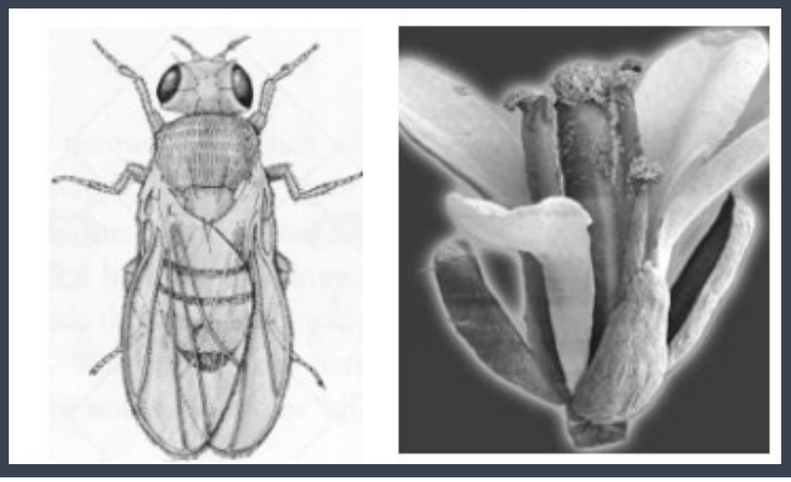


image credit: http://www.genomesize.com/rgregory/genome.html

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

CELERA
an Applera Corporation Business

(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

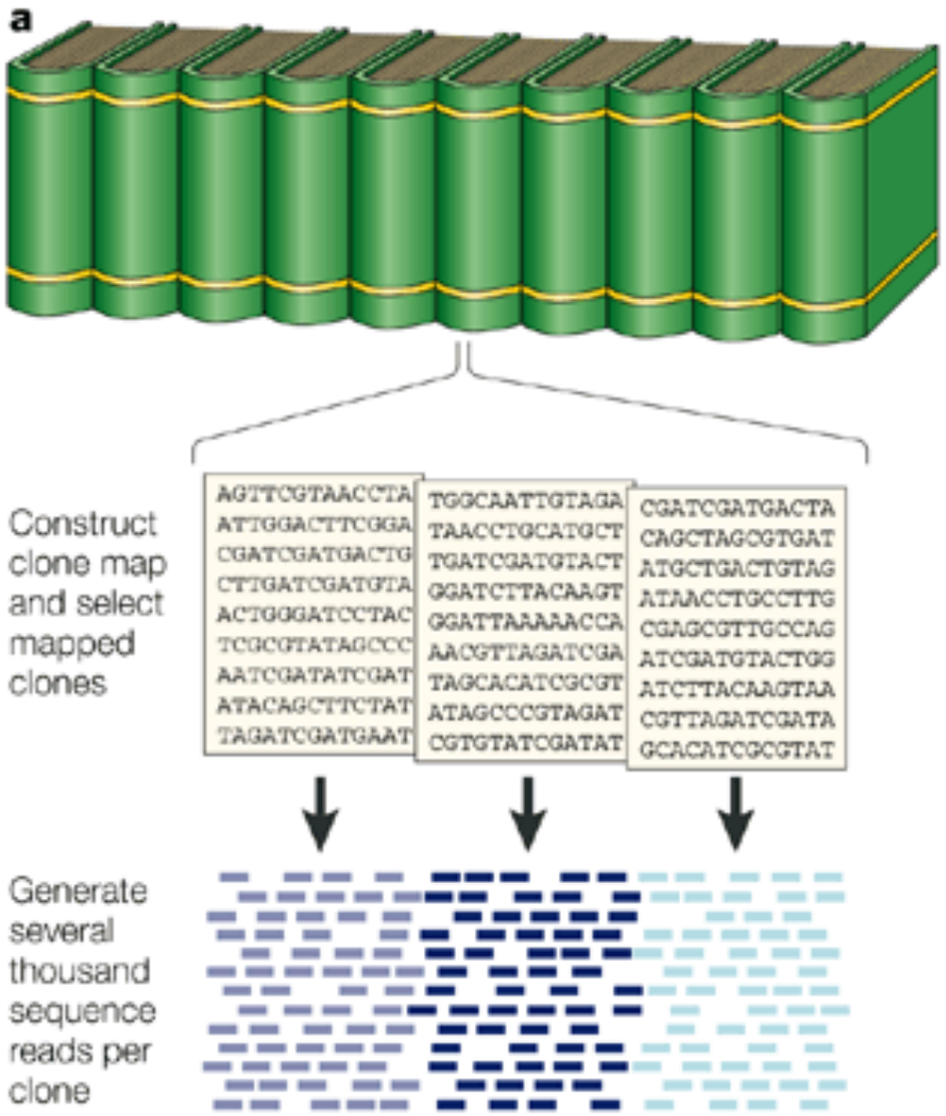
(1998)

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

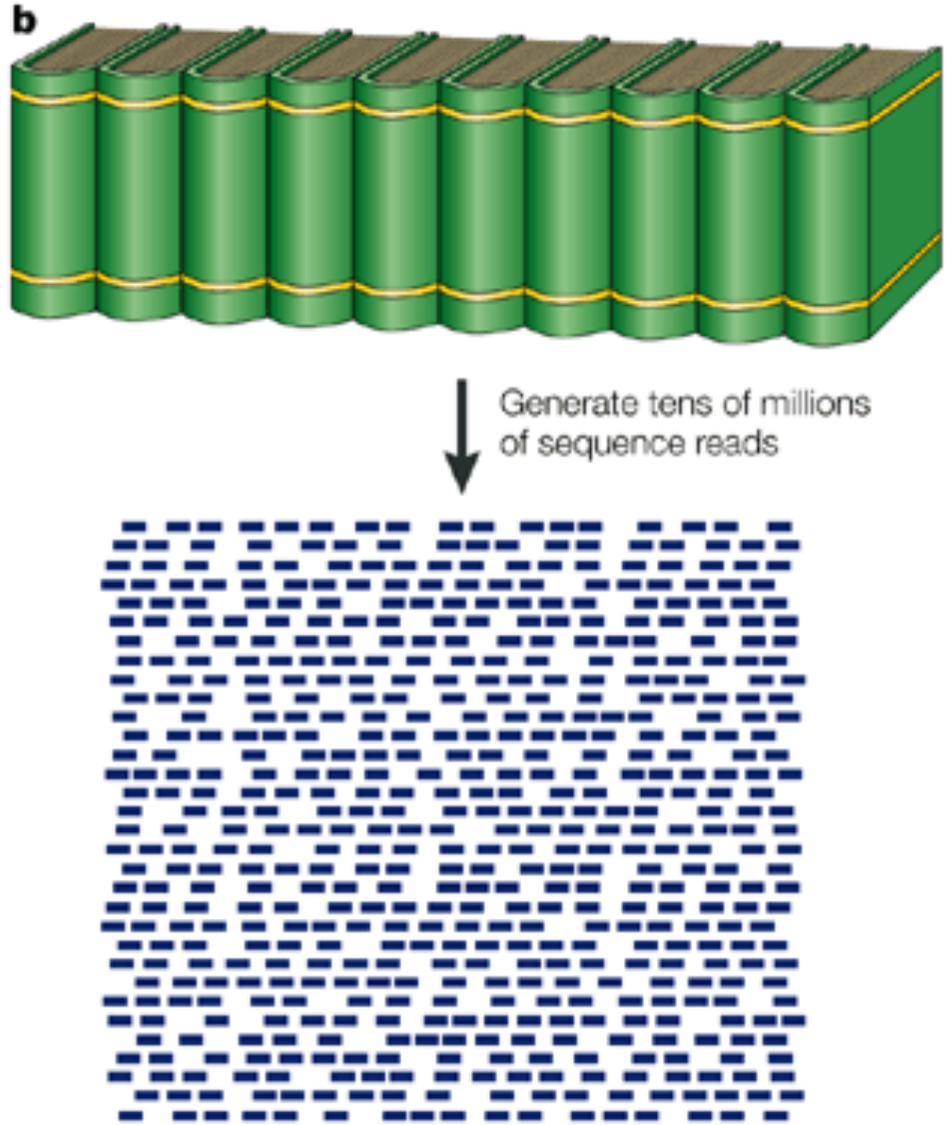
Human Genome Project

Heirarchical Shotgun Sequencing



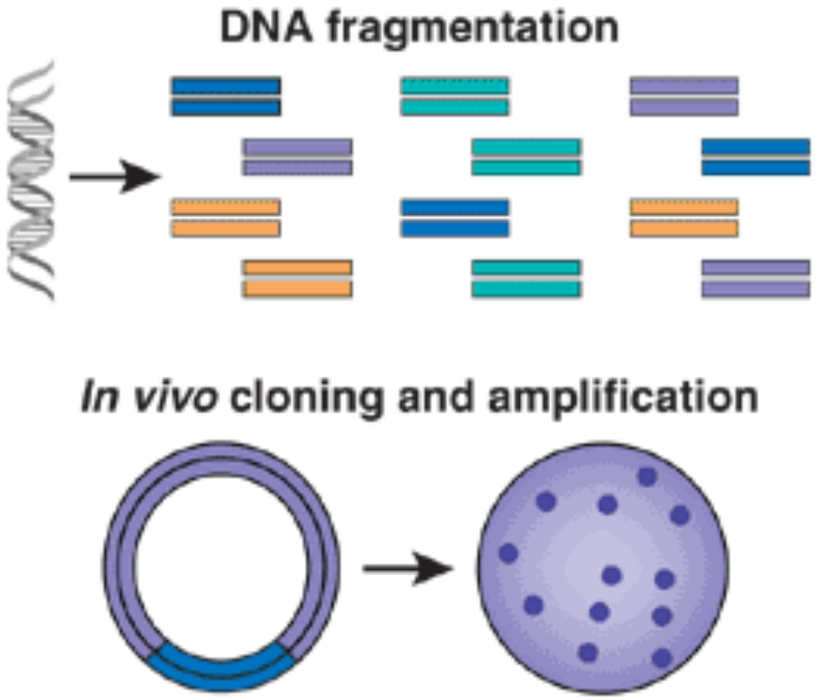
Celera Genomics

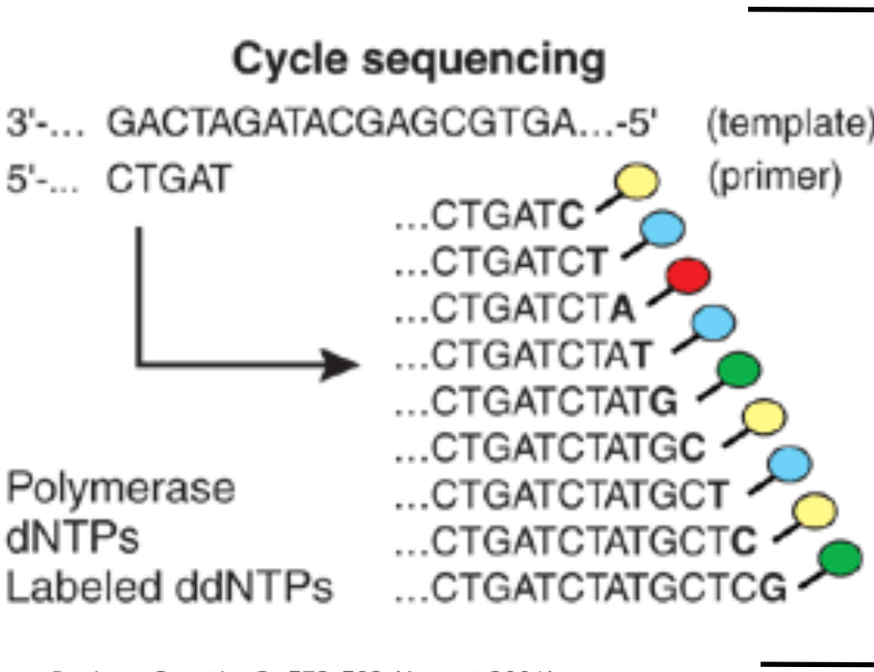
Whole Genome Shotgun Sequencing

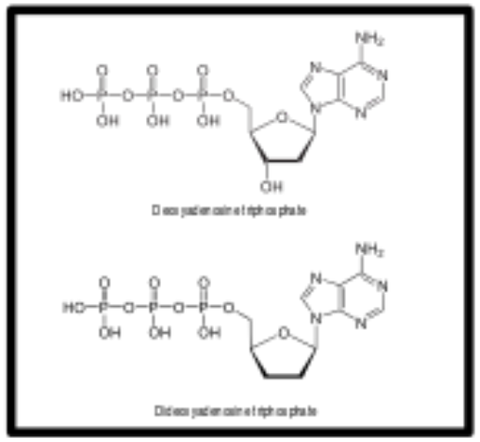


Nature Reviews Genetics 2, 573-583 (August 2001)

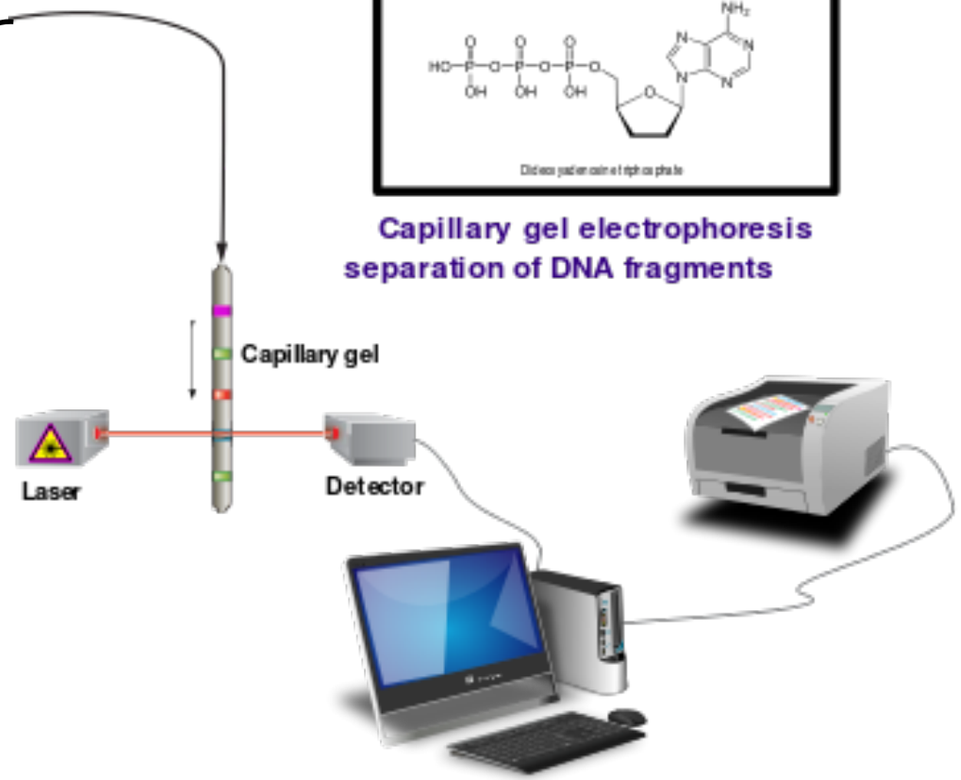
Sequencing of the Human Genome

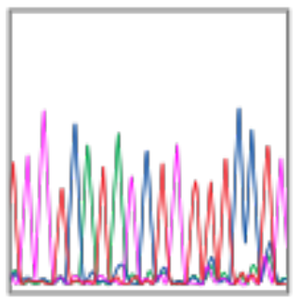






Capillary gel electrophoresis separation of DNA fragments

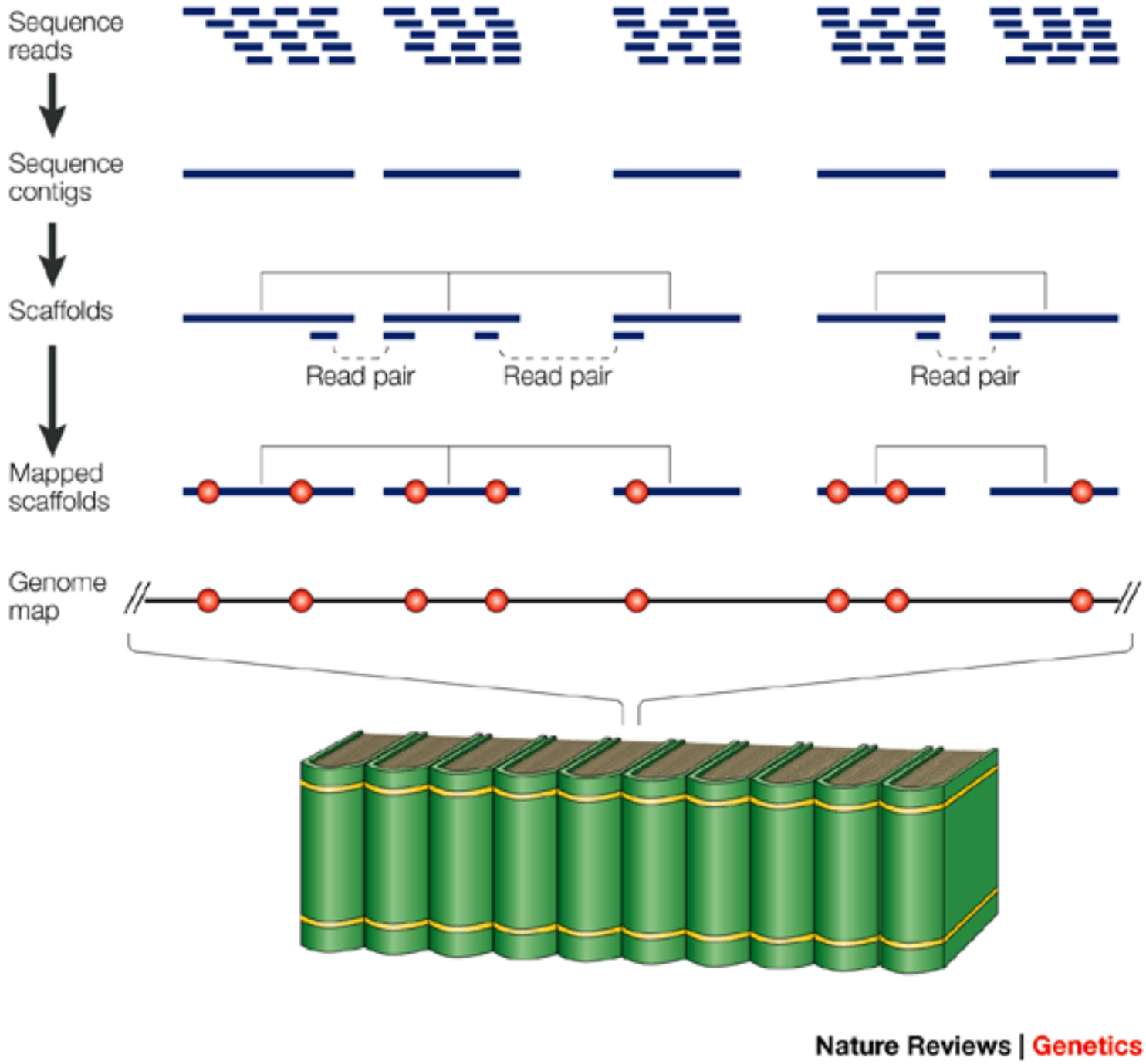




Laser detection of flourochromes and computational sequence analysis

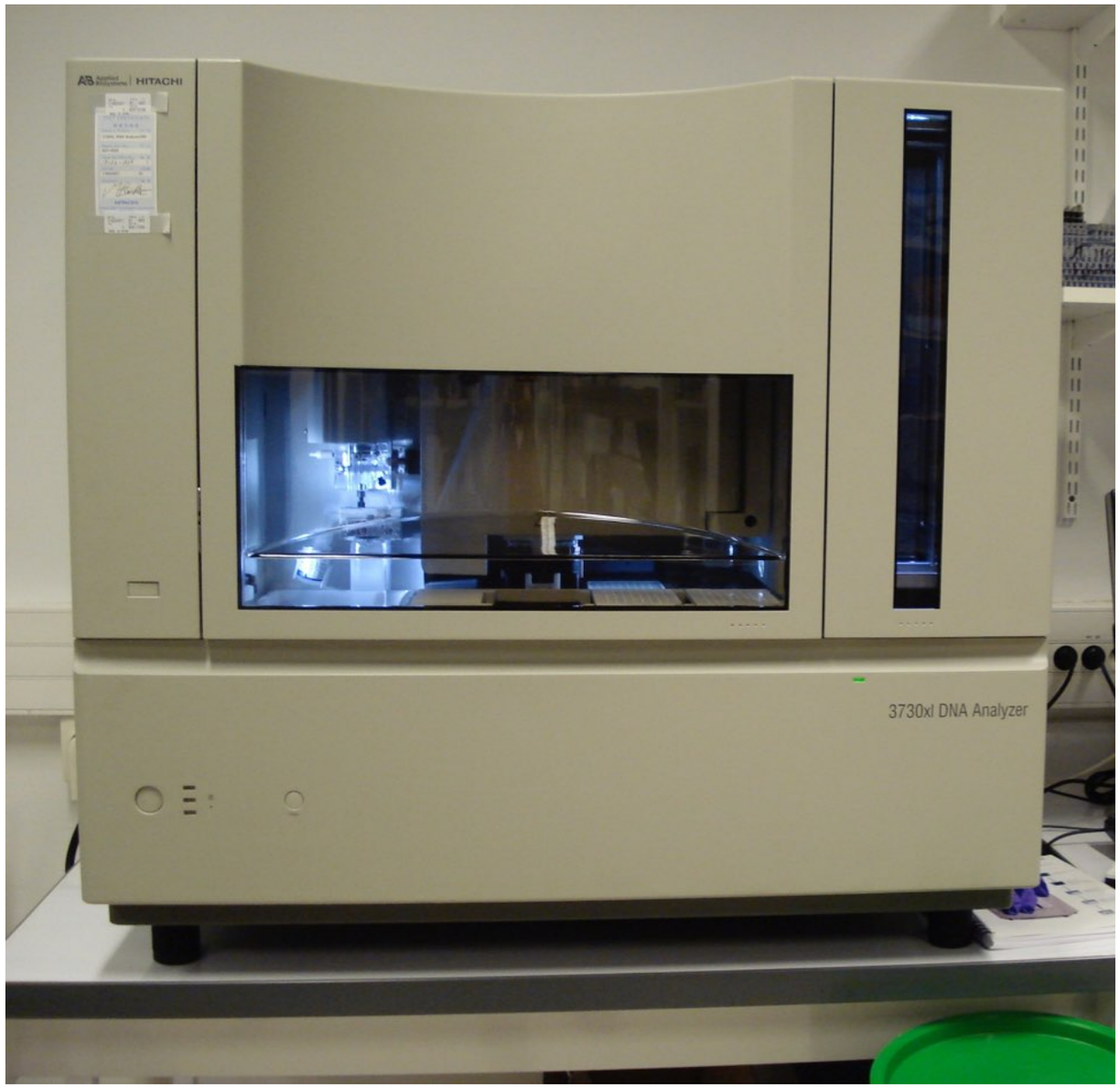
Nature Reviews Genetics **2**, 573-583 (August 2001)

Sequencing of the Human Genome



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

Sequencing of the Human Genome

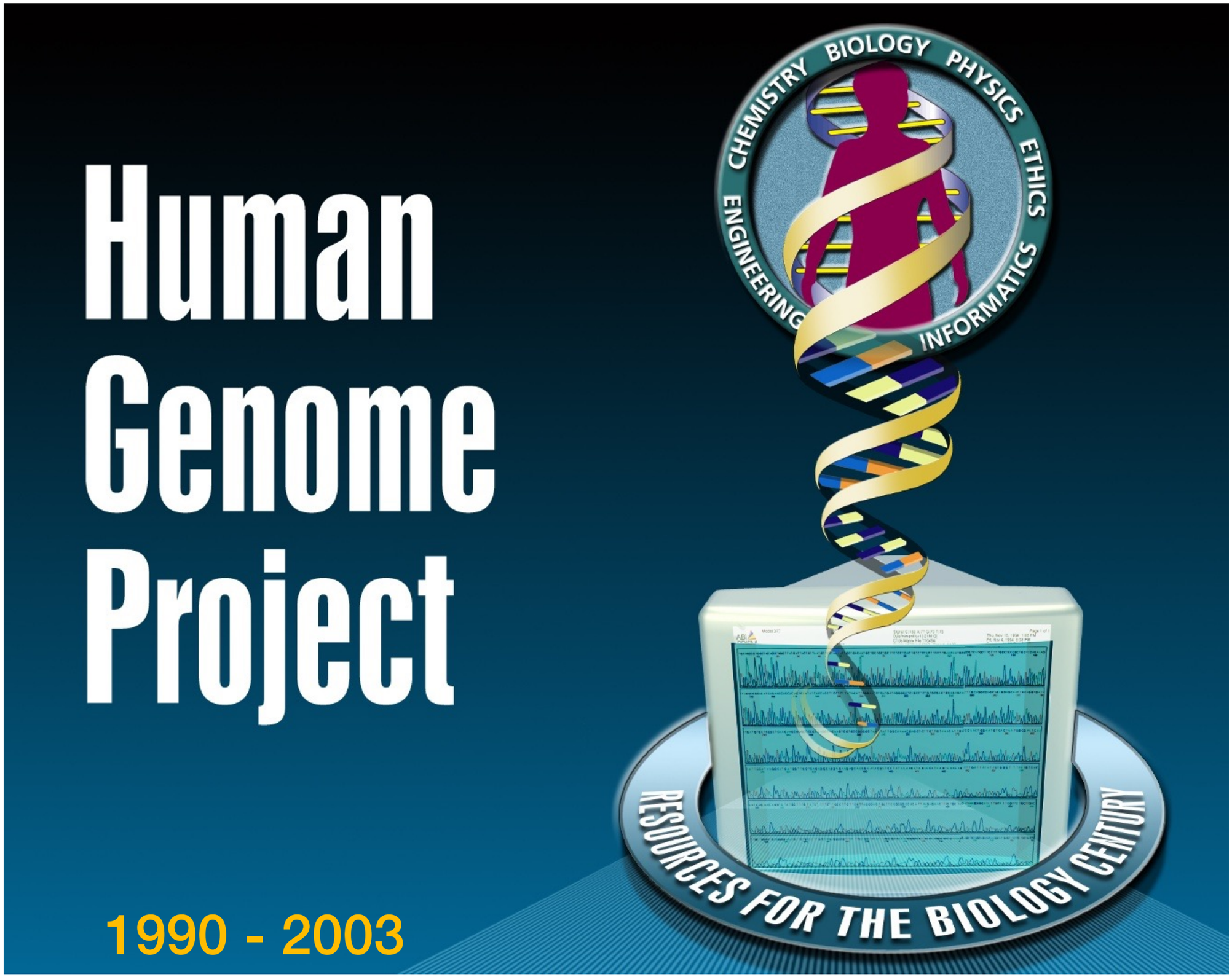


http://www.pasteur.fr/ip/portal/action/WebdriveActionEvent/oid/01s-00001u-01p

Cycle (Sanger) sequencing generated:

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

Sequence production was rate limiting, not analysis

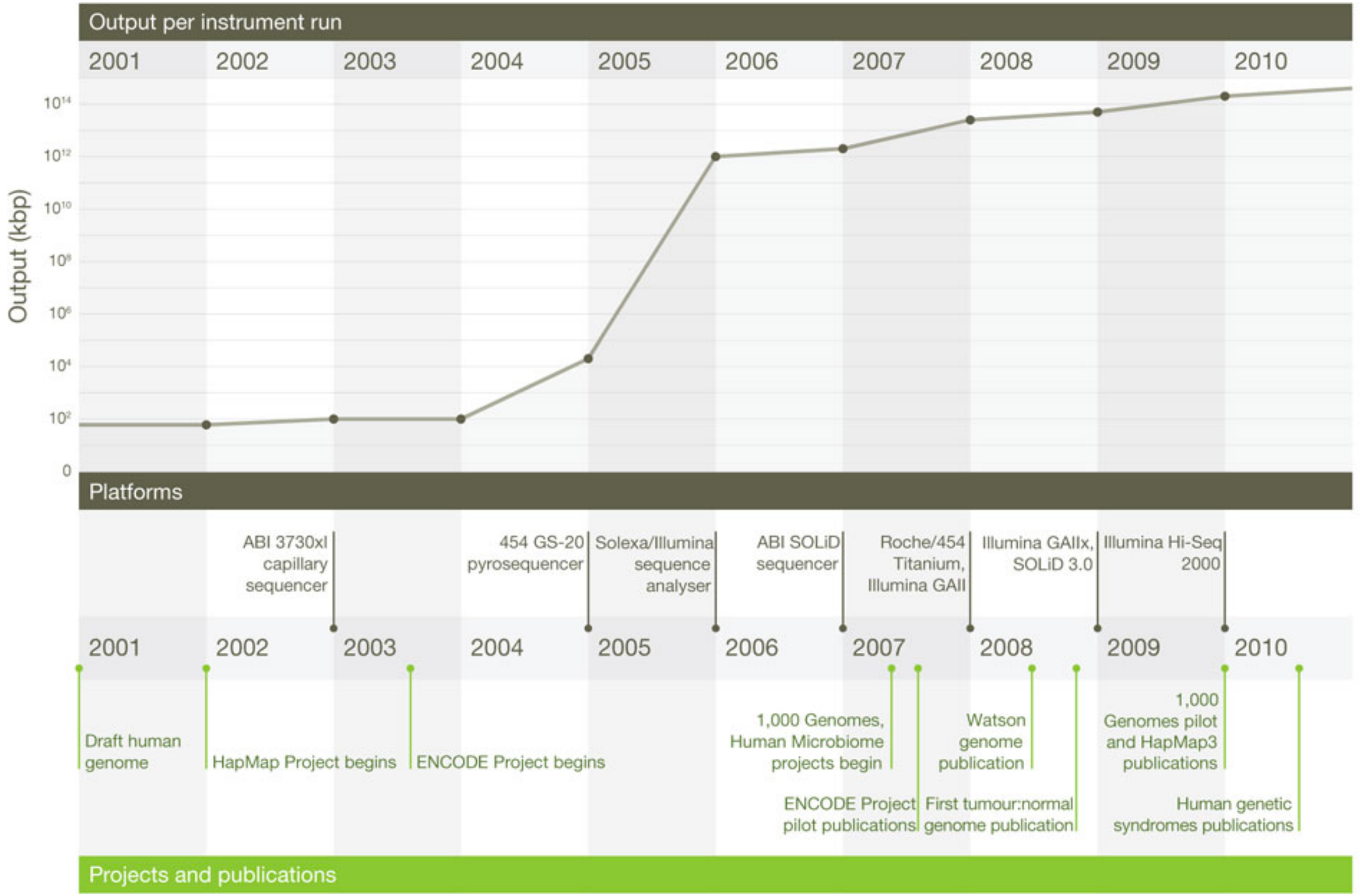


http://blogs.swa-jkt.com/swa/10318/2014/03/18/biology-human-genome-project-prompt/

Comparative Genomics Research

Highly conserved regions of DNA likely to be functional

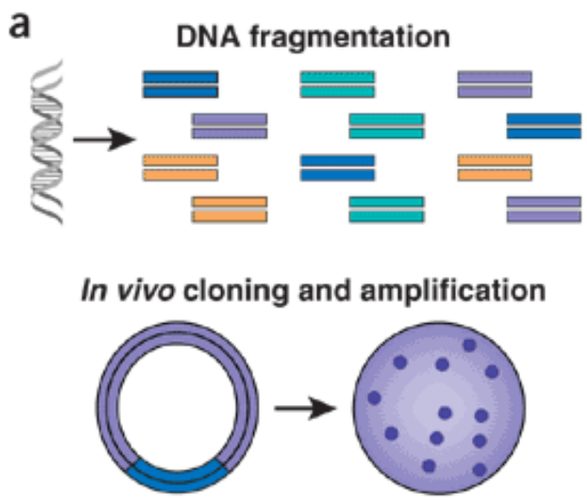
Advancements in Sequencing Technology

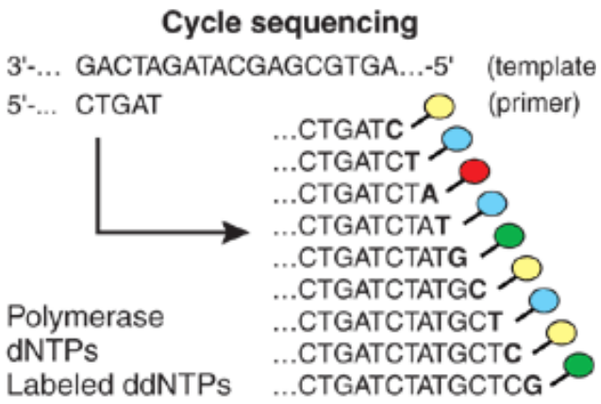


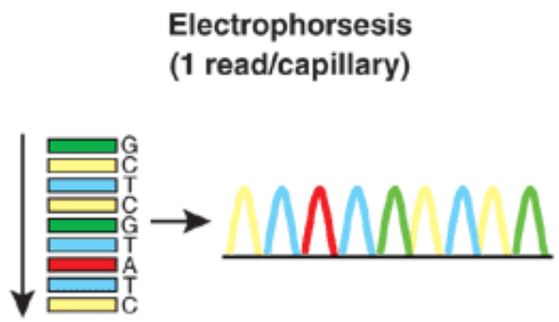
Nature **470**, 198–203 (10 February 2011)

NGS Technologies

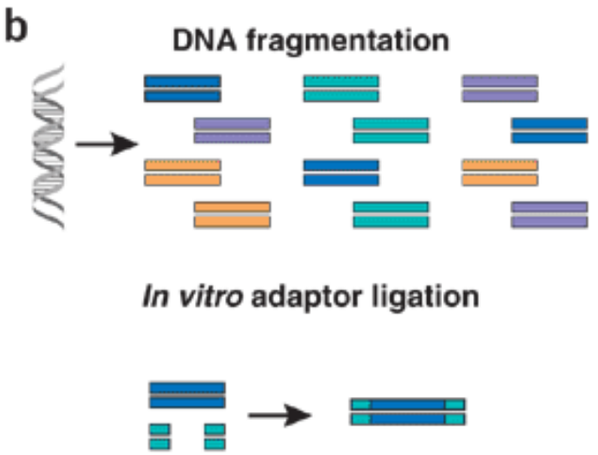
Cycle Sequencing

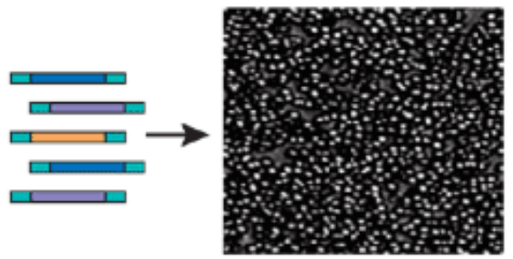


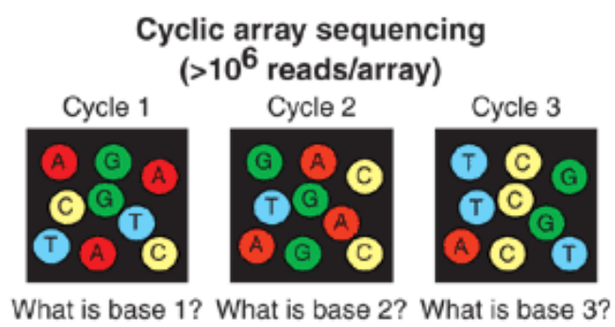




NGS Sequencing

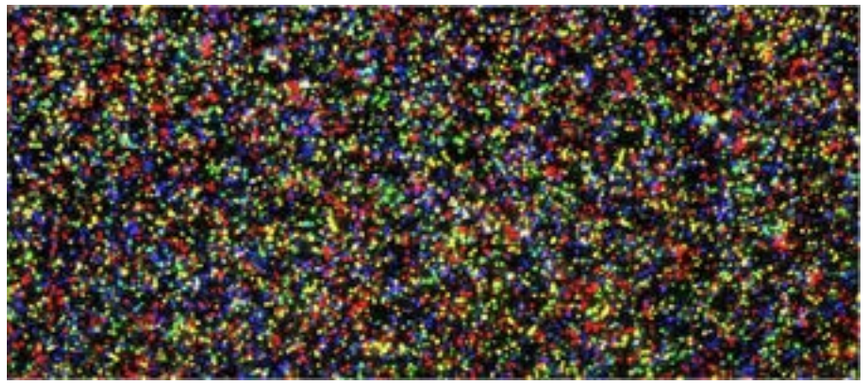






Nature Reviews Genetics **2**, 573-583 (August 2001)

NGS Technologies



Generate:

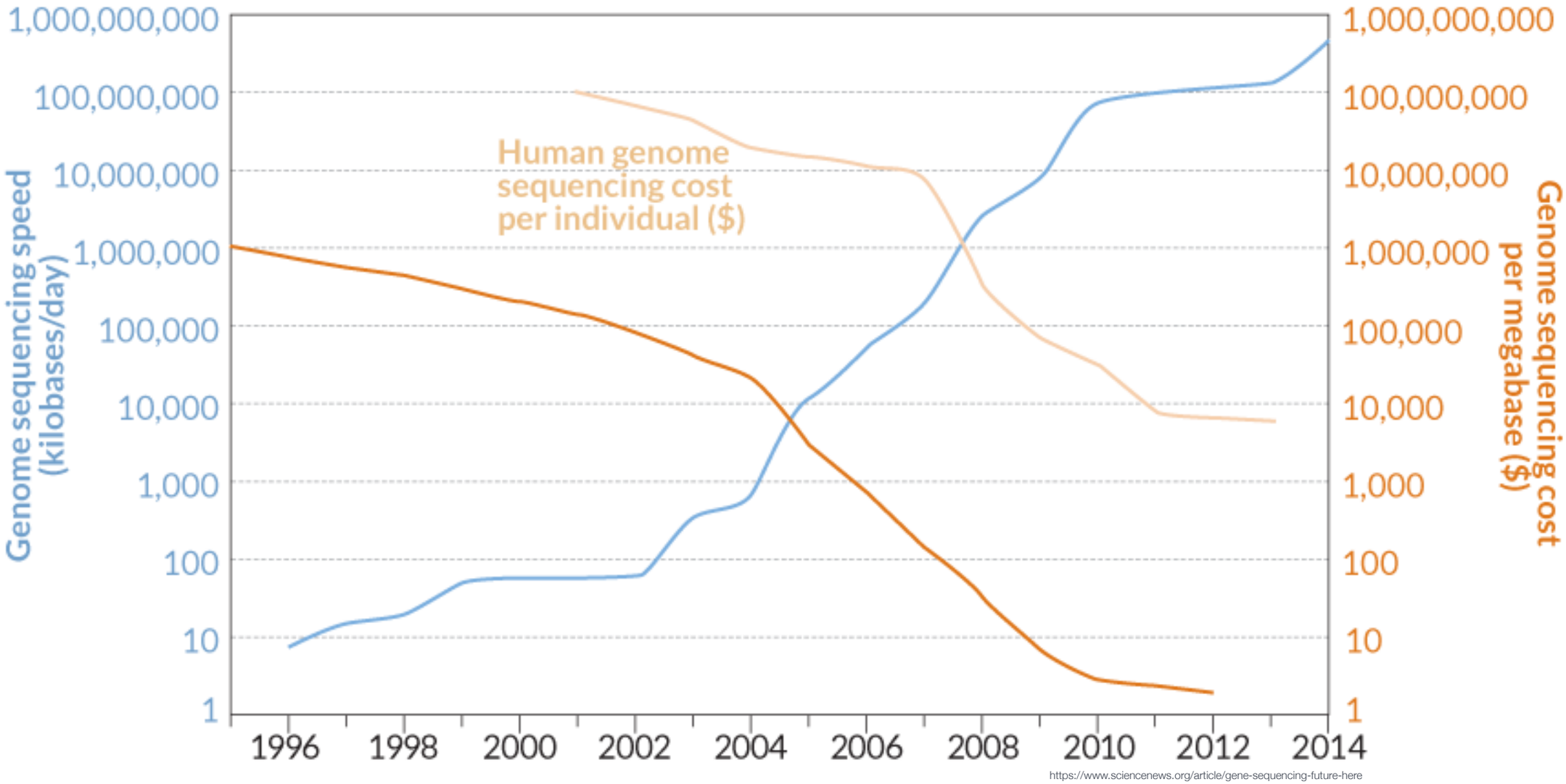
- Illumina HiSeq2500: 125 bp / read (250 bp / read Rapid-Run Mode)
- > **100 billion bp** / day

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

The Genomic Era: Individual Projects

Credit: Wellcome Library, London]

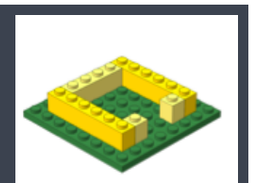
- Only a few experiments = vast amounts of data
- Data generation straightforward, but analysis requires bioinformatics expertise

Bioinformatics in the Genomic Era

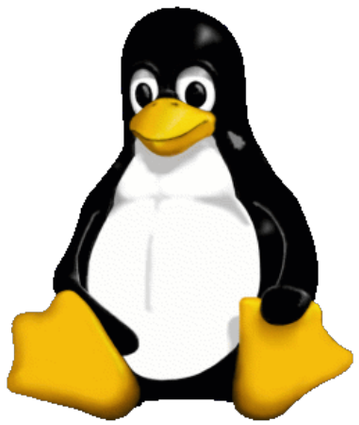
Alliances between experimentalists and computational biologists

Bioinformatics Toolkit

Programming languages



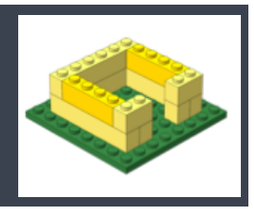
Programming languages are critical in genomic analyses:



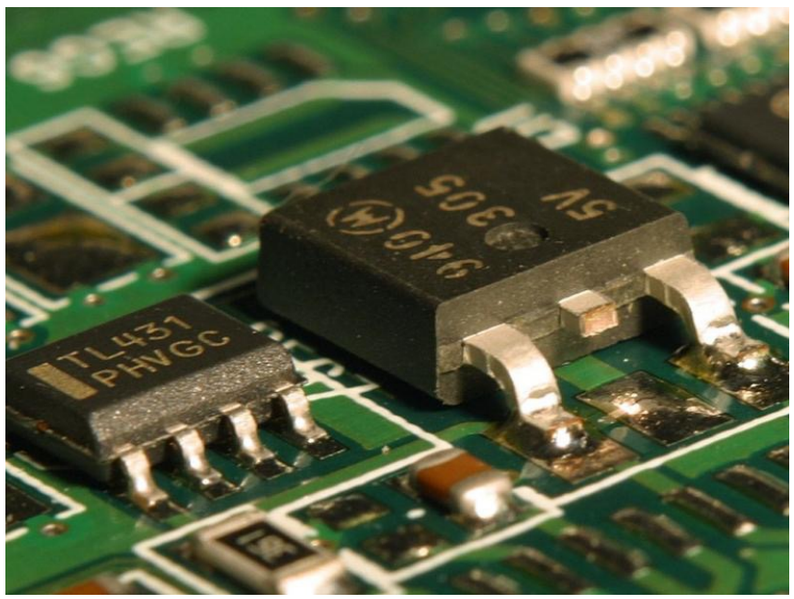
- **Bash:** command line language used for interacting with *Linux / Unix-based* operating systems.
  - Attaining sequencing data
  - Accessing computing resources and analysis tools
  - Basic data manipulation, creating scripts for running tools

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

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

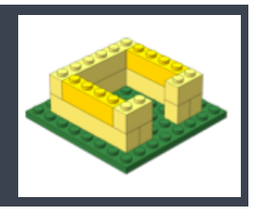
Memory/RAM (Random Access Memory)

- Large datasets ≈ lots of RAM to perform analysis

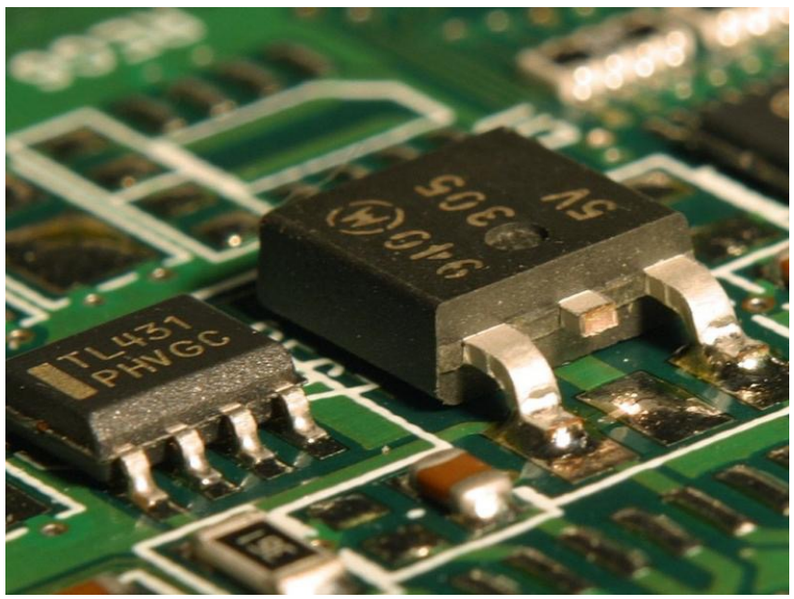
CPU (Central Processing Unit)

- 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)

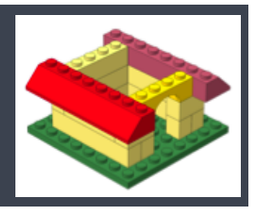
- Large datasets ≈ lots of RAM to perform analysis

CPU (Central Processing Unit)

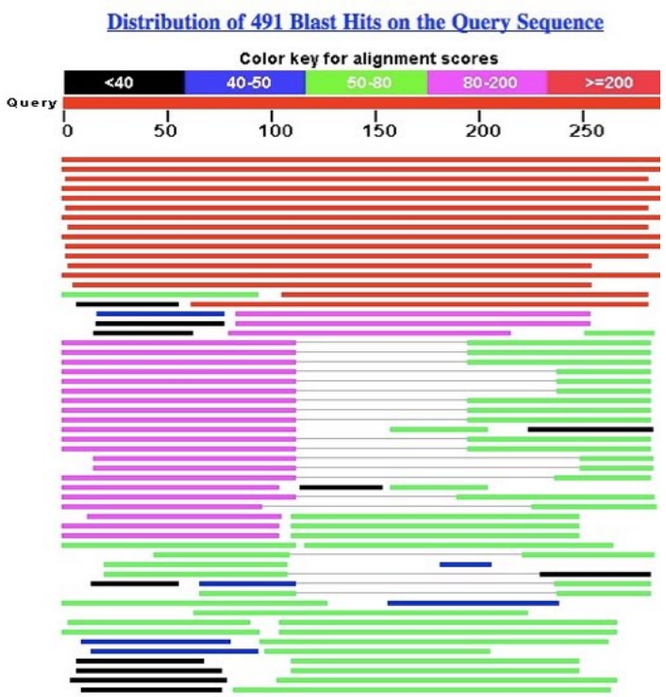
- Large datasets ≈ lots of time to perform analysis

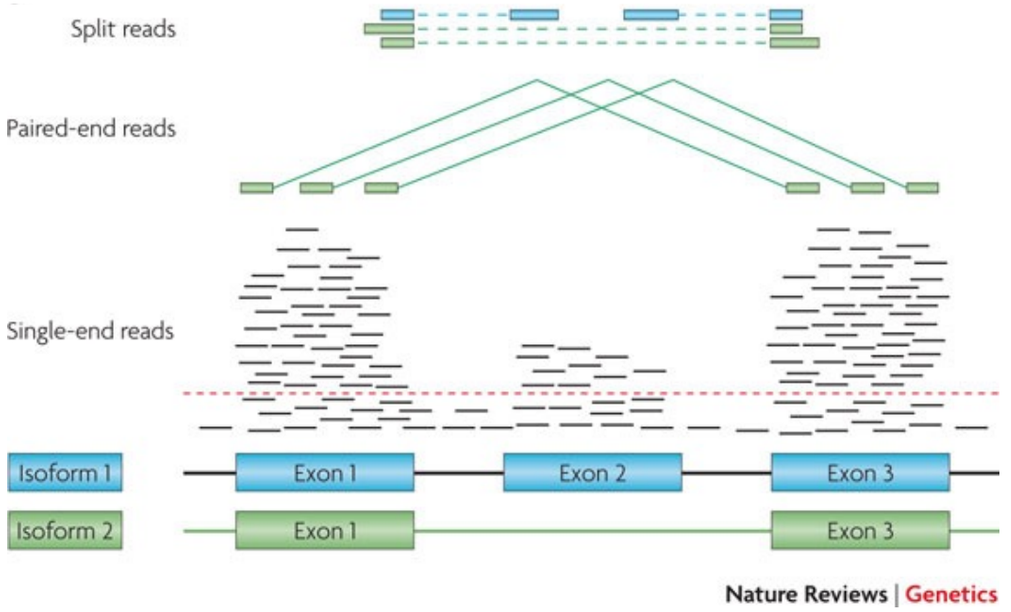
Solution: Amazon Cloud or computing cluster

NGS Analysis Tools and Workflows

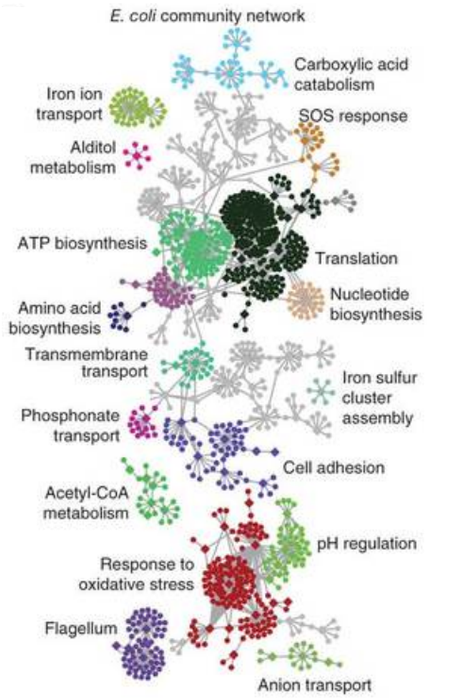


Large genomics datasets require software (tools) to perform each of the steps in an NGS analysis workflow.



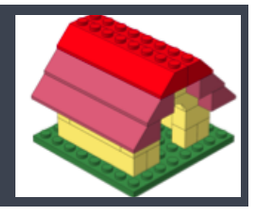


Nature Reviews Genetics **11**, 559-571 (August 2010)



Nature Methods **9**, 796–804 (2012)

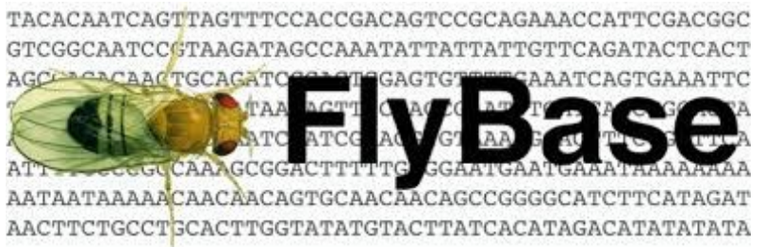
Genome Databases



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



FlyBase

e!Ensembl

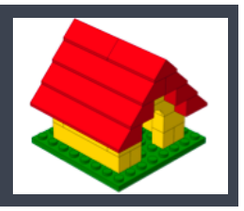
UCSC Genome Bioinformatics

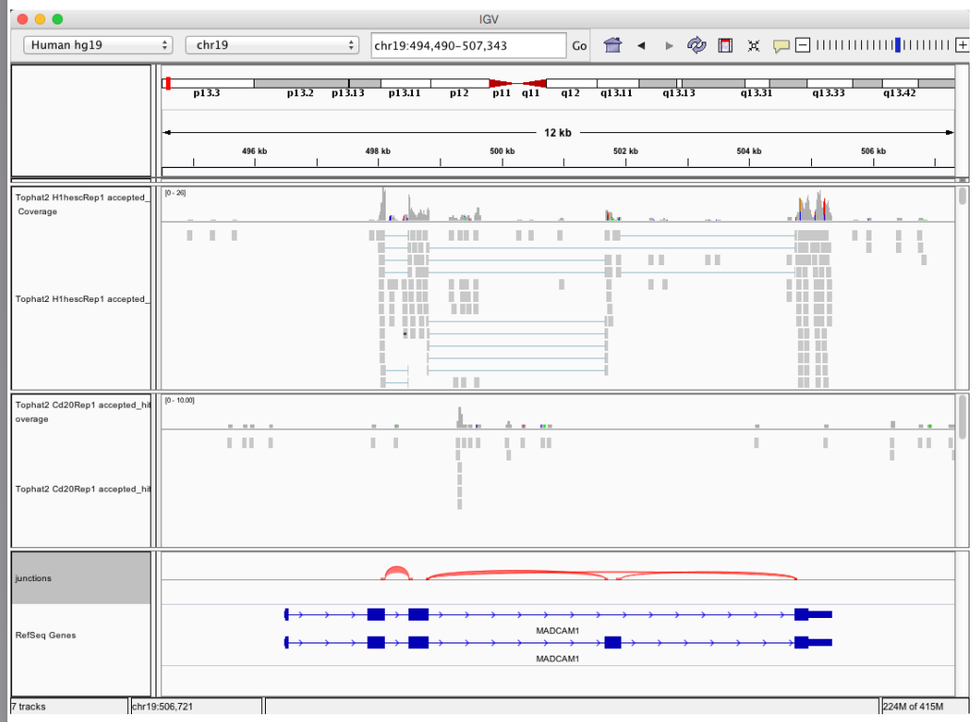
ViPR
Virus Pathogen Resource

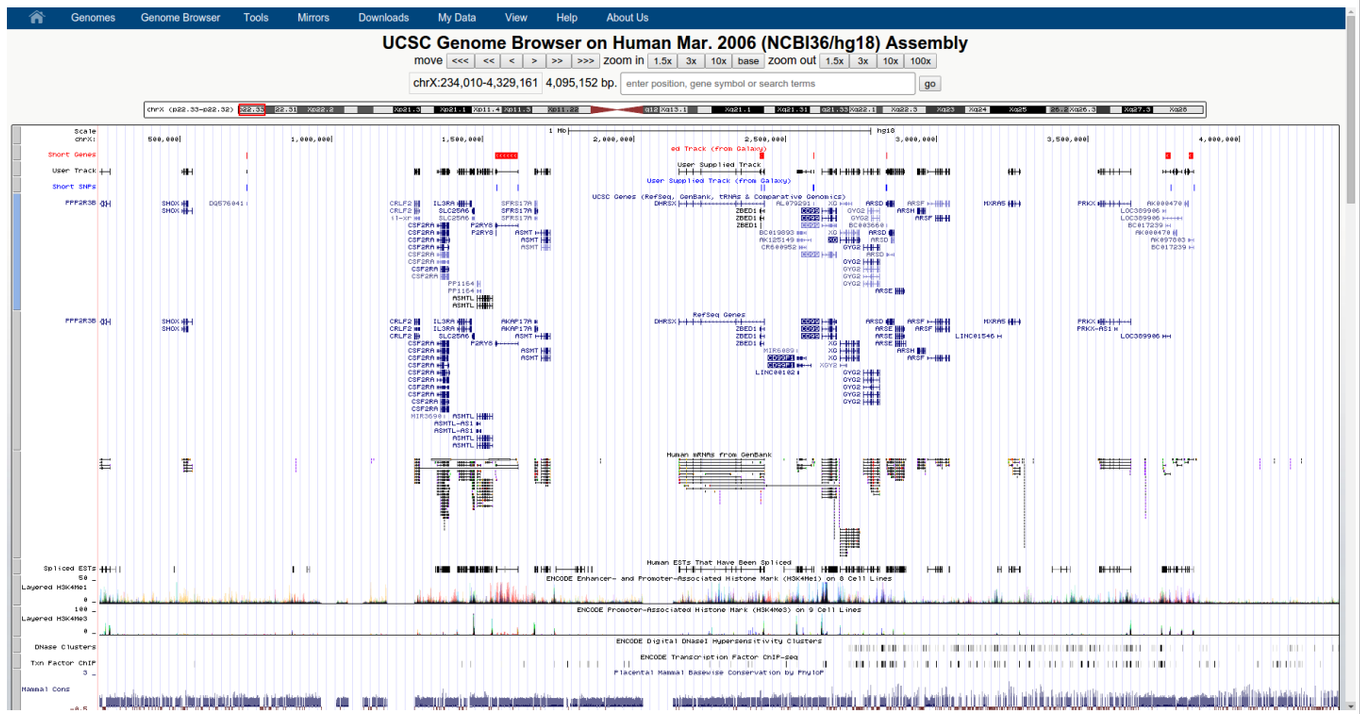
PATRIC
Pathosystems Resource Integration Center

WormBase

Genome Browsers







Visualize genomic data from:

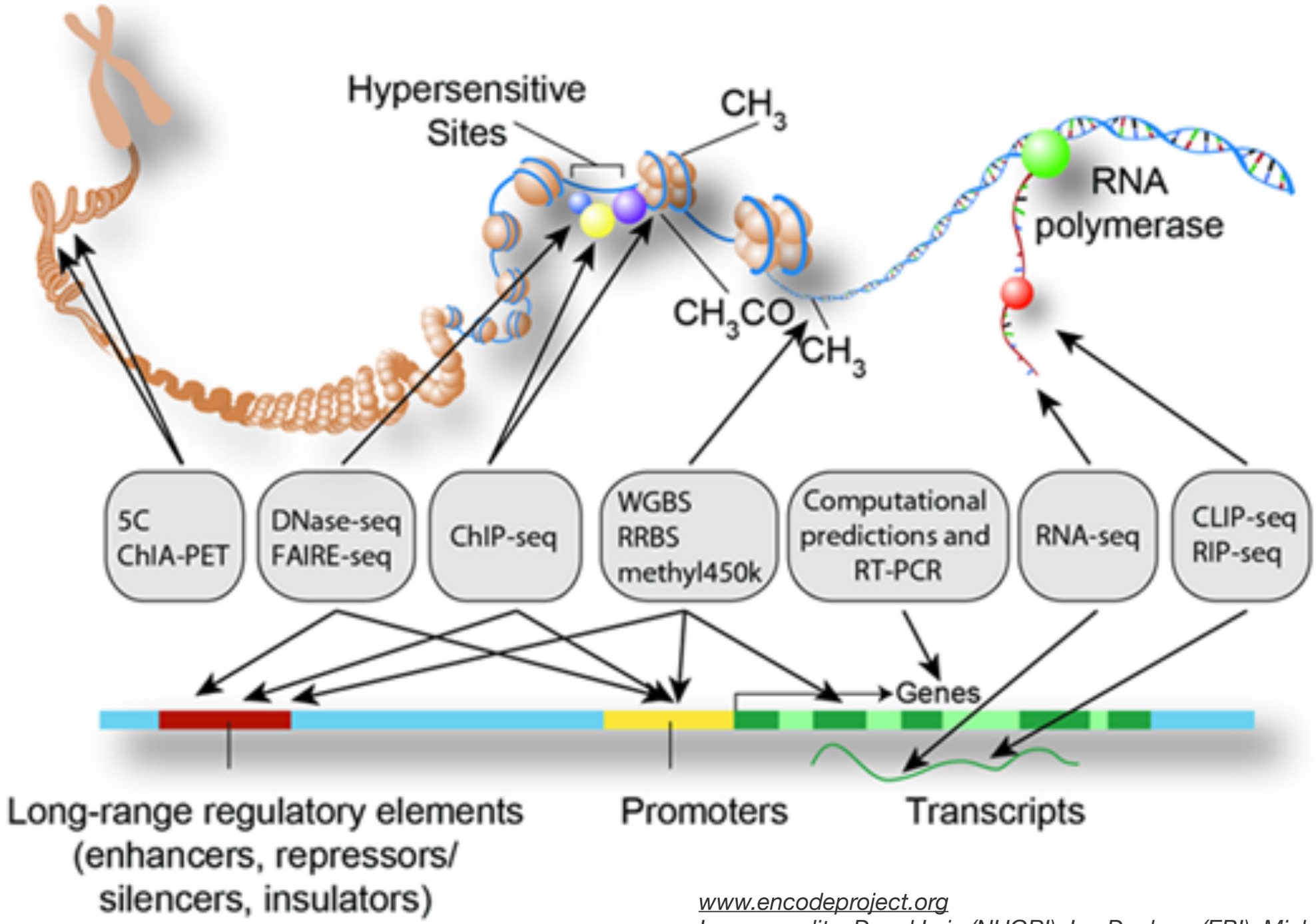
- ***Genome databases:*** entire genomes, regulation sequences, gene predictions and structures, and data from comparative analyses.
- ***Your own analyses***

Community

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

NGS Applications

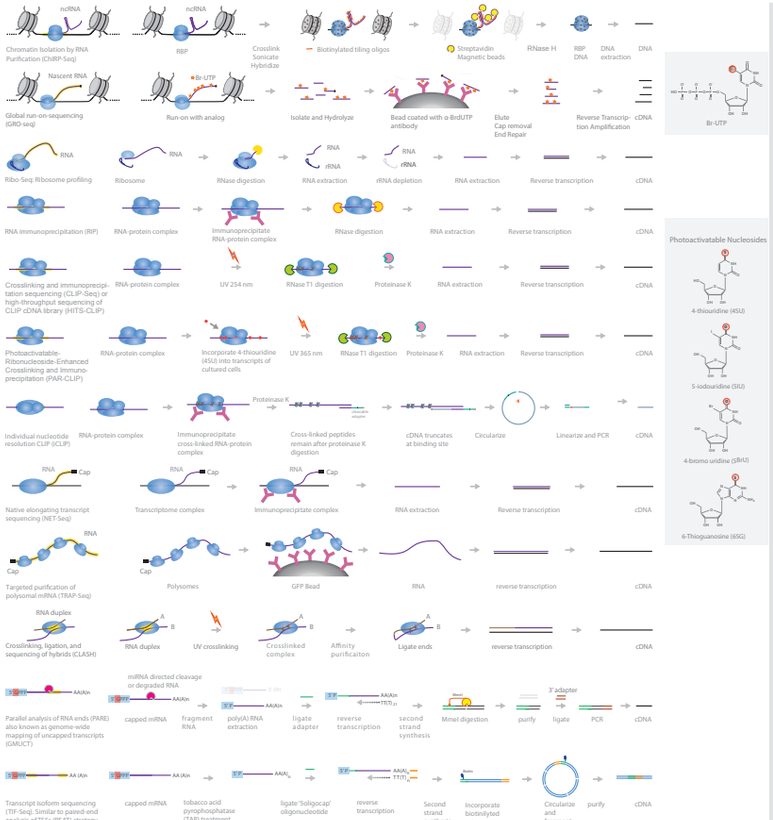
NGS Applications



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

RNA Transcription

ChIRP-Seq
GRO-seq
Ribo-Seq ARTseq™
RIP-Seq
HITS-CLIP CLIP-Seq
PAR-CLIP
iCLIP
NET-Seq
TRAP-Seq
CLASH
PARE-Seq GMUCT
TIF-Seq PEAT

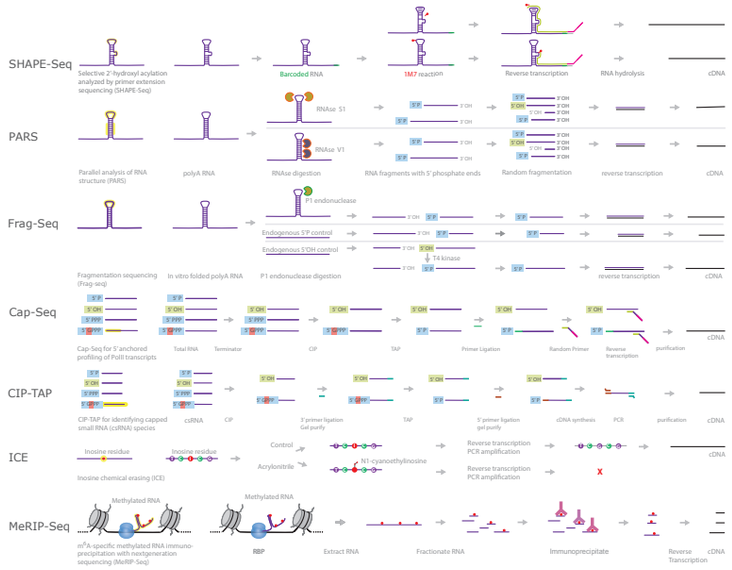


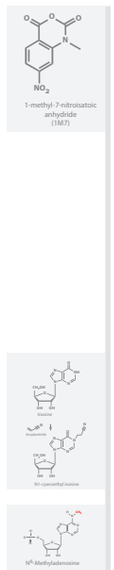
RNA Low-Level Detection

Digital RNA
Quartz-Seq
DP-Seq
Smart-seq
Smart-seq2
UMI Method
CEL-Seq
STRT-Seq

RNA Structure

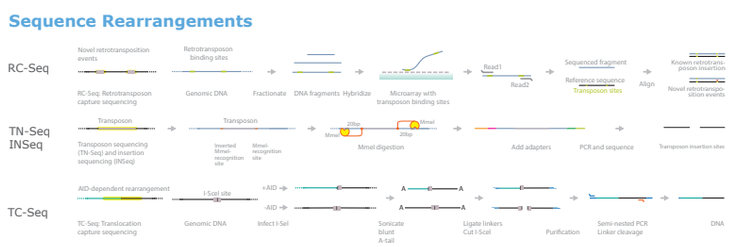
SHAPE-Seq
PARS
Frag-Seq
Cap-Seq
CIP-TAP
ICE
MeRIP-Seq





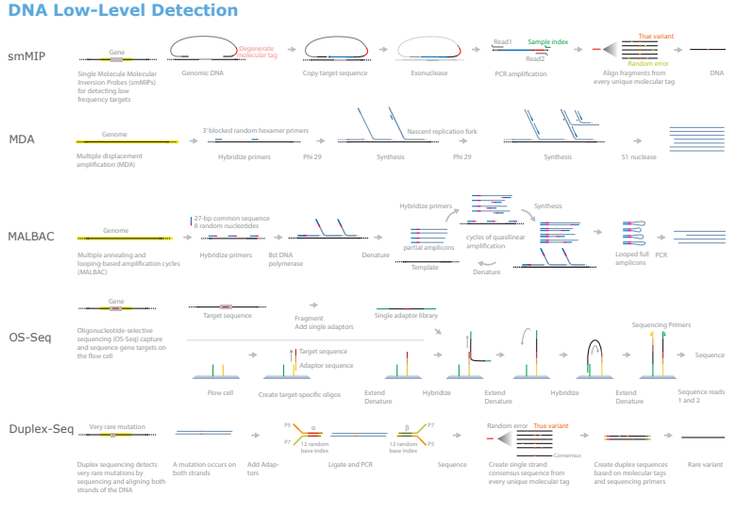
Sequence Rearrangements

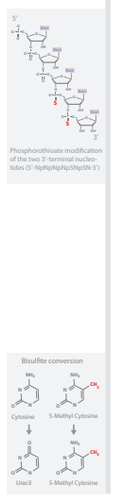
RC-Seq
TN-Seq INSeq
TC-Seq



DNA Low-Level Detection

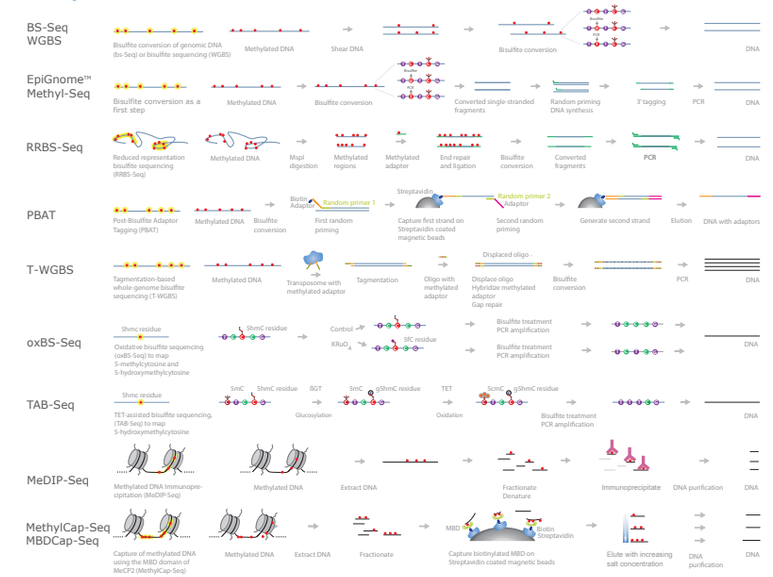
smMIP
MDA
MALBAC
OS-Seq
Duplex-Seq





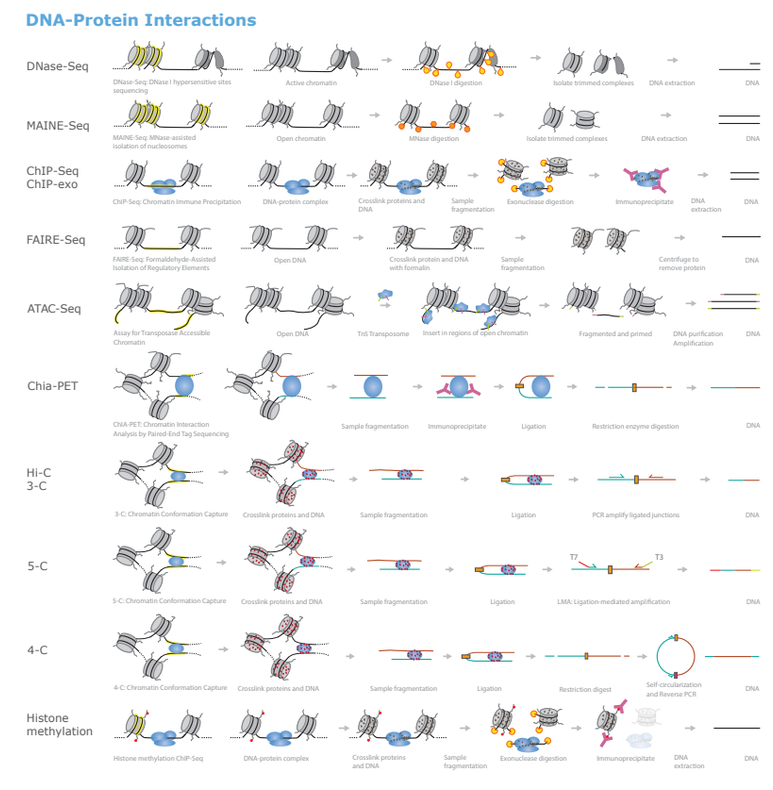
Methylation

BS-Seq WGBS
EpiGnome™ Methyl-Seq
RRBS-Seq
PBAT
T-WGBS
oxBS-Seq
TAB-Seq
MeDIP-Seq
MethylCap-Seq MBDCap-Seq

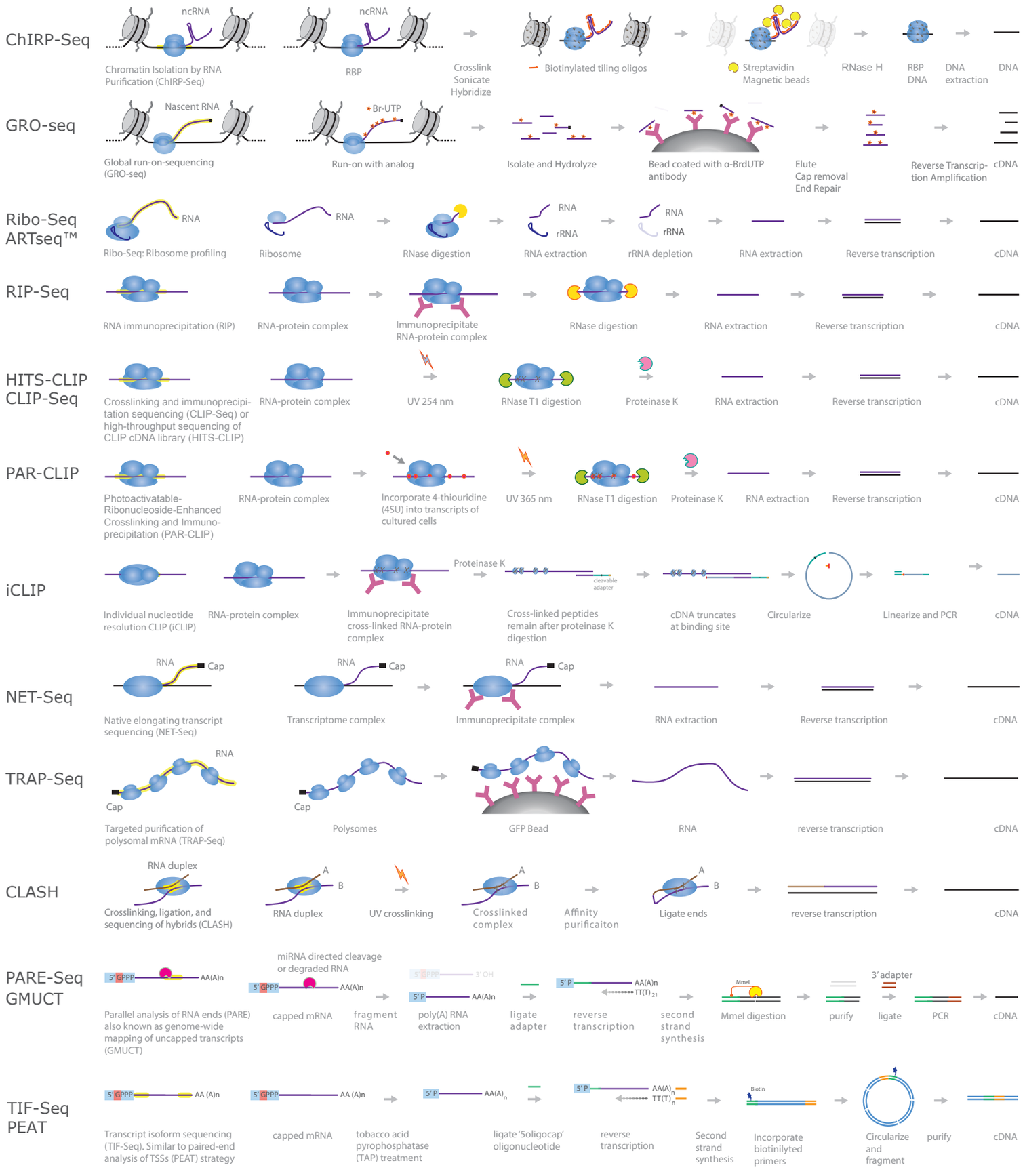


DNA-Protein Interactions

DNase-Seq
MAINE-Seq
ChIP-Seq ChIP-exo
FAIRE-Seq
ATAC-Seq
Chia-PET
Hi-C 3-C
5-C
4-C
Histone methylation
Key



RNA Transcription



Expectations

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