Overview of genomics

Pakorn Aiewsakun

Department of Microbiology

Faculty of Science, Mahidol University

Outline

- What is genomics?
- Sequencing technologies
- What can we do with whole genome data?
- What are we going to do in this workshop?

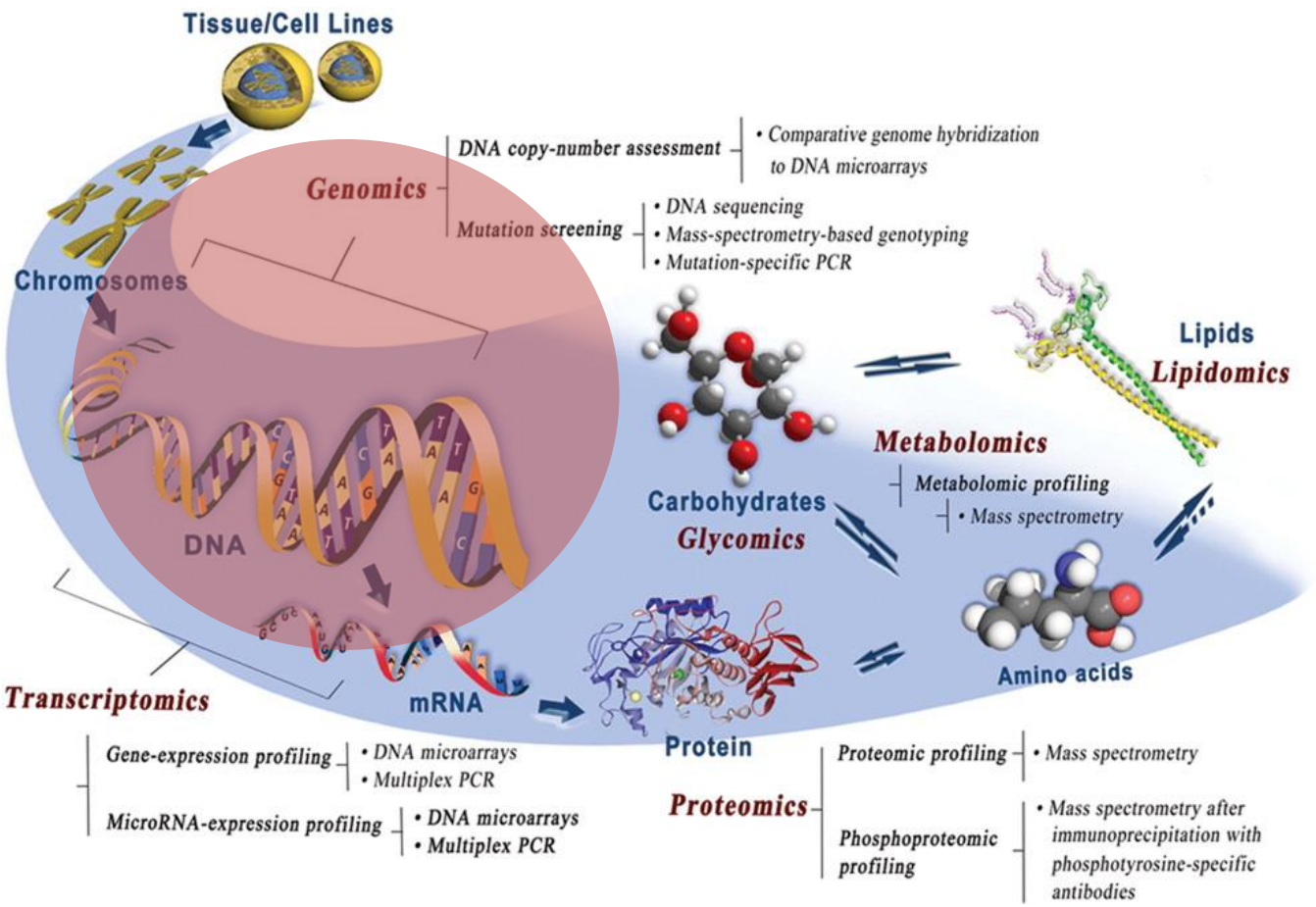
What is genomics?

The study of whole genomes of organisms

What is genomics?

- **Genomics** is the study of whole genomes of organisms

- **Genomics** is a field of biology focusing on the structure, function, evolution, mapping, and editing of genomes.
https://en.wikipedia.org/wiki/Genomics

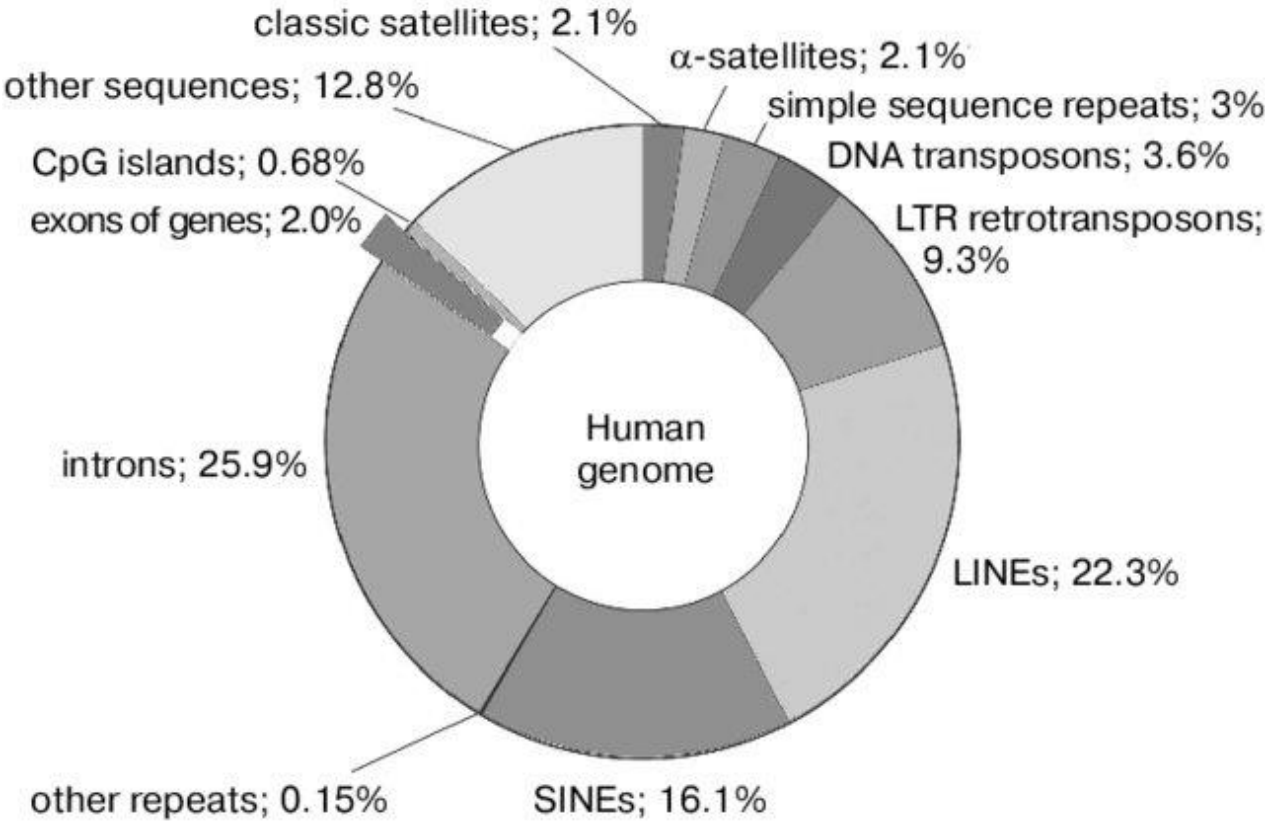


Wu et al., 2011, Journal of Dental Research

What is genomics?

- A **genome** is an organism's (abstract/consensus) complete set of DNA, including all of its genes (*and beyond*!).

- A **genome** is a set of information telling how an organism grows and develops



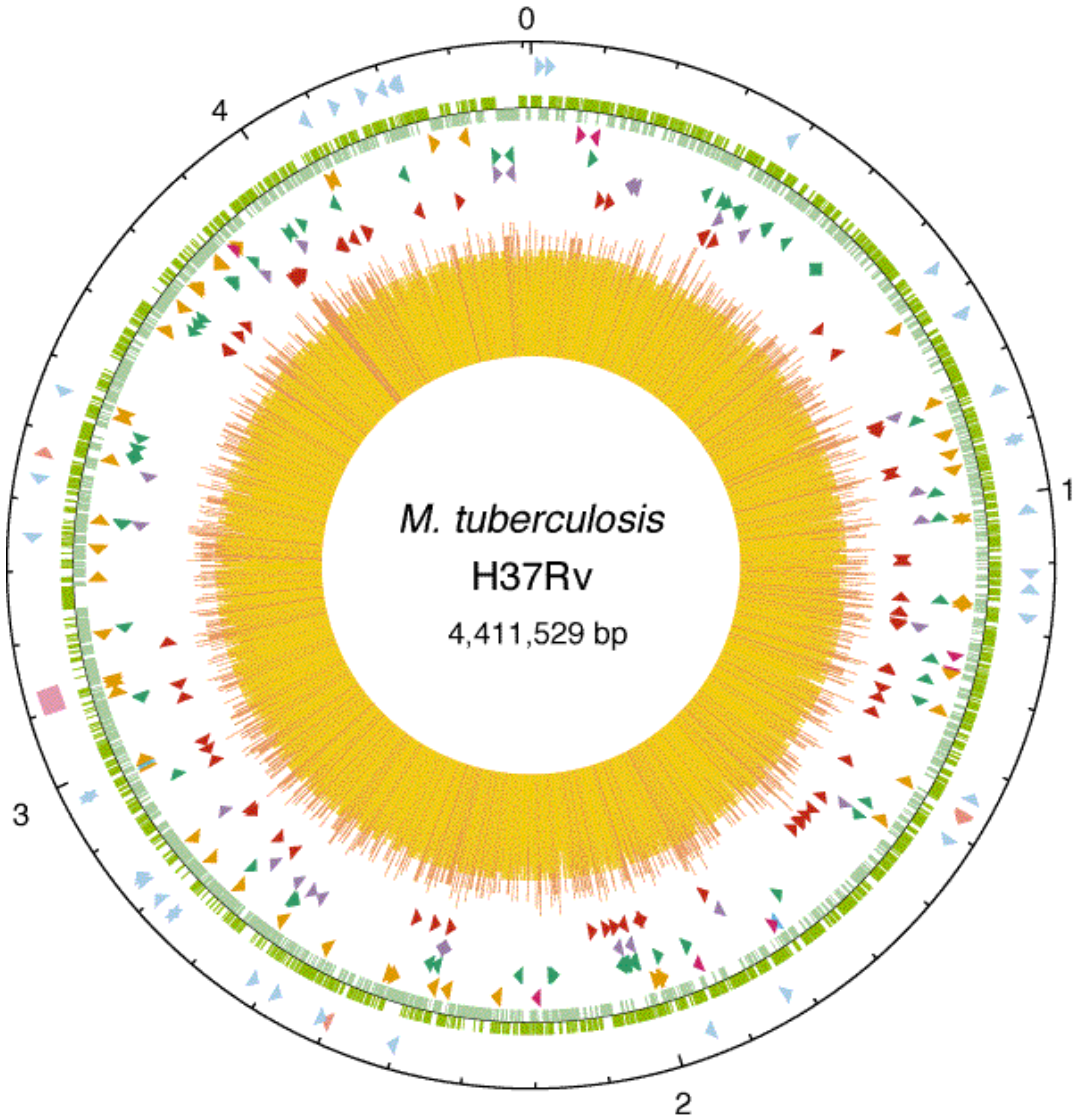
Human genome content

Patrushev & Minkevich, 2008, Biochemistry (Moscow)

What is genomics?

- A **genome** is an organism's (<u>abstract/consensus</u>) complete set of DNA, including all of its genes (*and beyond*!).

- A **genome** is a set of information telling how an organism grows and develops

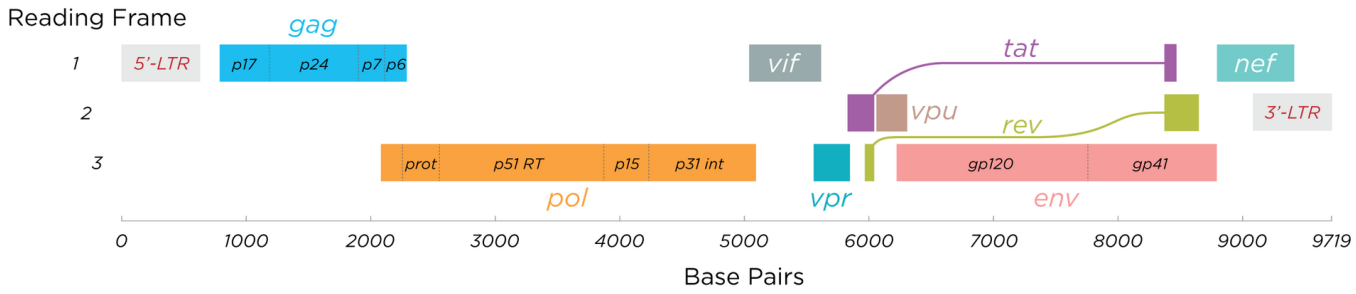


The complete genome sequence of the best-characterized strain of *Mycobacterium tuberculosis*, H37Rv

Cole et al., 1998, Nature

What is genomics?

- A **genome** is an organism's (abstract/consensus) complete set of DNA, including all of its genes (*and beyond*!).



- A **genome** is a set of information telling how an organism grows and develops

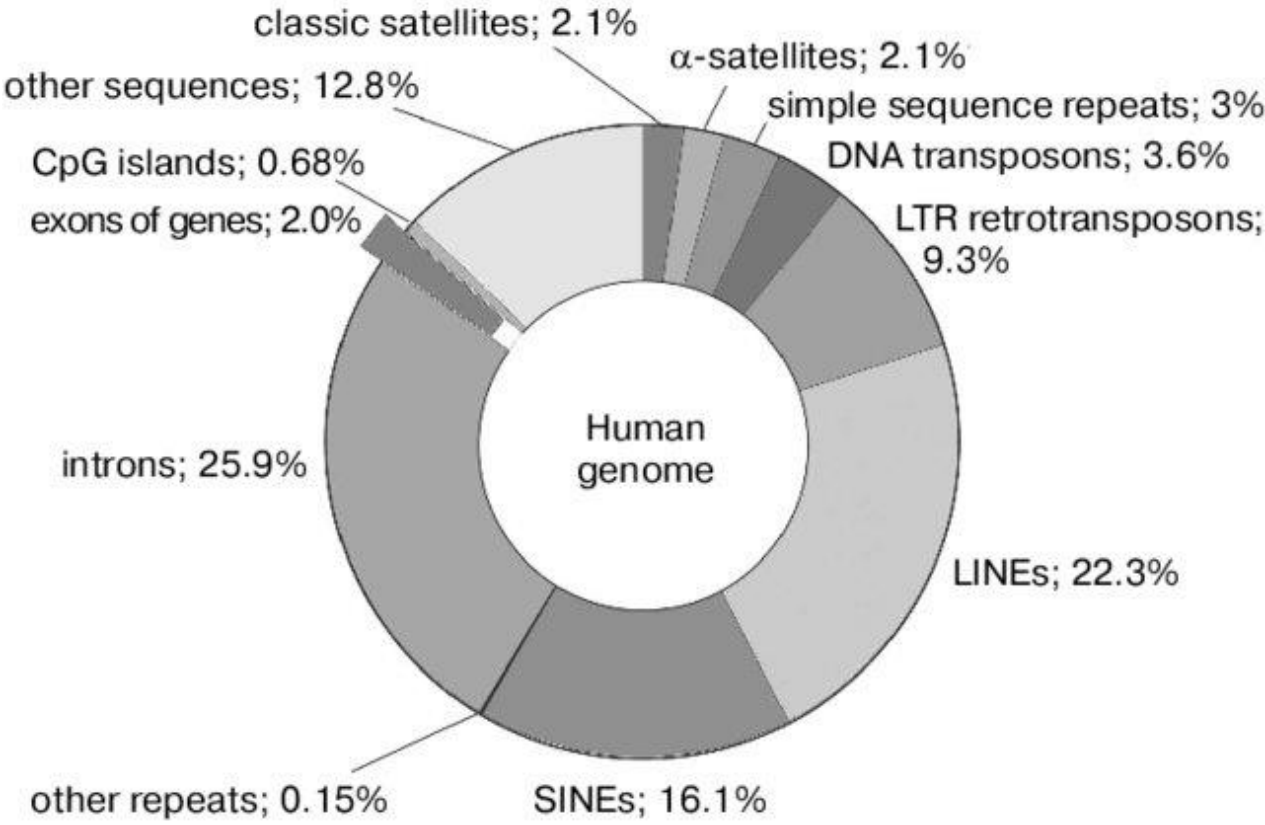
Structure of the RNA genome of HIV-1

Thomas Splettstoesser (www.scistyle.com), 214

What is genomics?

- **Genomics ≠ Genetics**!

- **Genetics** is the study of <u>individual</u> genes and their roles in <u>inheritance</u>
- **Genomics** aims at the collective characterisation and quantification of <u>all</u> of an organism's genes (*and other genomic elements*), their interrelations (*and modifications*), and influence on the organism

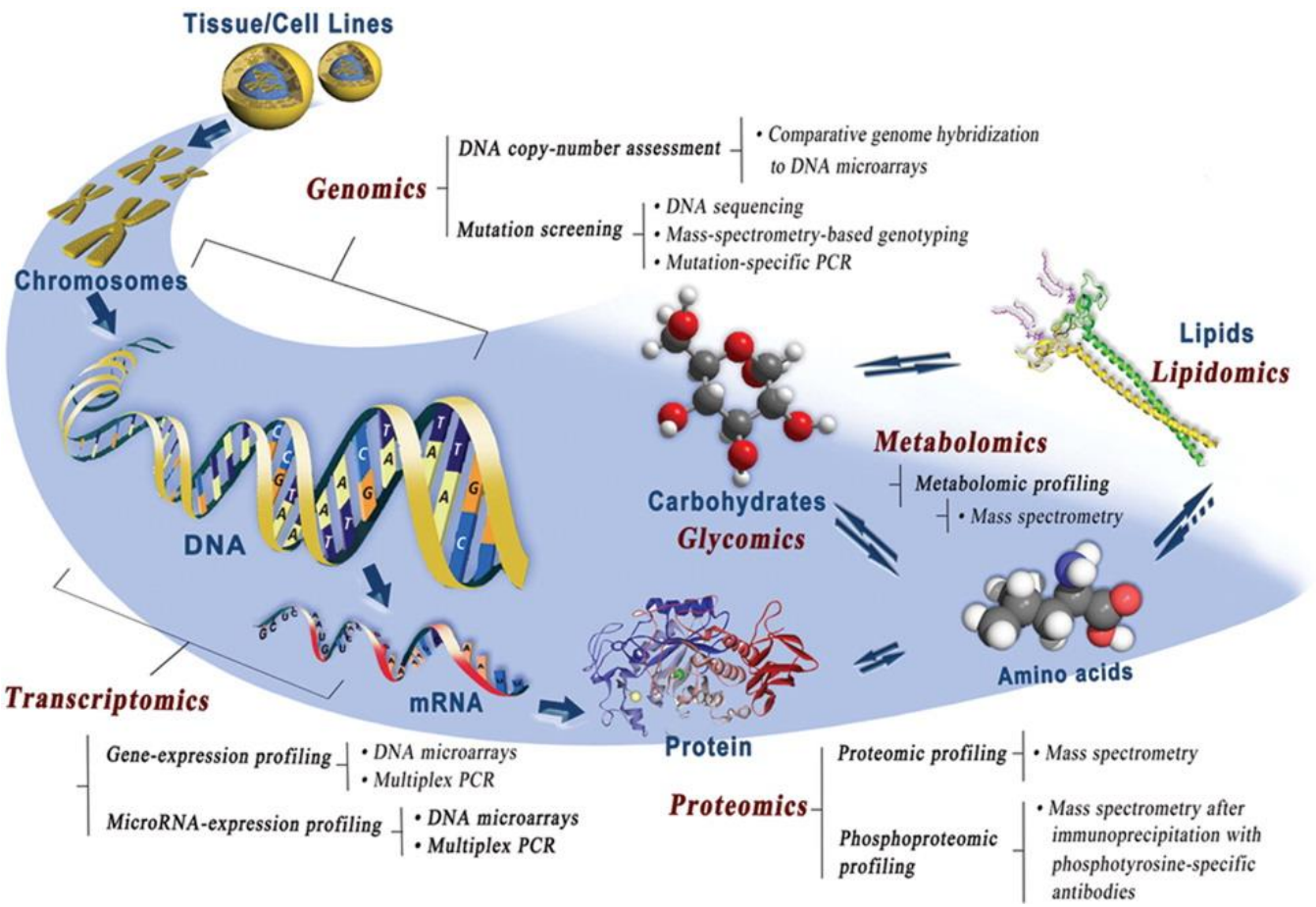


Human genome content

Patrushev & Minkevich, 2008, Biochemistry (Moscow)

What is genomics?

- Advances in genomics have triggered a revolution in discovery-based research and **systems biology**
  - An interdisciplinary field of study that focuses on **complex interactions between components of biological systems**, and **how these interactions give rise to the function and behaviour of that system as a whole.**

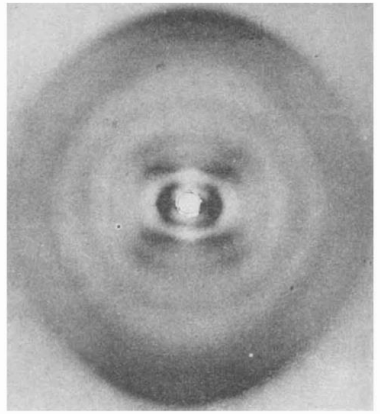


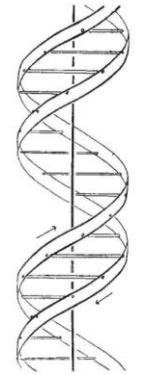
Wu et al., 2011, Journal of Dental Research

Sequencing technologies

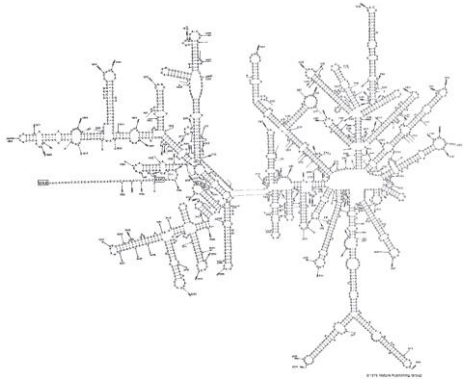
Advancement in sequencing technologies allows genomics to flourish

A brief history of nucleotide sequencing





This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis



1976
First whole R
phage MS2-R
Walter Fiers

80 years

10 years

1860

1869
(Accidental!) **discovery of DNA**
Friedrich Miescher

1870

1950

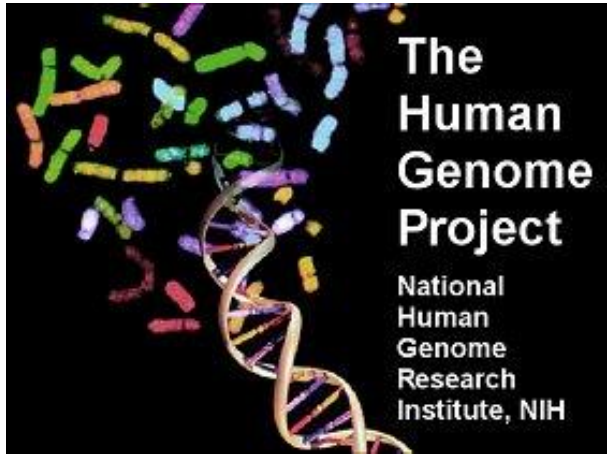
1953
DNA structure
Rosalind Franklin VS Maurice Wilkins
James Watson & Francis Crick

1960

1970

1972
First gene sequence
phage MS2 coat protein
Walter Fiers et al

A brief history of nucleotide sequencing



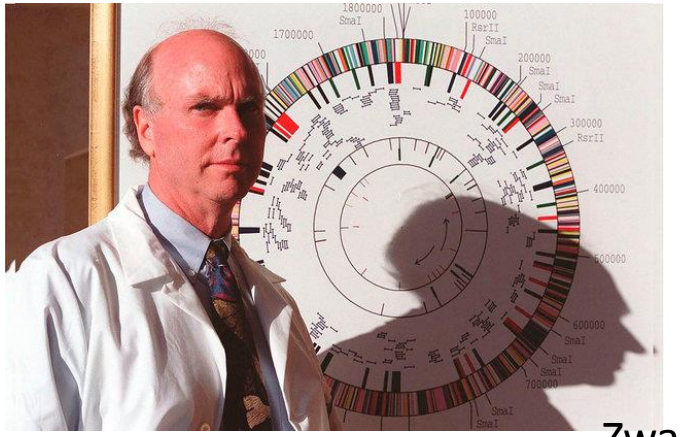
1990-2003
Human Genome Project (3 GB)
$2.7 bn and 13 years
NIH

1998
First multicellular (animal) genome
C. Elegans (97 Mb)
C. elegans Sequencing Consortium

1980 | 1990 | 2000 | 2010

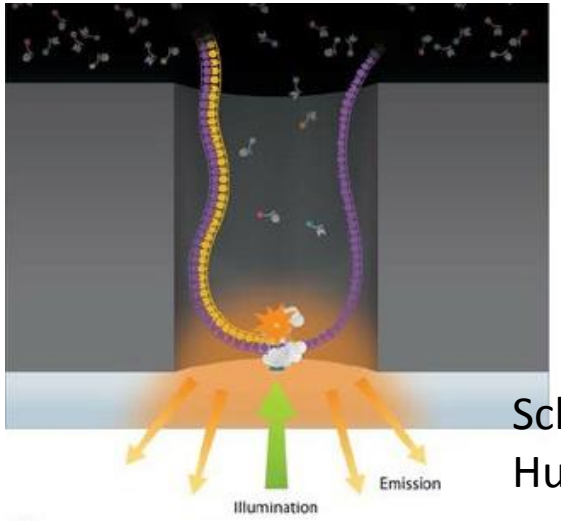
1986
Automated sequencing machine
Leroy Hood
Applied Biosystems

1995
First bacterial genome
H. Influenzae (1.8 Mb)
Craig Venter



Zwart, 2018, IJJS

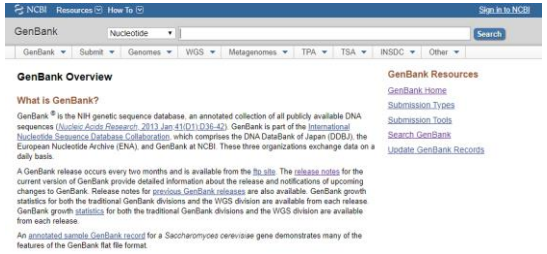
A brief history of nucleotide sequencing



2011
Third generation sequencing
Pacific Biosciences of California

Schadt et al., 2010,
Human Molecular Genetics

10/2020
698,688,094,046 bases
219,055,207 sequences
traditional GenBank records



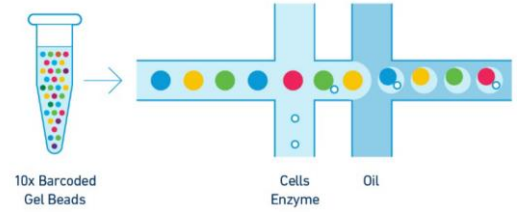
2000

2010

2020

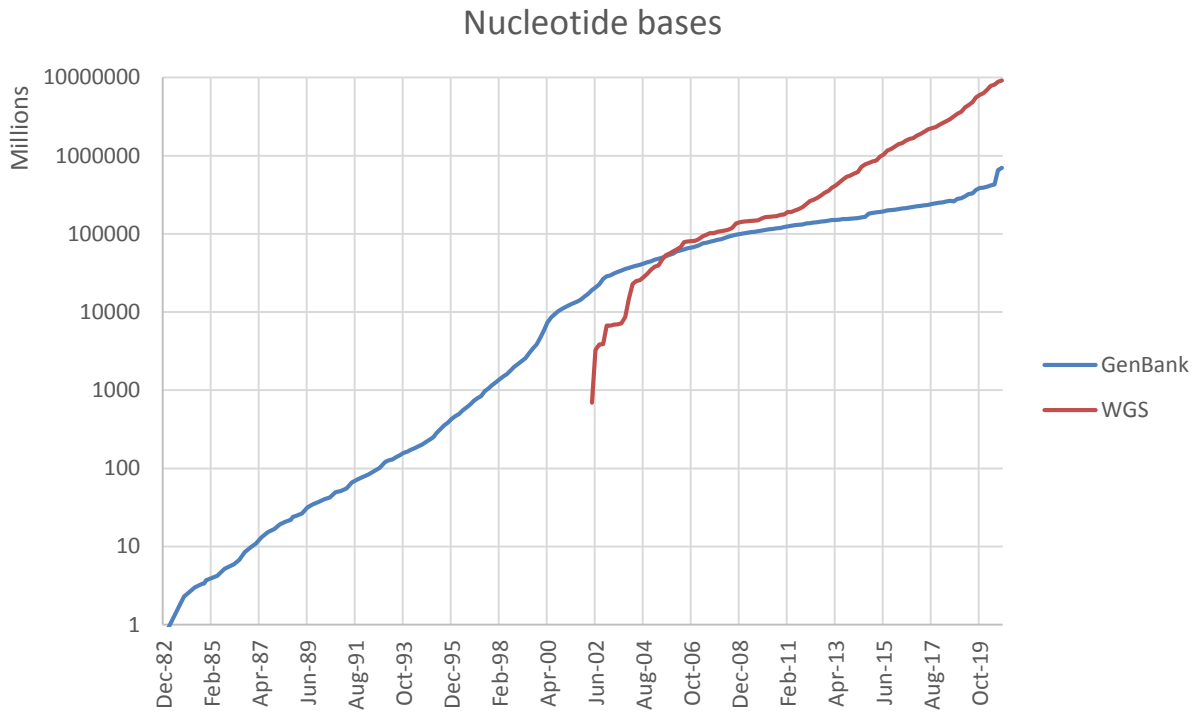
2004
Second generation sequencing
454 pyrosequencer
Roche

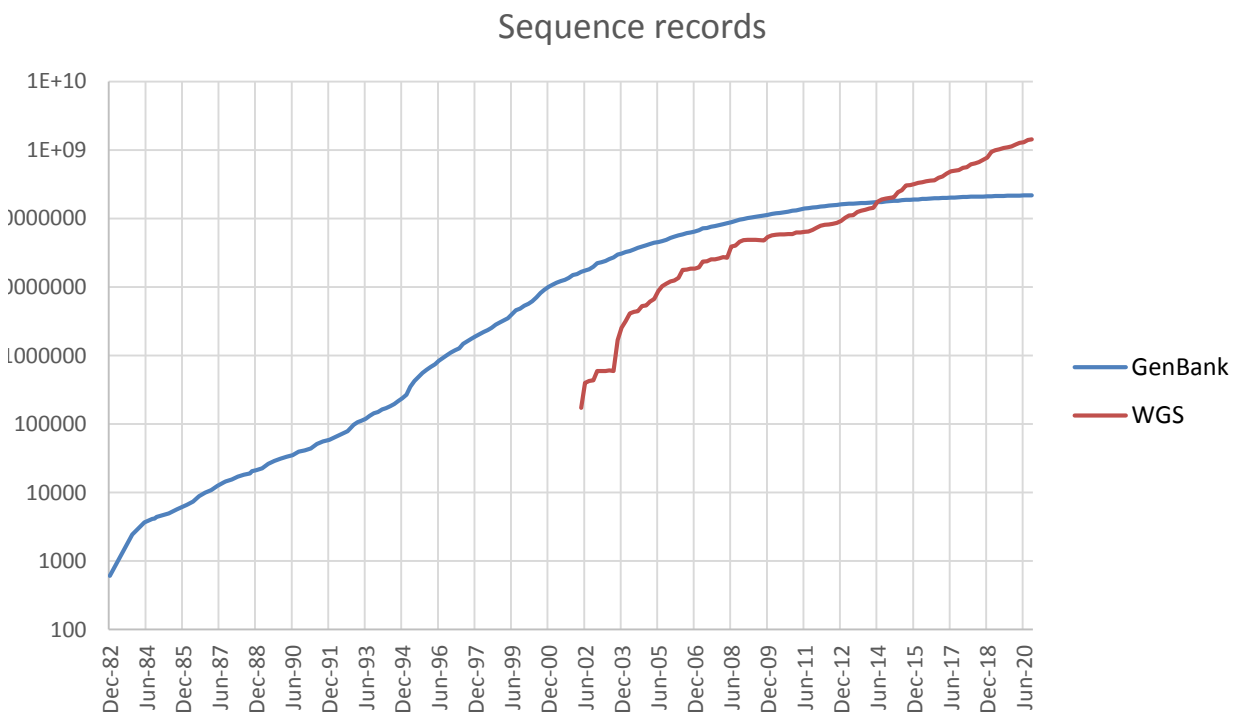
2013
Single cell sequencing
Method of the year
Nature Publishing Group



Tekman et al., 2019, https://galaxyproject.github.io/

Explosive growth of nucleotide data

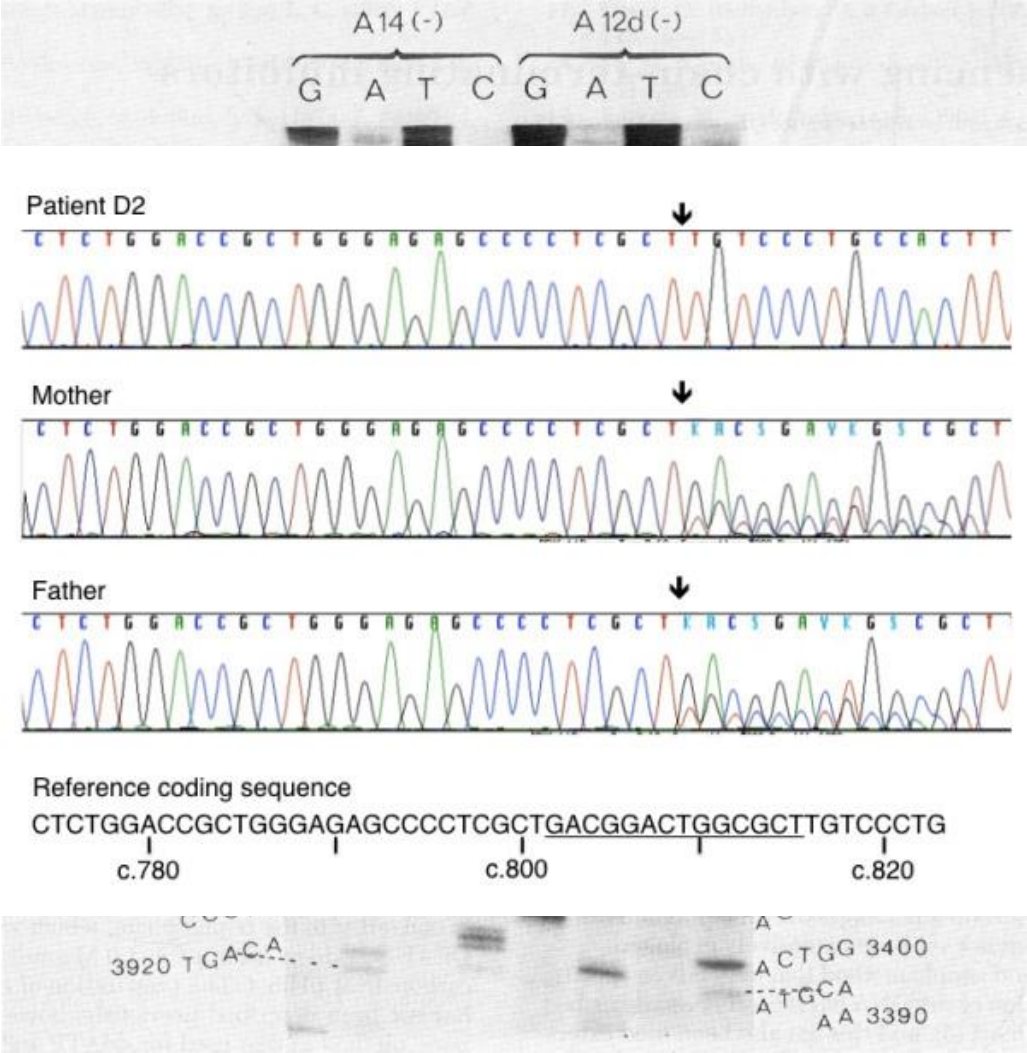




10/2020
NCBI

1st generation sequencing

- Sanger sequencing
  - <900 bp (500-600bp without enrichment)
  - Highly accurate
  - Relatively expensive / base
  - Time consuming and not scalable
  - Require knowing *a priori* what sort of sequences you will be sequencing



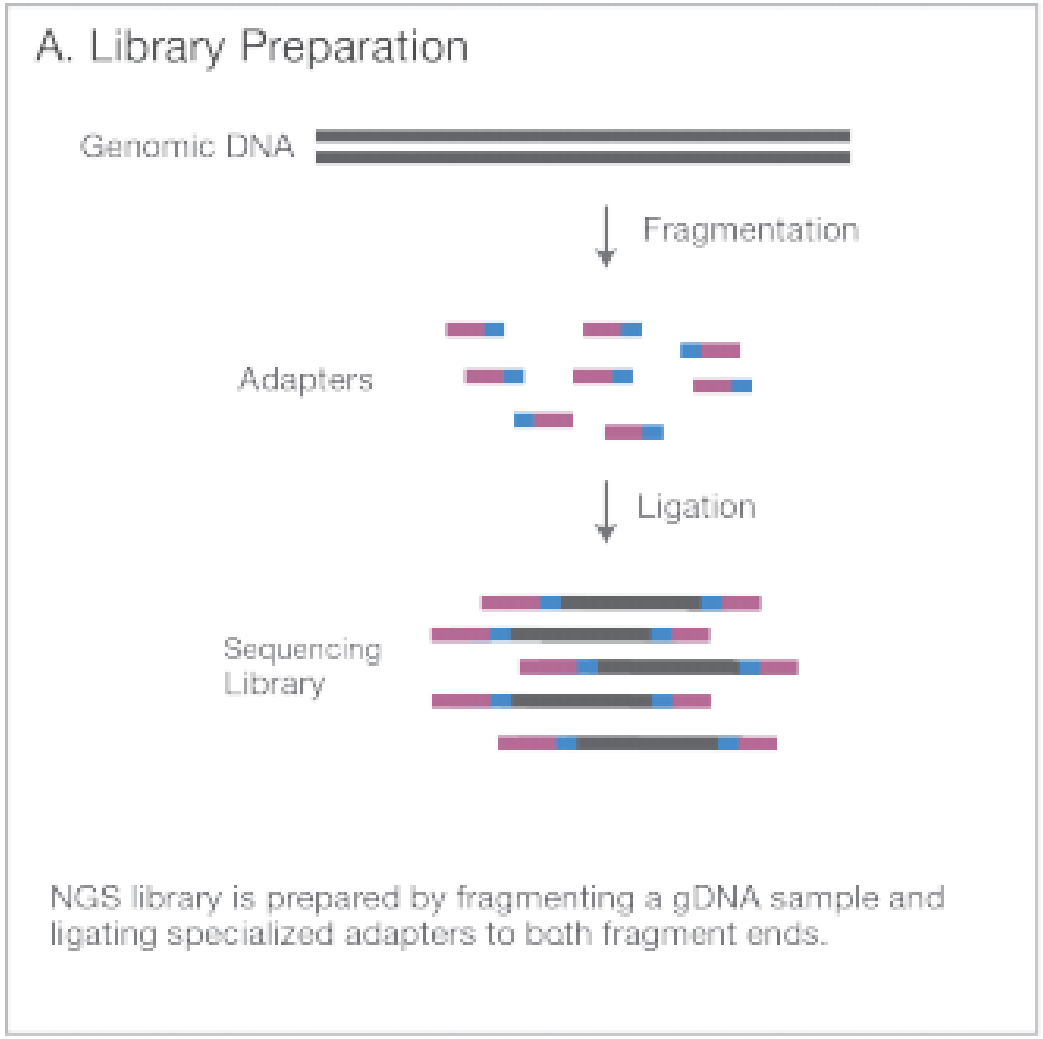
Levene et al, 2012, BMC Medicine

Sanger, Nicklen, & Coulson, 1977, PNAS

2nd generation sequencing

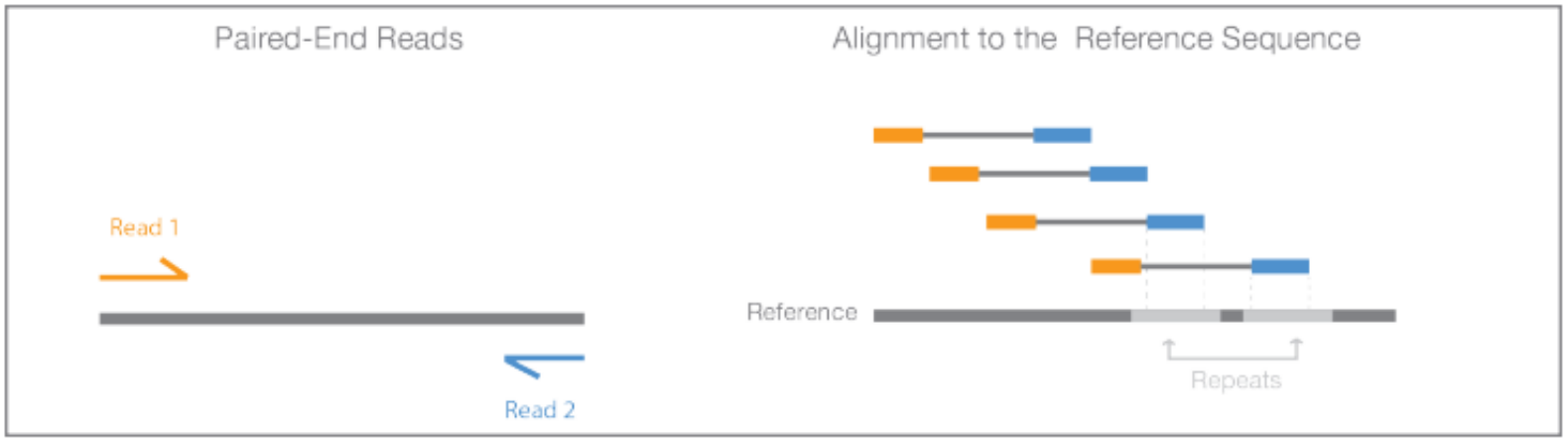
- Basic characteristics
  - Generating of millions of short and accurate **reads** (150-300 bp) in parallel
  - Low cost / base
  - You don't need to know anything about the sequences

  - 454 pyrosequncing
  - Ion torrent sequencing
  - **Illumina sequencing**
  - SOLiD sequencing



Illumina Inc, 2017, An introduction to Next-Generation Sequencing Technology

2nd generation sequencing



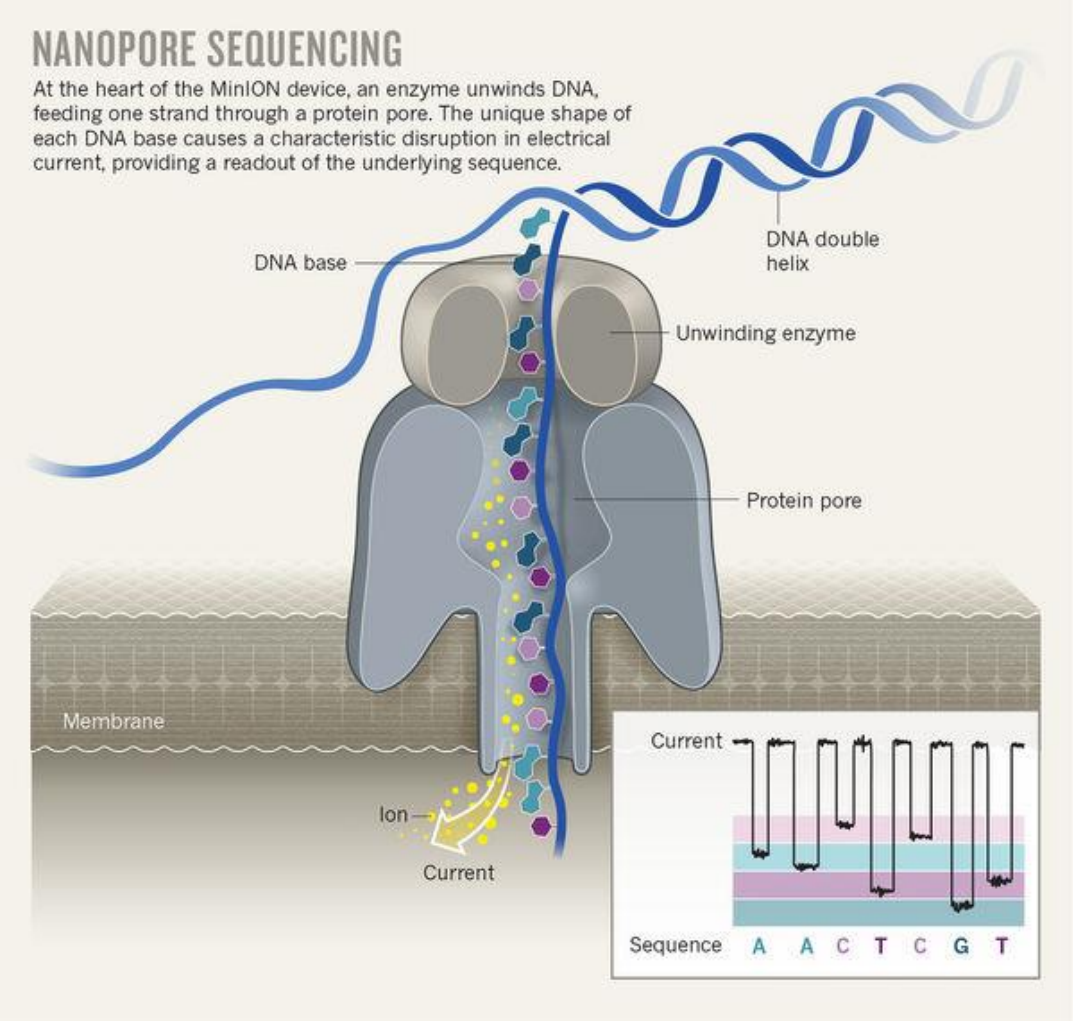
- **Illumina (MiSeq)**
  - ~150 bp/read
  - 99.9% accuracy
  - ≤ 1,000 Bbp/run (HiSeq)
  - < $5-150/1bn bp

Illumina Inc, 2017, An introduction to Next-Generation Sequencing Technology

3rd generation sequencing

- Basic characteristics
  - Produce very long reads (long-read sequencing)
  - Low cost / base
  - Low accuracy
  - No knowledge is required about the sequences
  - **Single molecule sequencing**

3rd generation sequencing



Perkel, 2017, Naturejobs

Single molecule real time sequencing (SMRT)

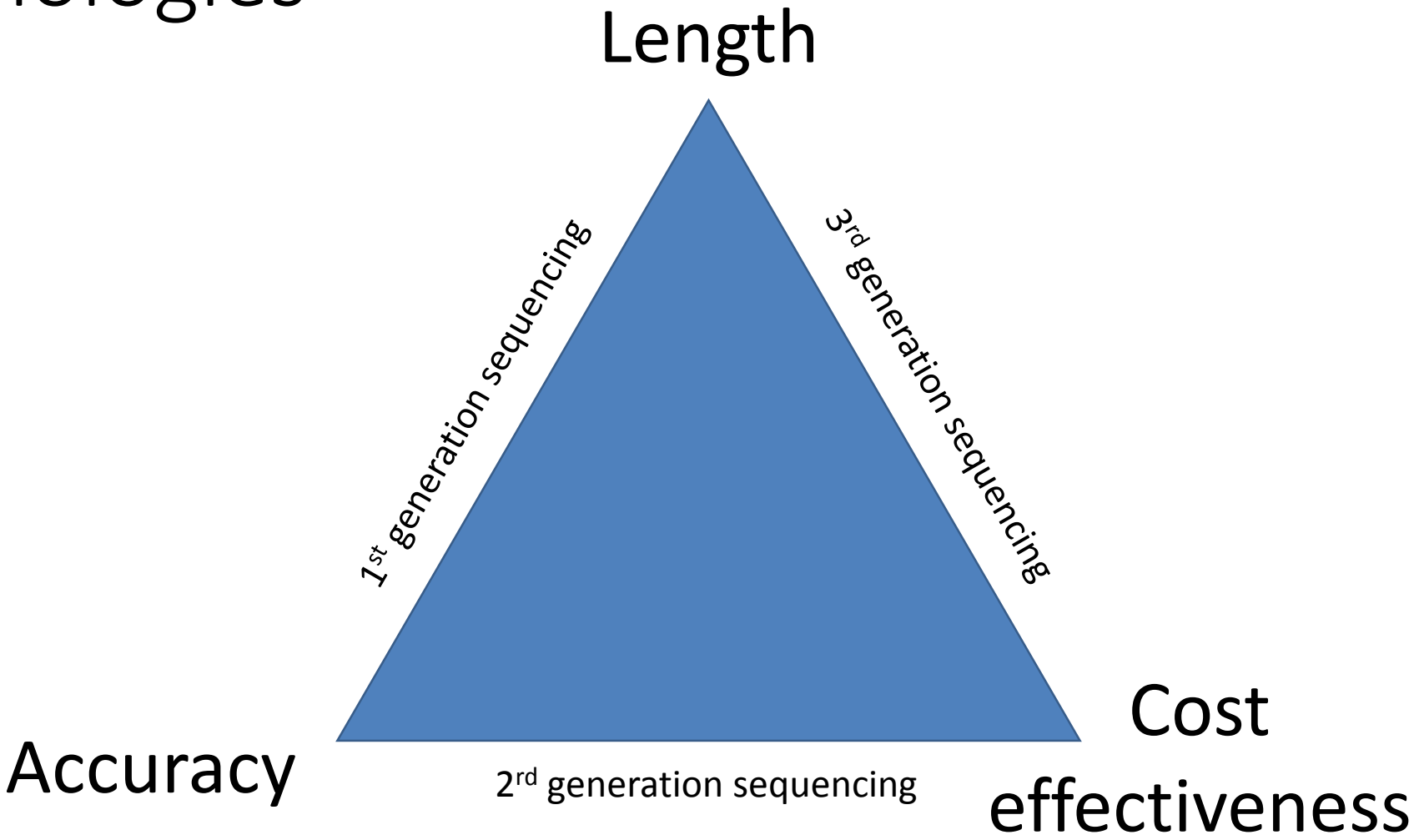
Whole genome/read (depend on Lib. Prep.)

92-97% accuracy

[Not too high] bp/run (depend on Lib. Prep.)

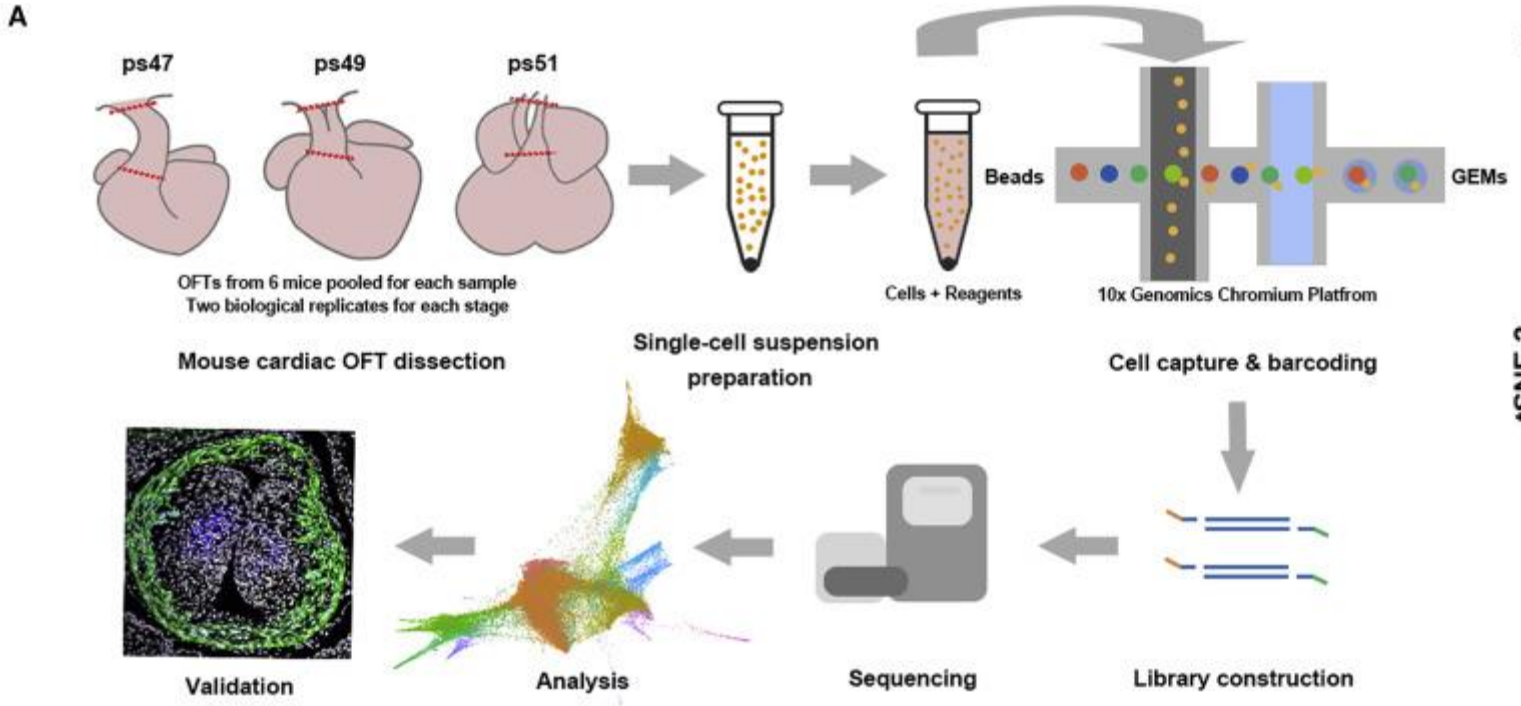
$7-100/1bn bp

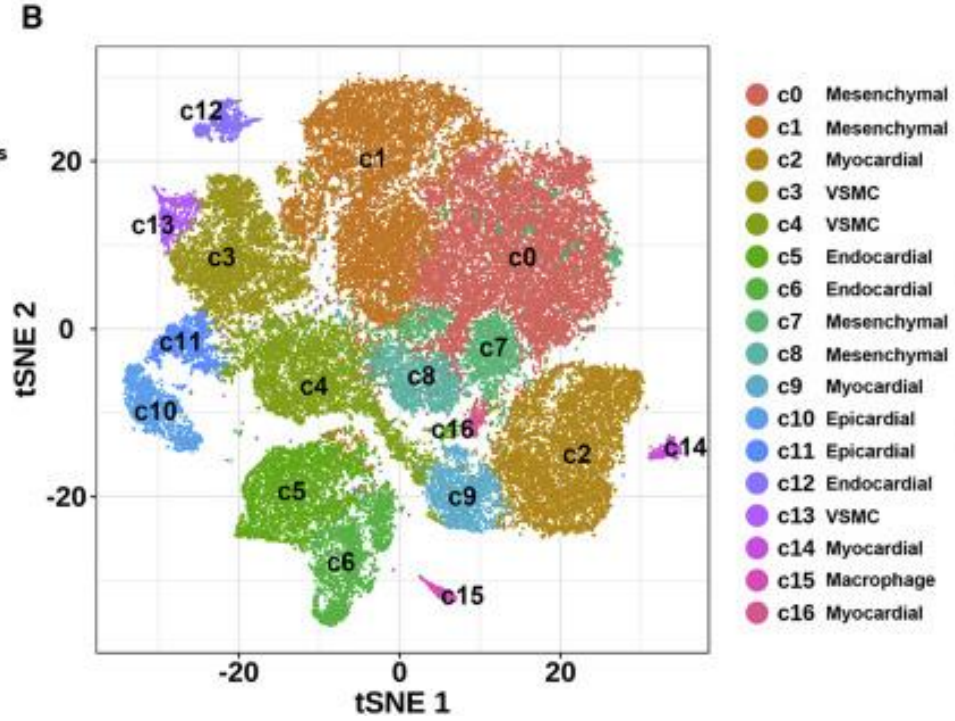
Comparison between the three sequencing technologies



Some cool sequencing tricks

- Single cell sequencing
  - Single cell isolation + SGS = examines the sequence information from individual cells

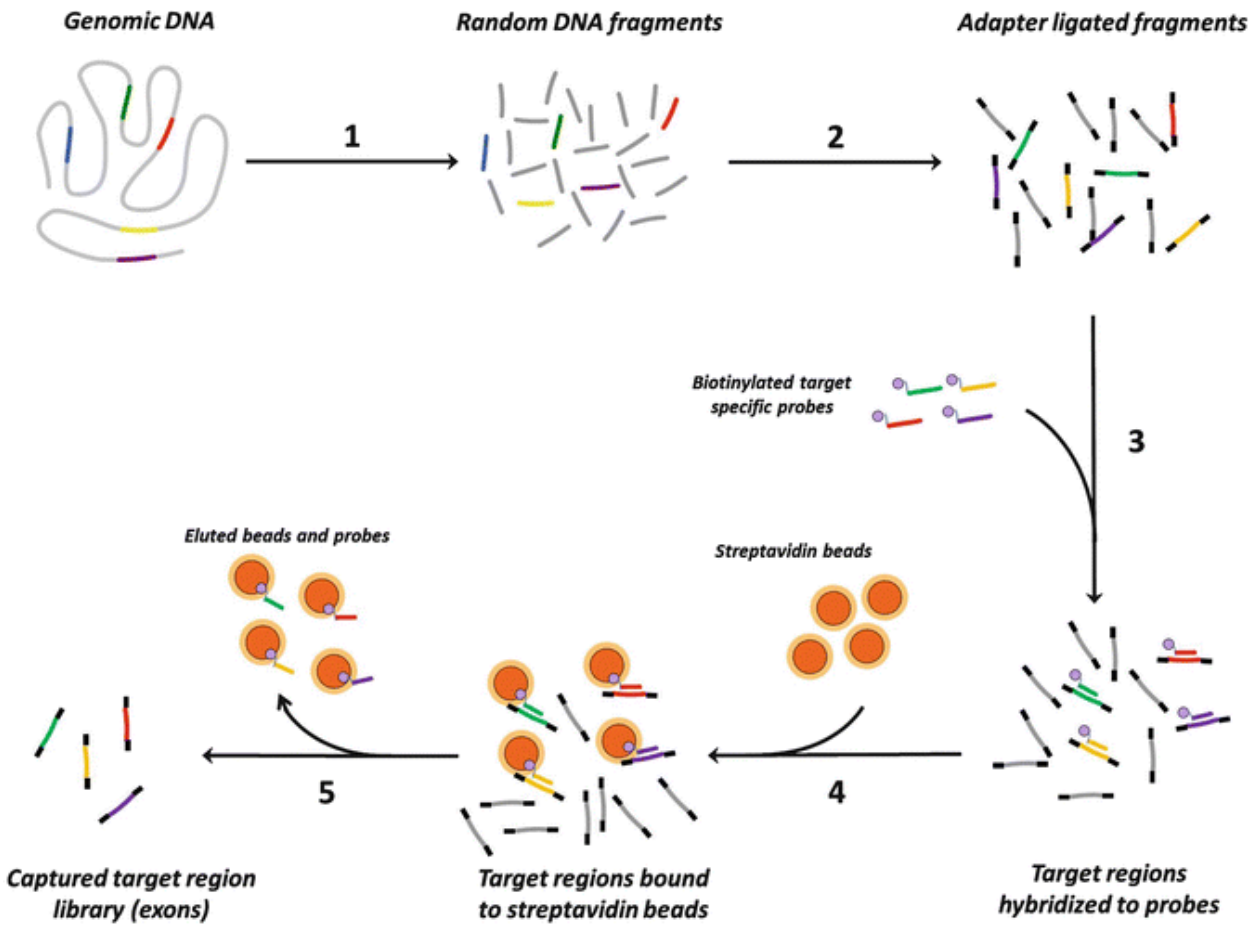




Lui et al., 2019, Cell Reports

Some cool sequencing tricks

- Target/probe enrichment sequencing
  - Sequence capture array (e.g. exon array) + SGS = exome sequencing



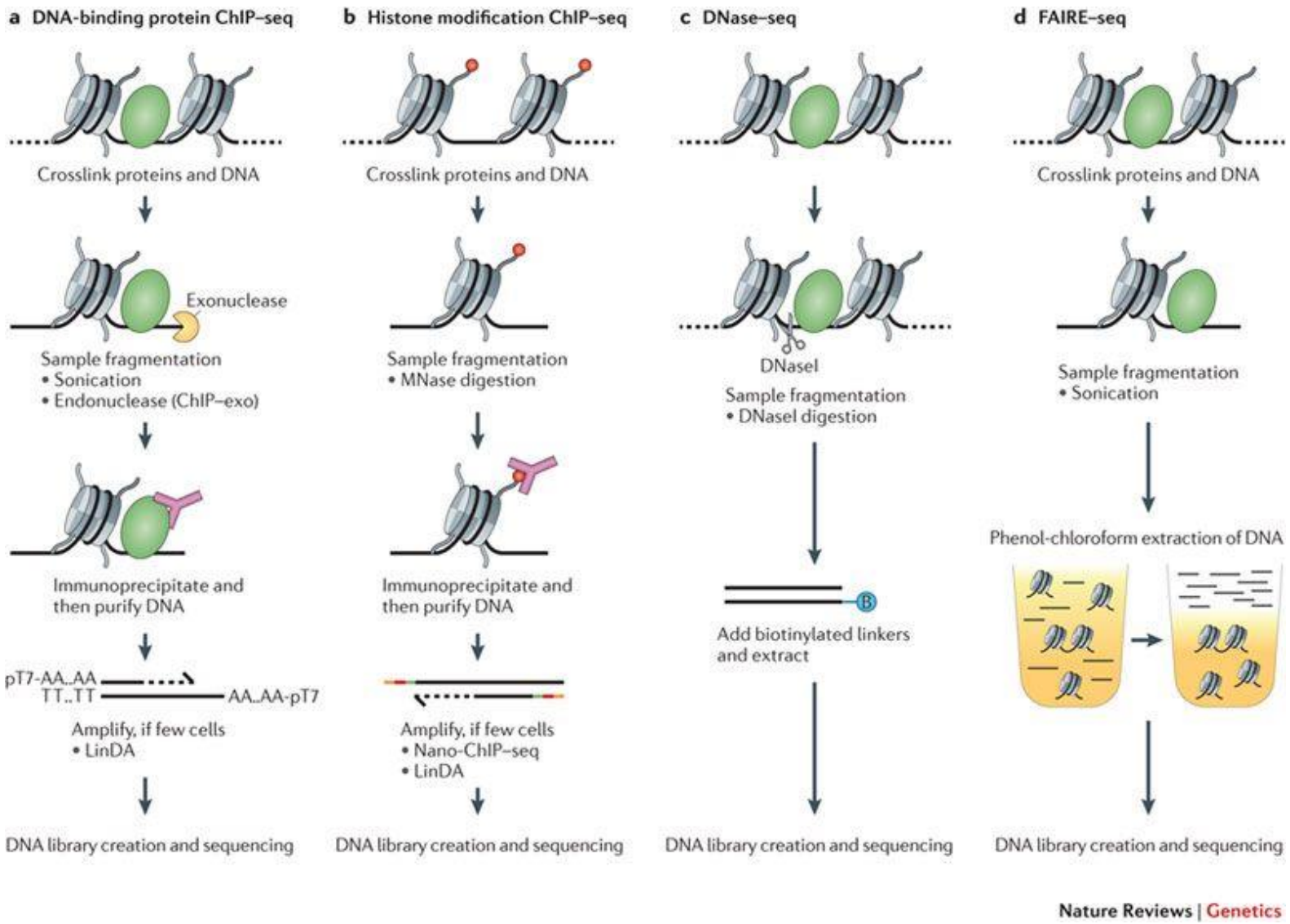
Ankala & Hegde, 2014, Genomic Applications in Pathology

Some cool sequencing tricks

- Chip-seq
  - Antibodies specific to your DNA-binding proteins of interest

    + immunoprecipitation

    + SGS = Chip seq



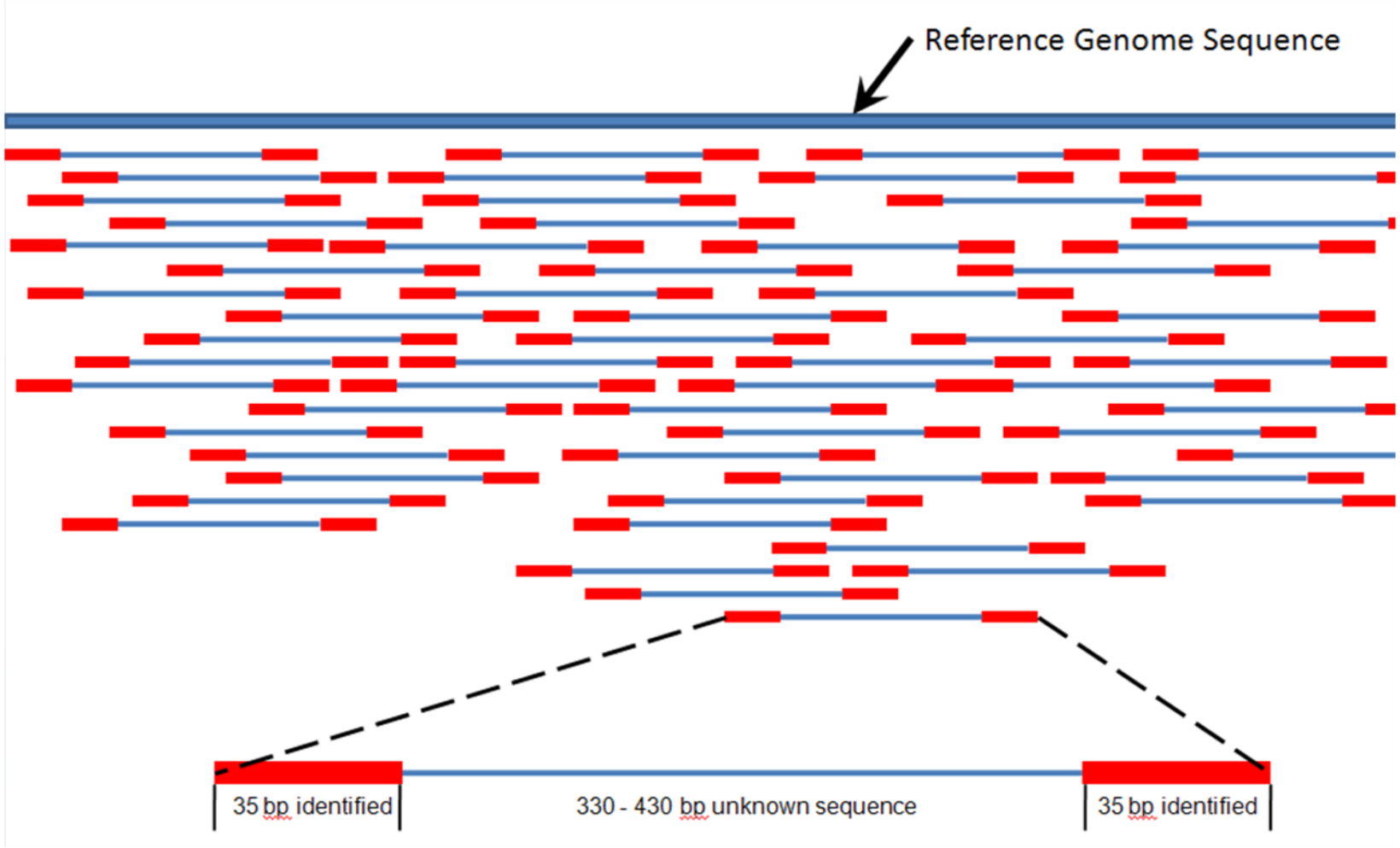
Furey, 2012, Nature Reviews Genetics

What can we do with whole genome data?

Key analyses techniques in genomics

Genome mapping (resequencing)

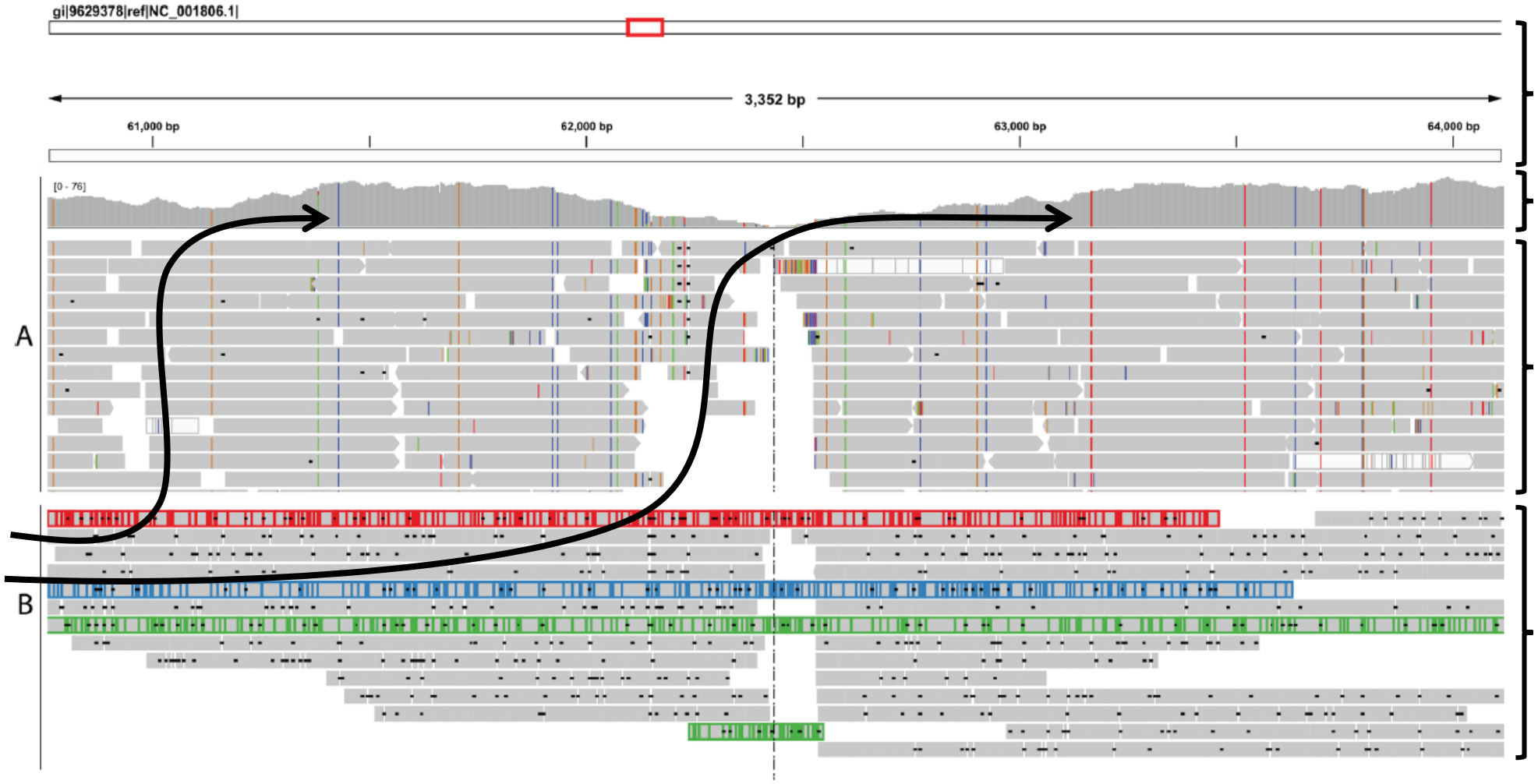
Read mapping diagram



Suspencewl, 2011, https://en.wikipedia.org/wiki/Genomics

Single nucleotide variant calling

Read mapping diagram



Reference genome

Read mapping depth

Read mapping

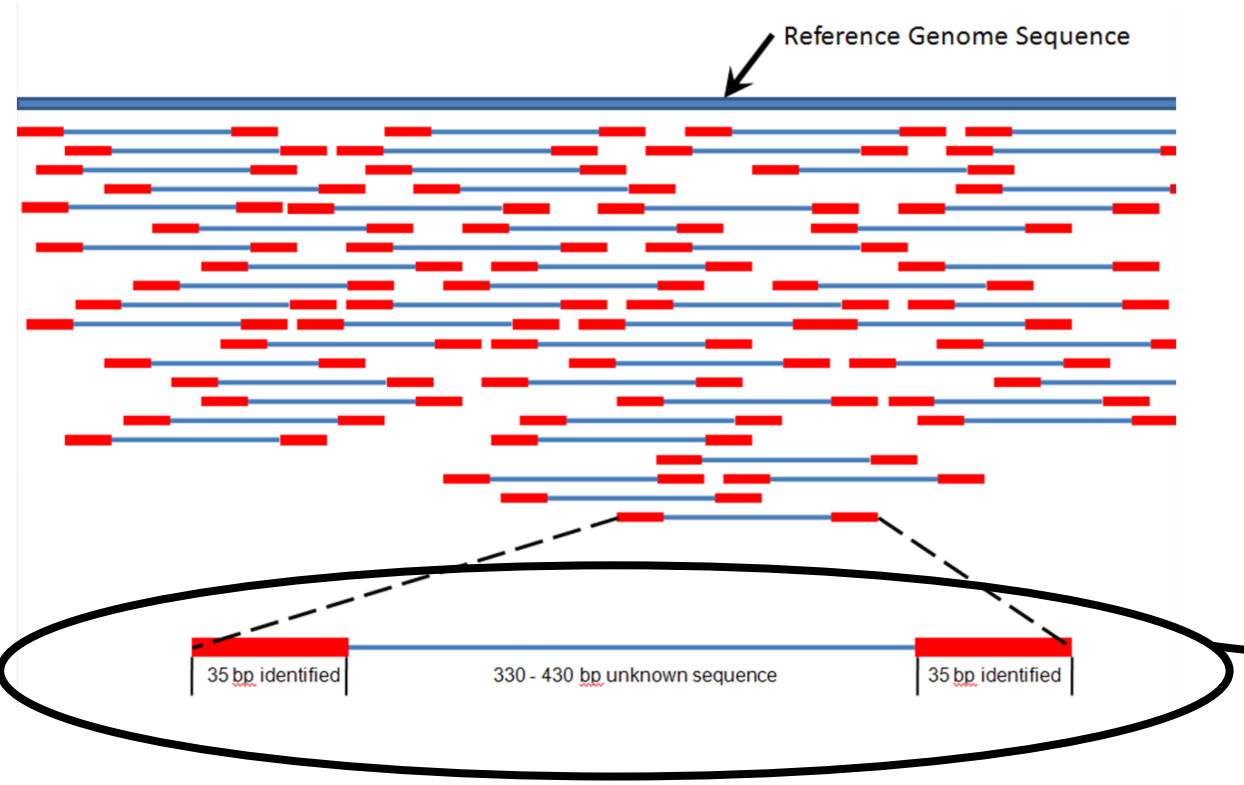
single nucleotide variants

Karamitros et al., 2016, PLOS ONE

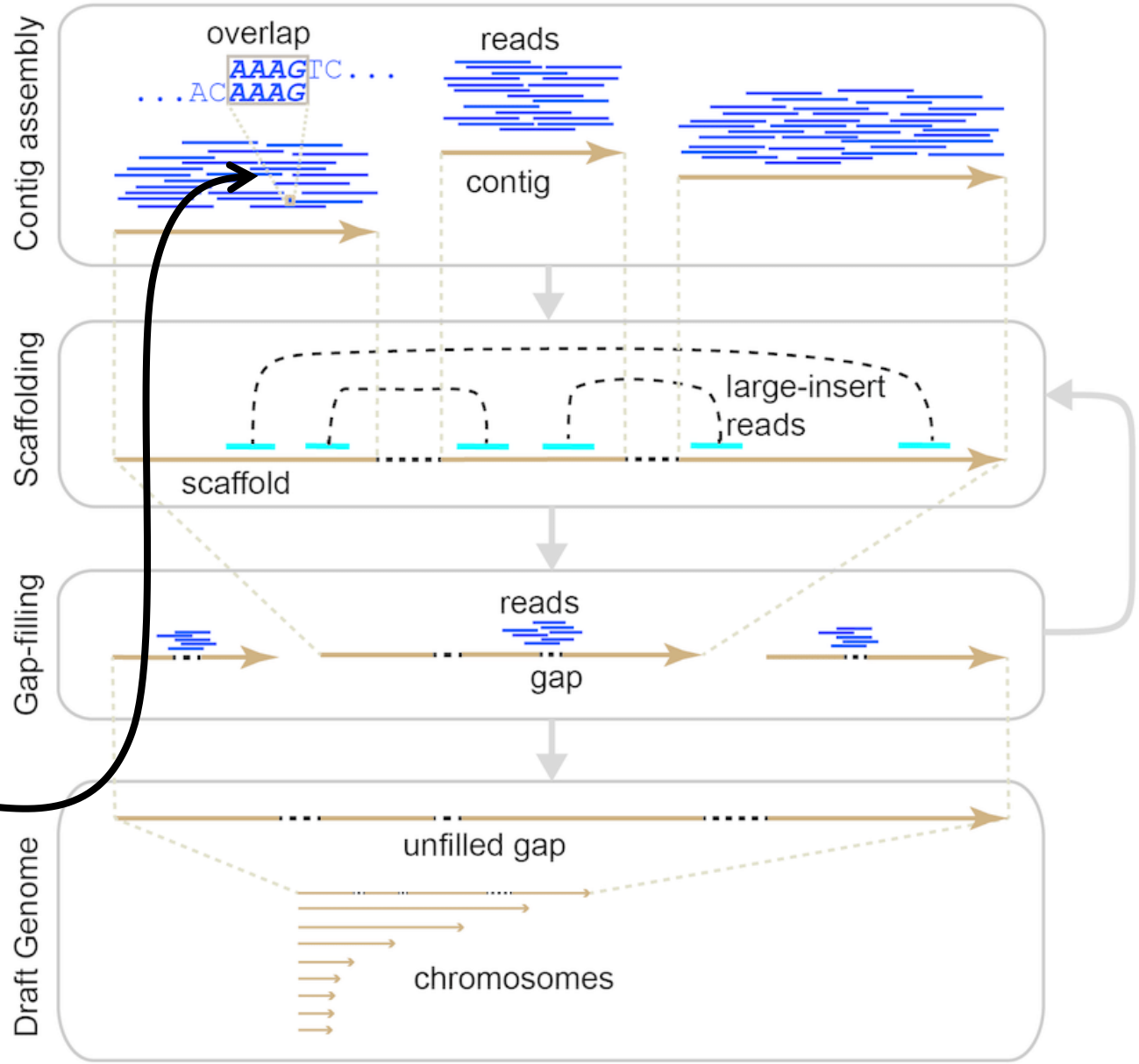
(*De novo*) genome assembly

```
         1                                                                    86
Read 2   GAAGAAGAGGTAGTAGTTAGGTCTAAA
Read 3             TAGTAGTTAGGTCTGAAAATTTCTCAAACAA
Read 4                        CTGAAAATTTCTCAAACAATGCAAAAACCATAATAGTACAG
Read 5                                       CAATGCAAAAACCATAATACTACAGCTGACGGAAGCTGTAG
Read 1                                                     ATAGTACAGCTGACGGAAGCTGTAGAAATTAA...



Contig   GAAGAAGAGGTAGTAGTTAGGTCTGAAAATTTCTCAAACAATGCAAAAACCATAATAGTACAGCTGACGGAAGCTGTAGAAATTAA...
```

(*De novo*) genome assembly



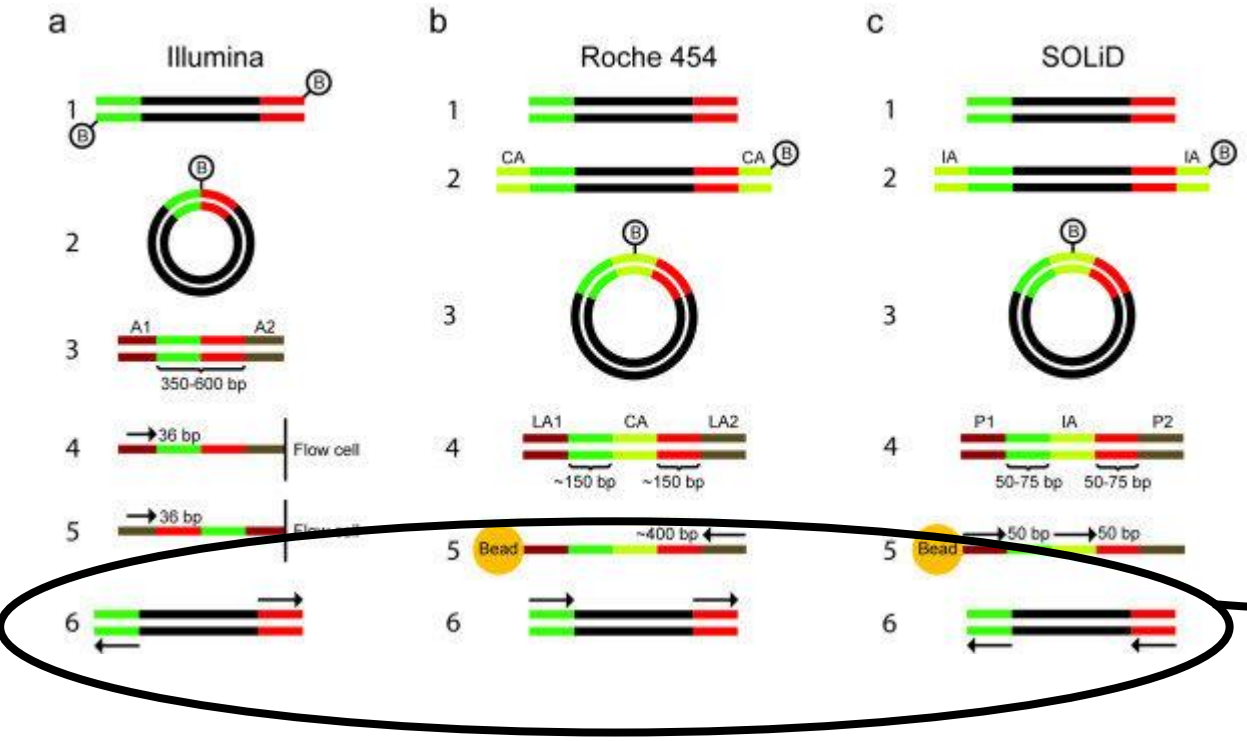
Paired end read



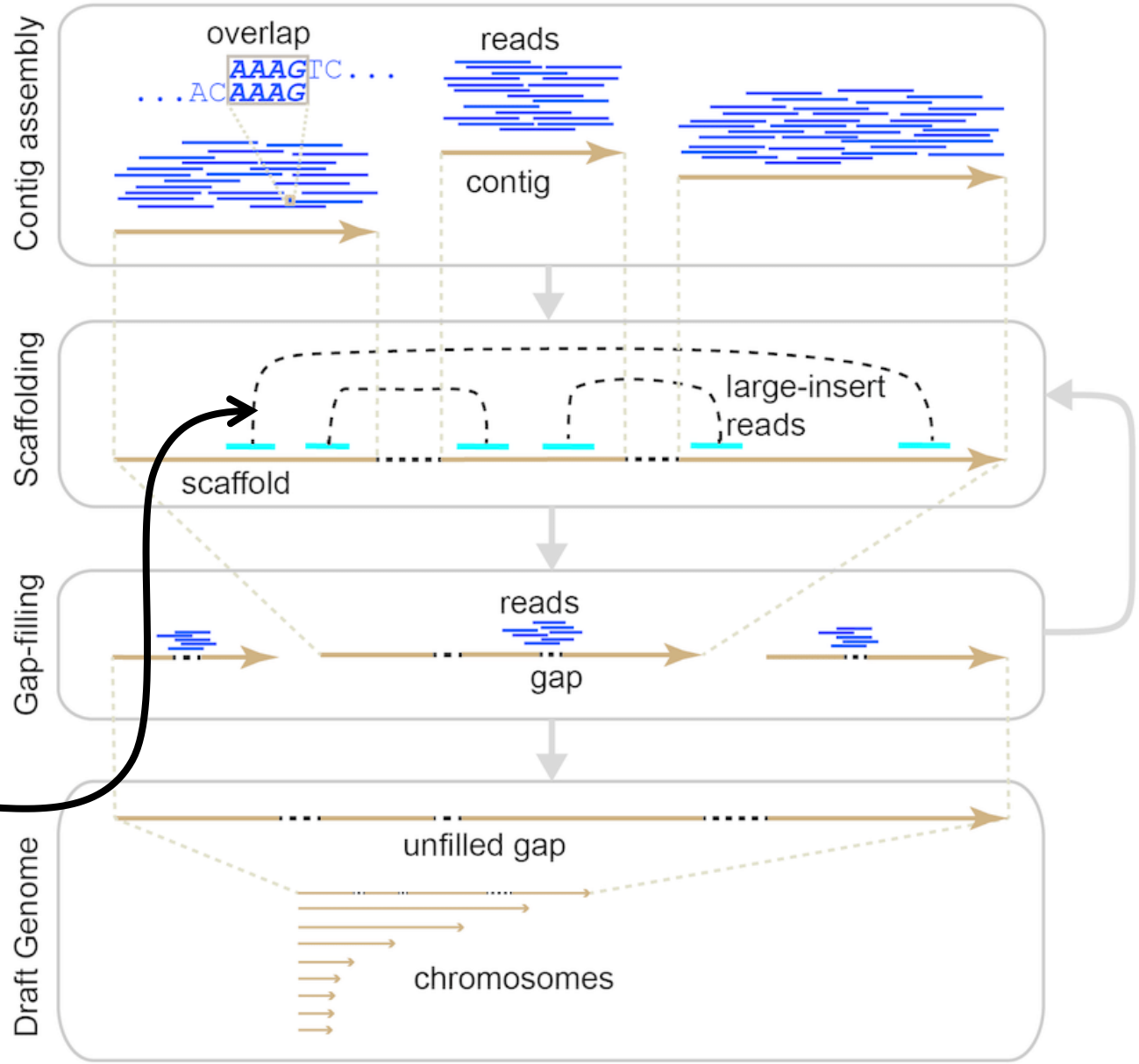
Sohn & Nam, 2016, Briefings in Bioinformatics

(*De novo*) genome assembly

mate-pair sequencing libraries



Paired end read



Berglund et al., 2011, Investigative Genetics

Sohn & Nam, 2016, Briefings in Bioinformatics

Genome annotation

.. Only work with viruses and bacteria

- ORFfinder (finding open reading frame)

Human immunodeficiency virus 1, complete genome

ORFs found: 51 Genetic code: 1 Start codon: 'ATG' only

Six-frame translation...

ORF7 (912 aa) Display ORF as... Mark

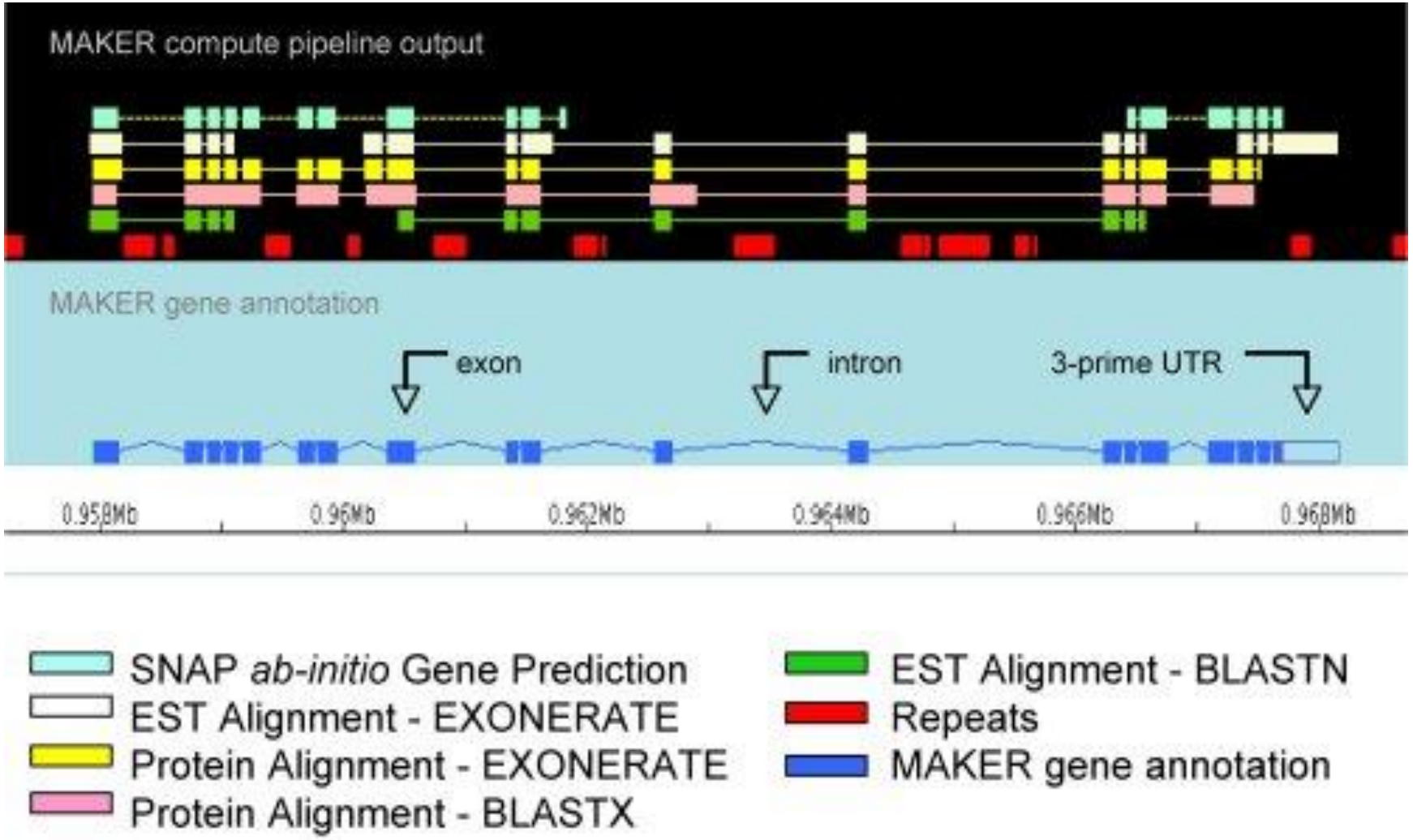
```
>lcl|ORF7
MSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNII
GRNLLTQIGCTLNFPISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVE
ICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQD
FWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPLDEDFRKYTAFTIPSIN
NETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDD
LYVGSDLEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFLWMGYELHP
DKWTVQPIVLPEKDSWTVNDIQKLVGKLNWASQIYPGIKVRQLCKLLRGT
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF7	**+**	**2**	**1904**	**4642**	**2739 | 912**
ORF10	+	2	5771	8341	2571 | 856
ORF11	+	3	336	1838	1503 | 500

Genome annotation

- MAKER, for eukaryotic gene prediction

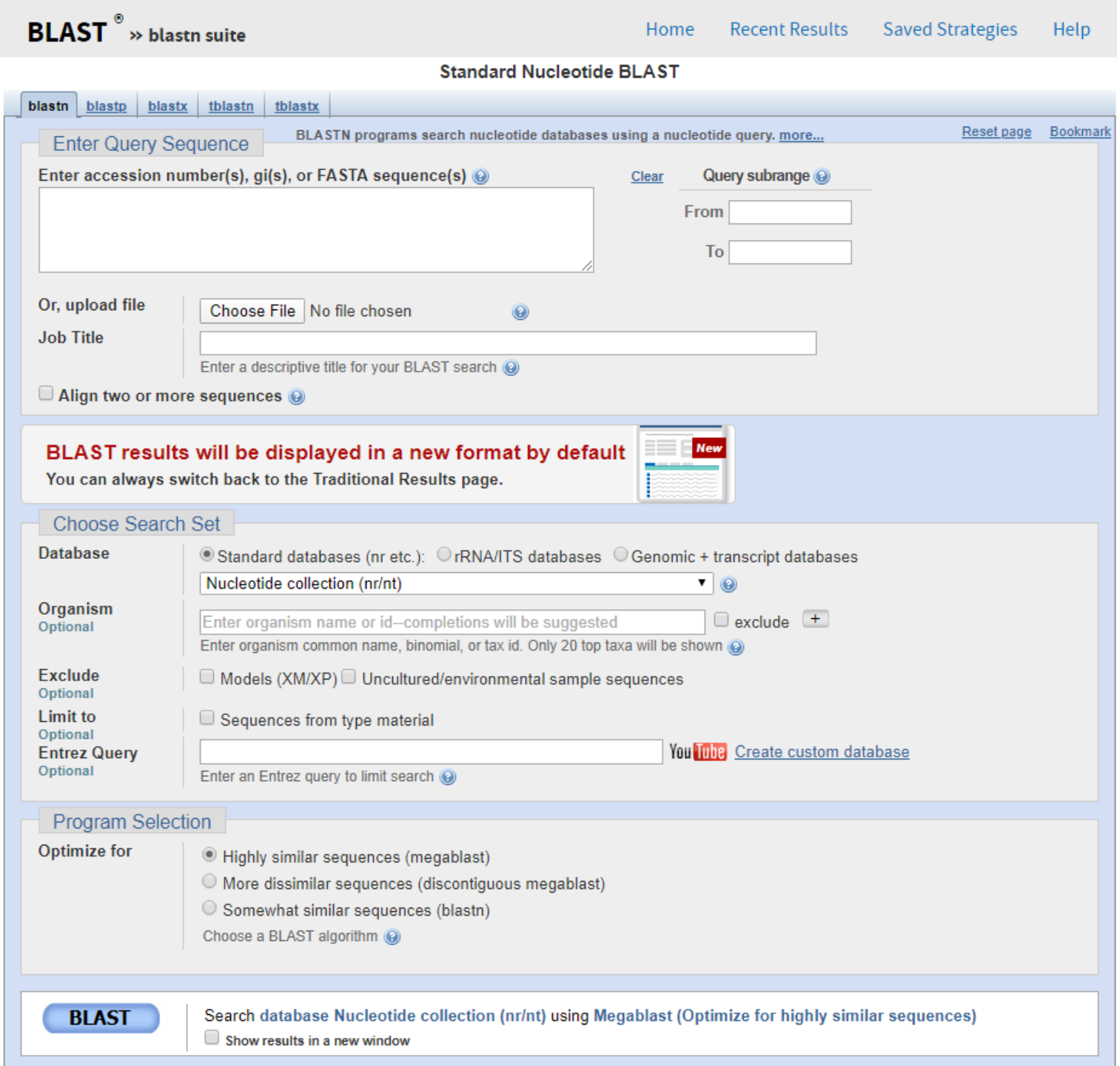


Genome annotation

- BLAST =

- Basic
- Local
- Alignment
- Search
- Tool

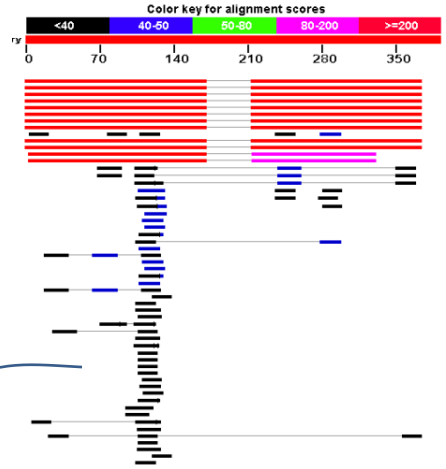
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome



Genome annotation

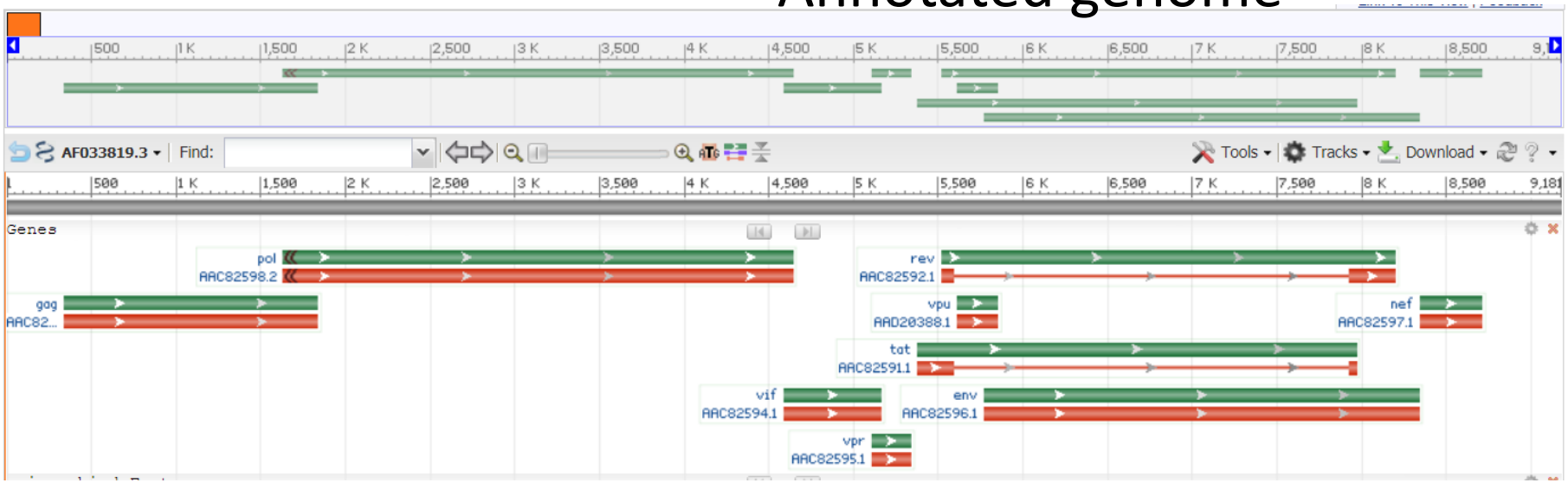
Gene finding

+



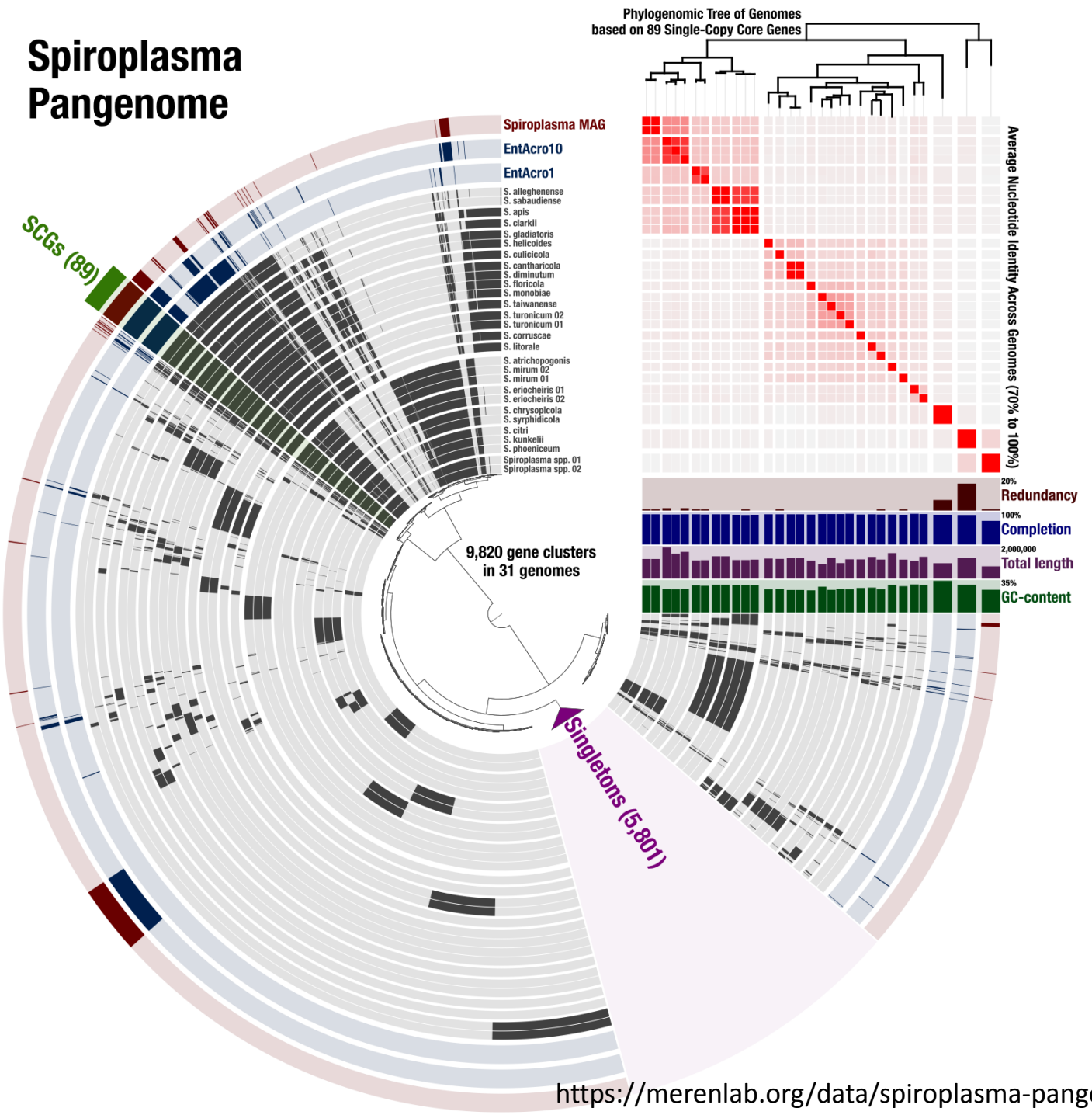
Similarity detection

Annotated genome



Pangenome analysis

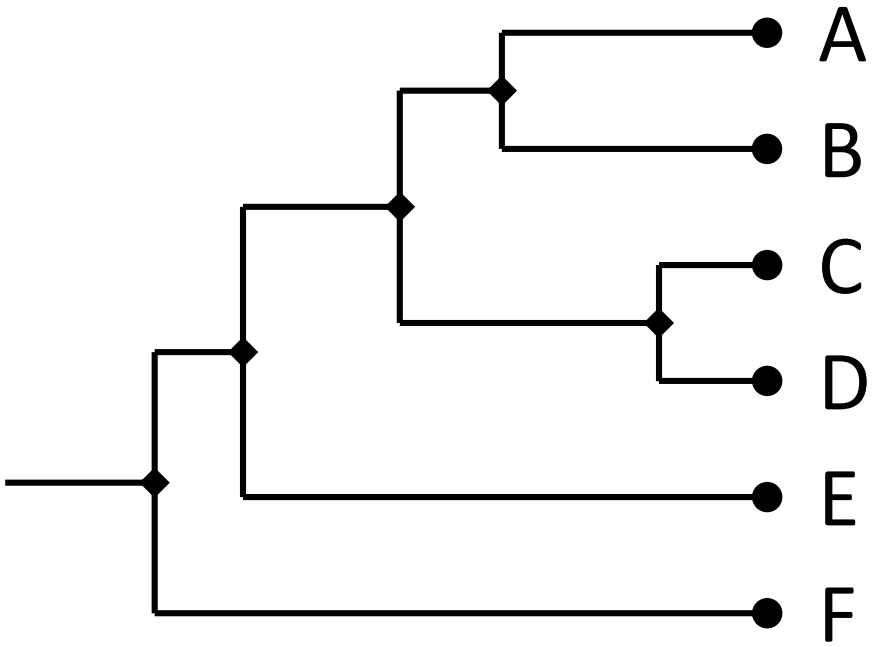
- The **pangenome** refers to a collection of genomic sequence found in the entire species or population rather than in a single individual
- The entire set of **core** and **accessory genes** for all strains within a clade

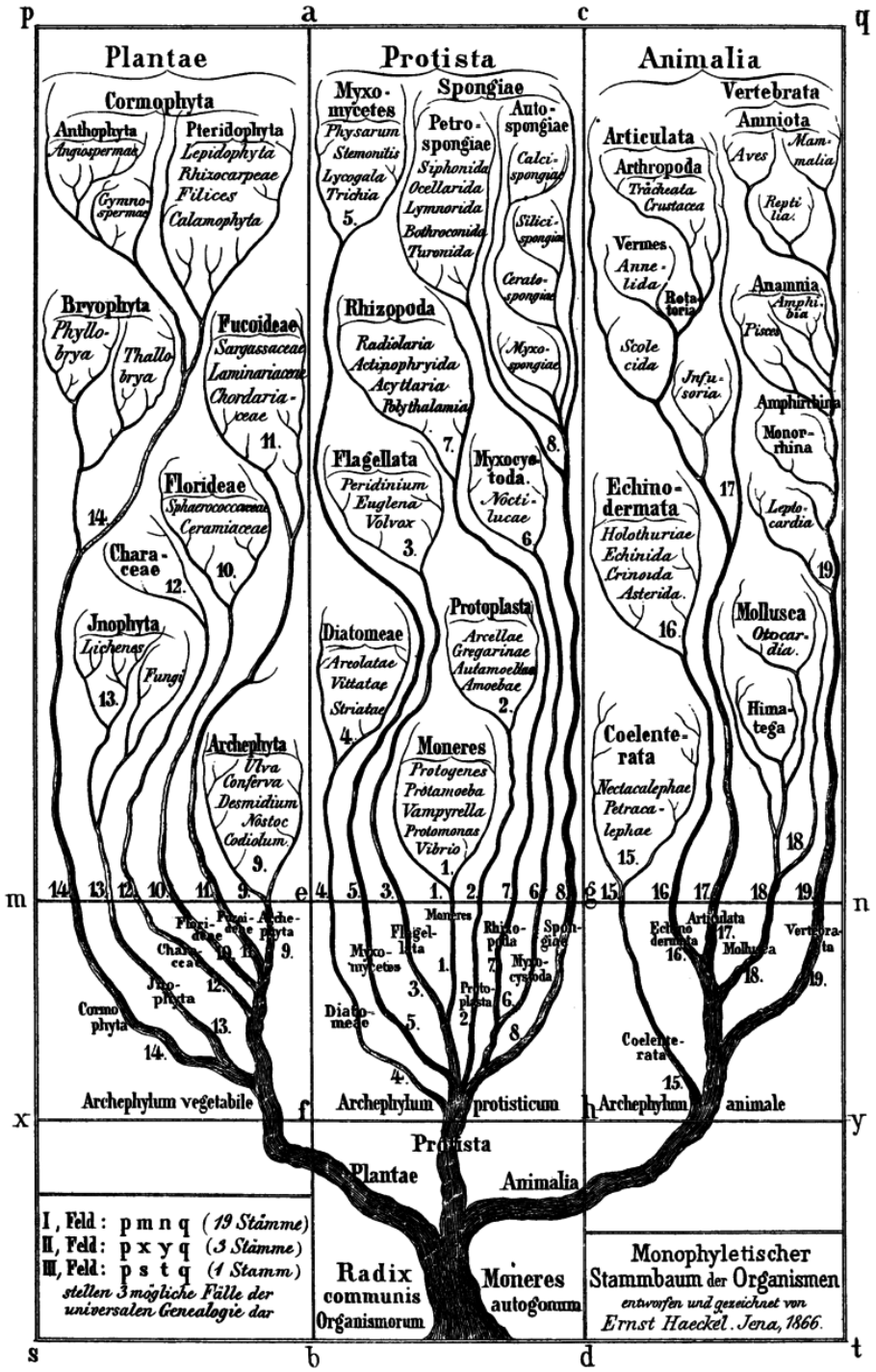


https://merenlab.org/data/spiroplasma-pangenome/

Phylogenetics analysis

- The study of evolutionary histories, or relationships of a set of organisms

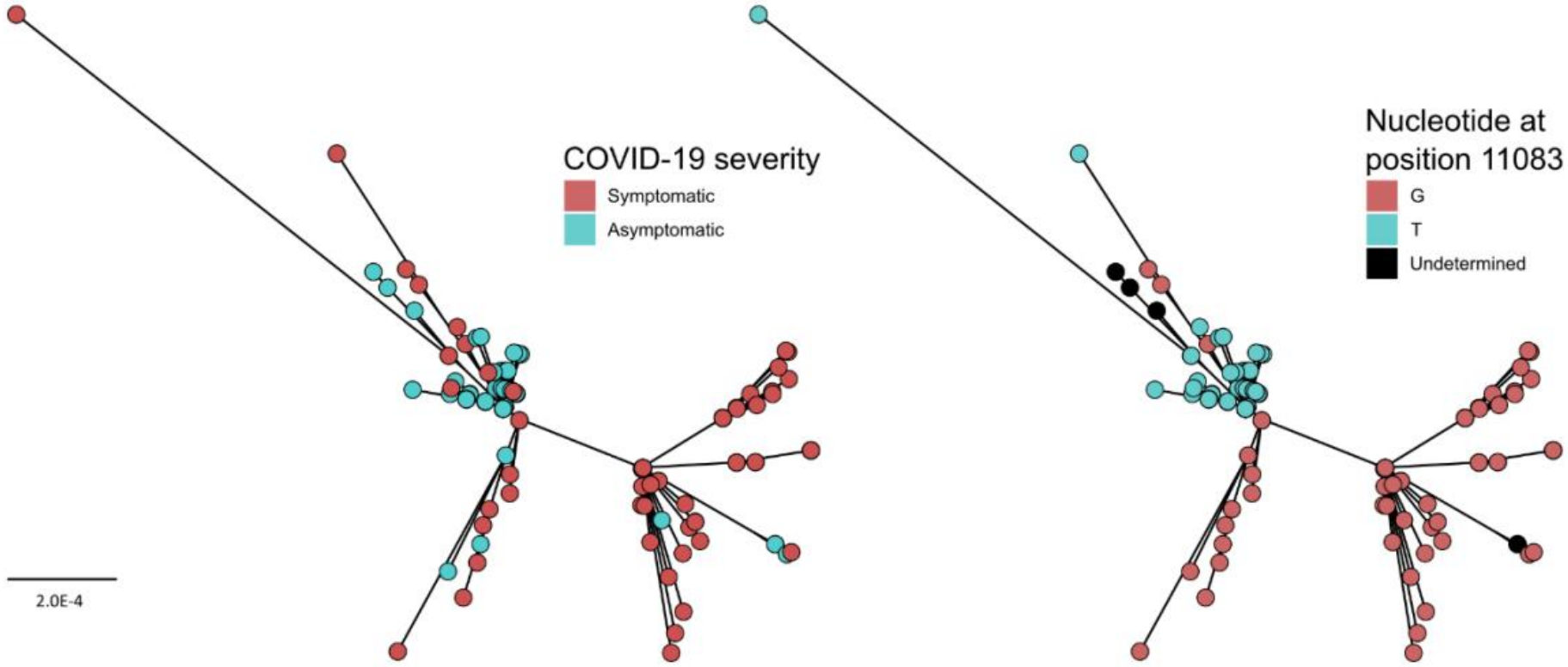




Phylogenetic tree suggested by Haeckel (1866)

Genome wide association study – GWAS

- A study that associates specific genetic variations with particular phenotypes, e.g. diseases



Key research areas

- **Comparative genomics**
  - *"How does a genome compare to other genomes of its close relatives?"*
- **Functional genomics**
  - *"What are the functions of all genomic elements?"*
- **Structural genomics**
  - *"What are the structure of all proteins encoded by this genome?"*
- **Metagenomics**
  - The study of all genetic materials in environmental samples
- **Epigenomics**
  - The study of the complete set of modifications on the genetic material

What are we going to do in this workshop?

Short answer: Systematic biology

Classification of pathogen by using NGS data and making SNP barcodes