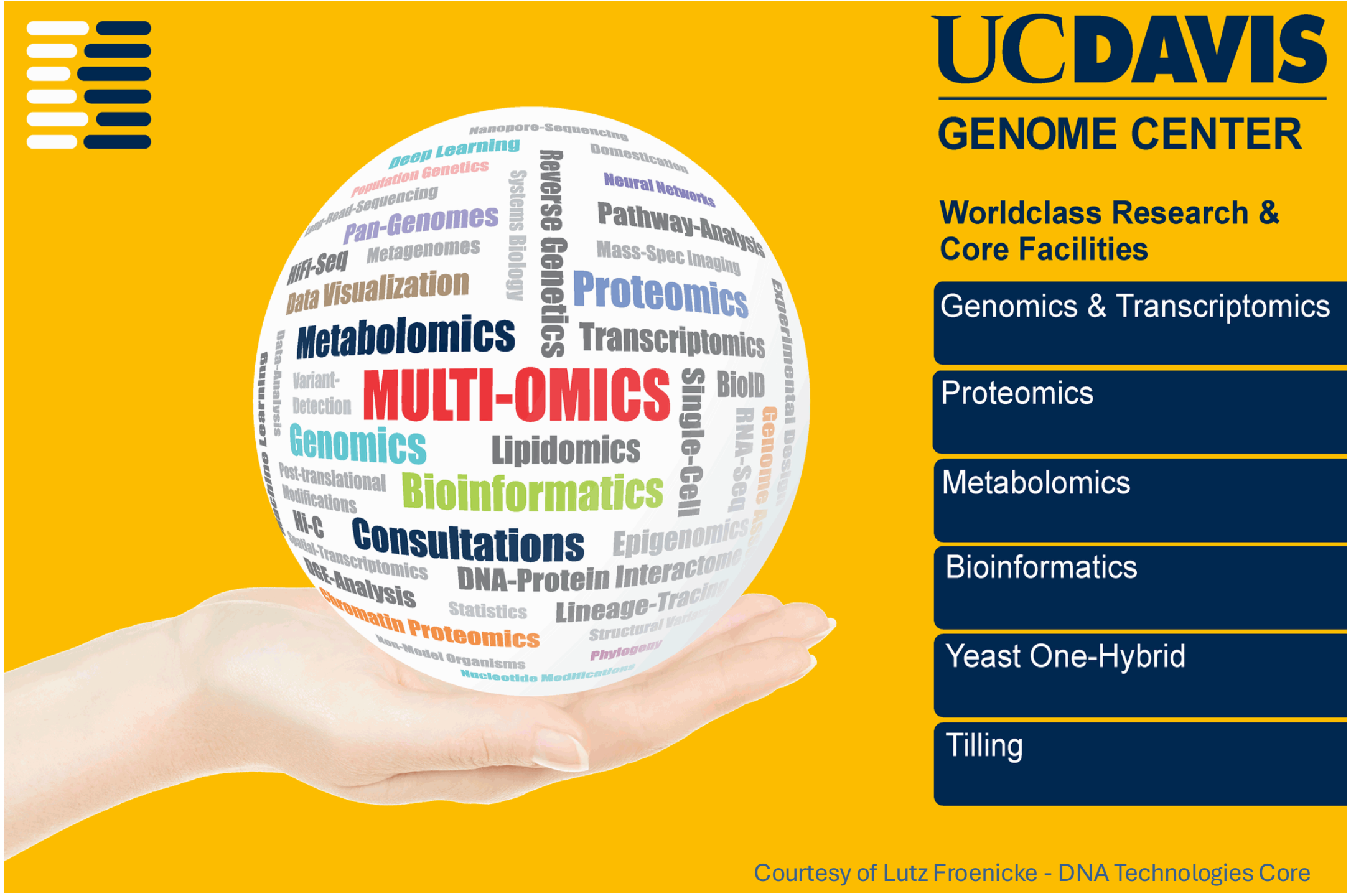
# Genome Assembly

Jie (Jessie) Li
PhD
Bioinformatics Core
Genome Center
UCDavis




UCDAVIS

GENOME CENTER

## Worldclass Research & Core Facilities

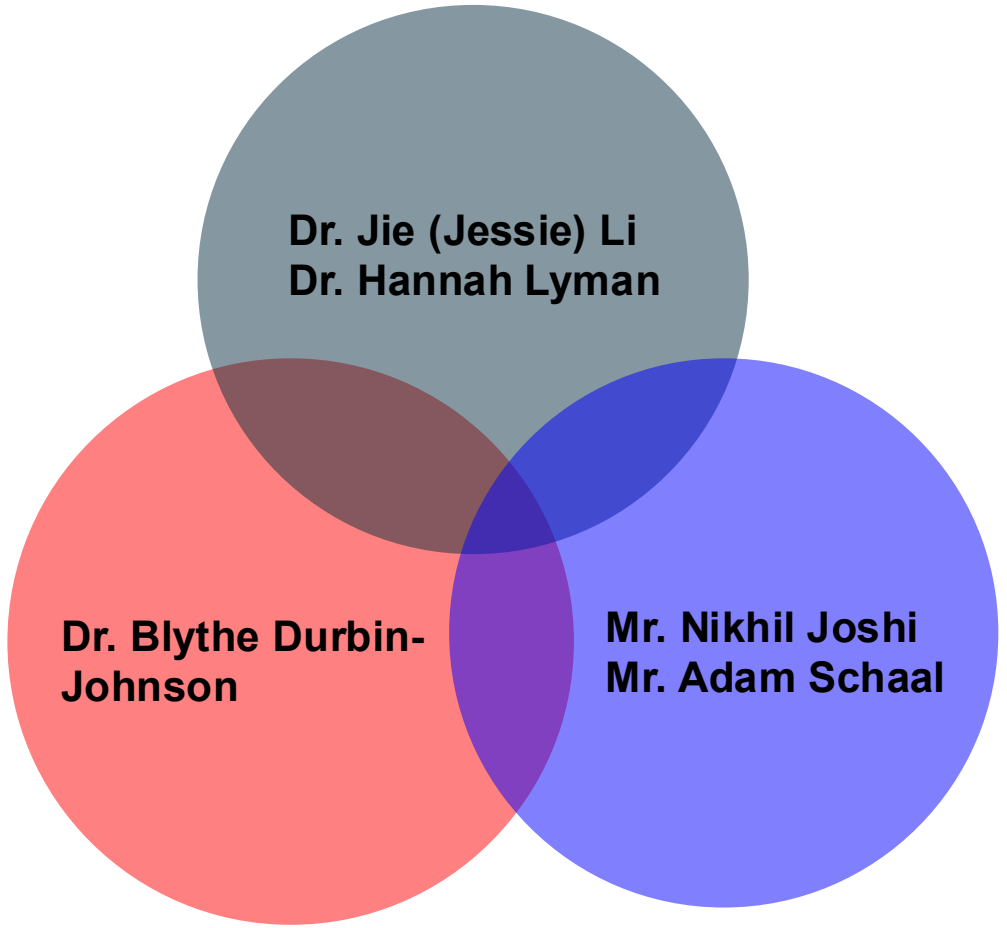
- Genomics & Transcriptomics
- Proteomics
- Metabolomics
- Bioinformatics
- Yeast One-Hybrid
- Tilling

Courtesy of Lutz Froenicke - DNA Technologies Core

# UC Davis Bioinformatics Core in the Genome Center

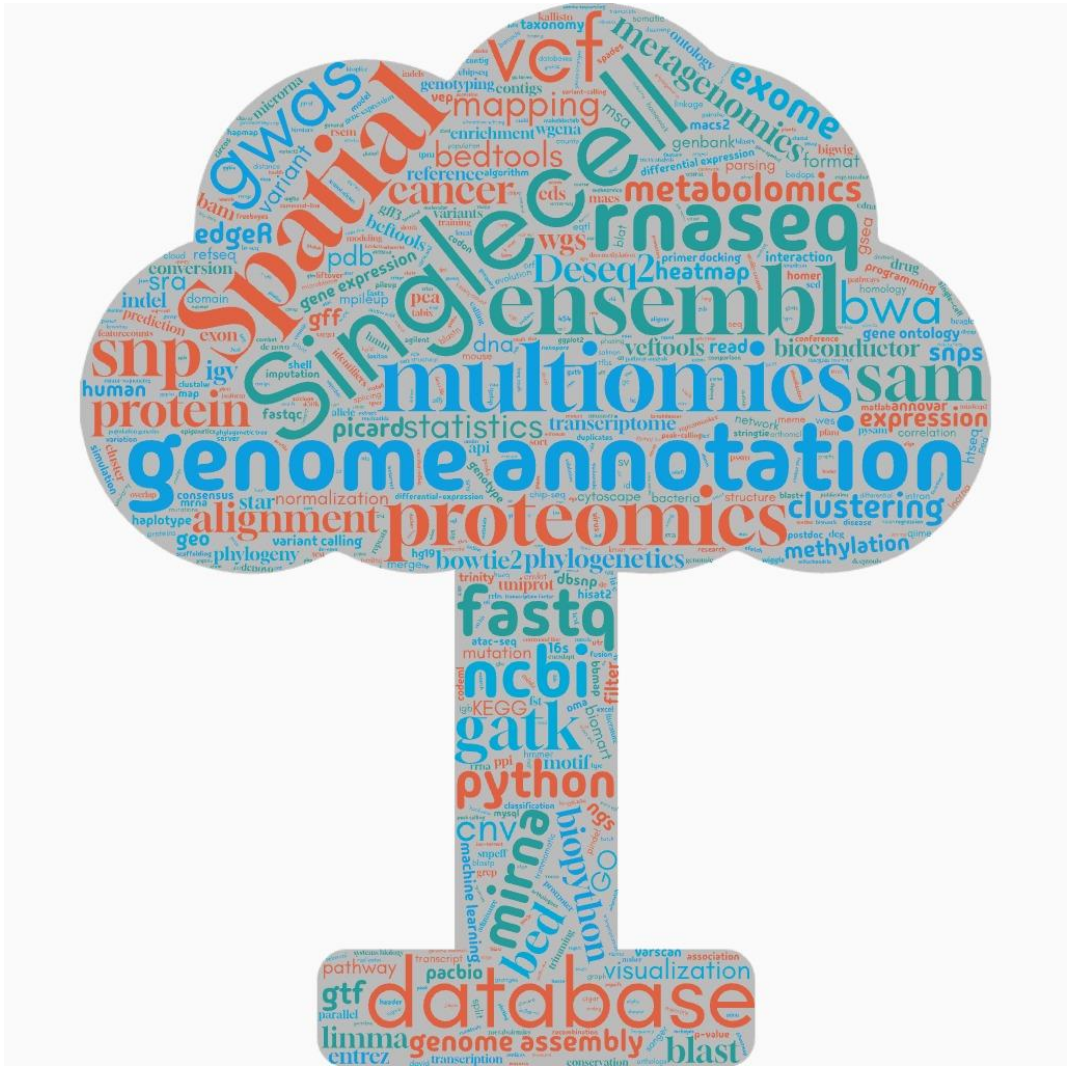
**Technical Director**
Jean Challacombe

**Scientific Director**
Ian Korf




UCDAVIS
Bioinformatics Core Facility
Genome Center

# UC Davis Bioinformatics Core in the Genome Center

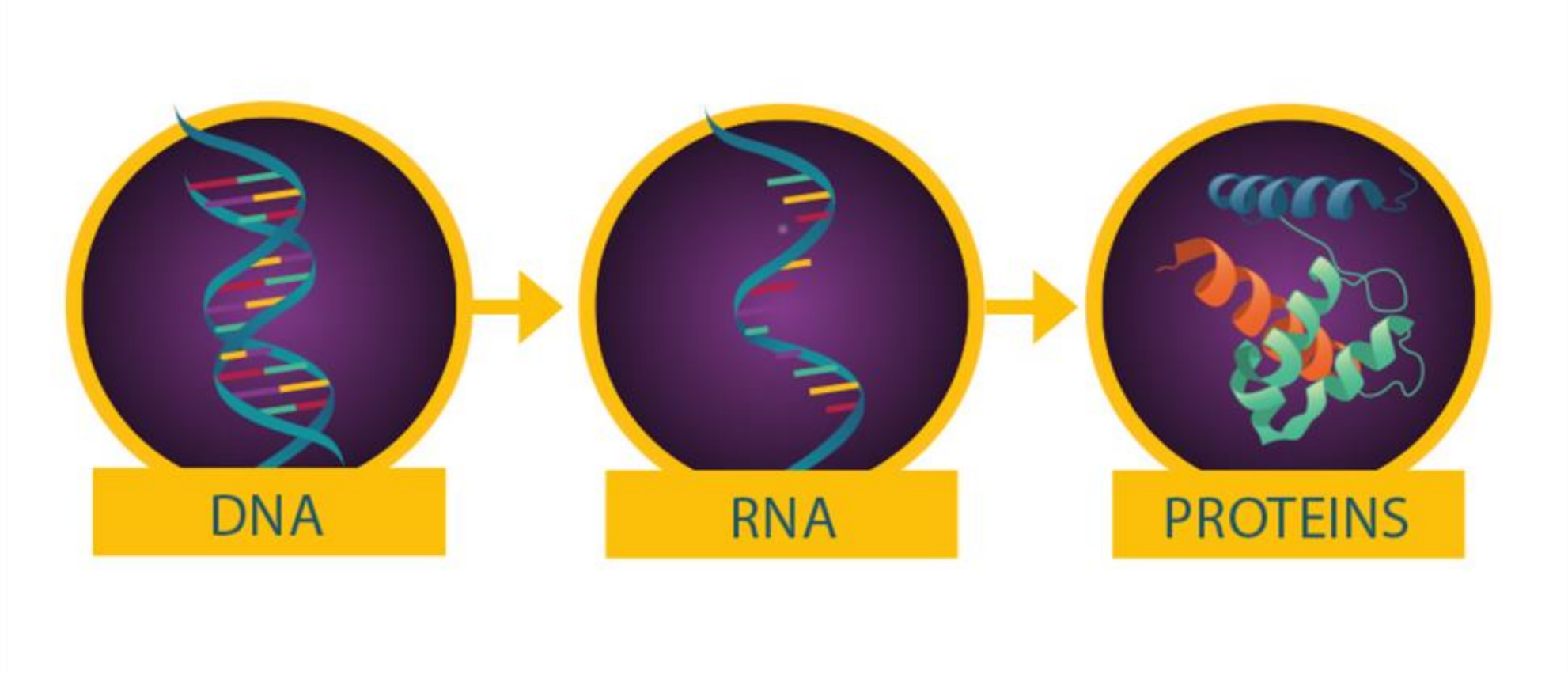



UCDAVIS
Bioinformatics Core Facility
Genome Center

# Outline

- Why assemble genomes and challenges
- Existing technologies for generating data
- Bioinformatics of genome assembly

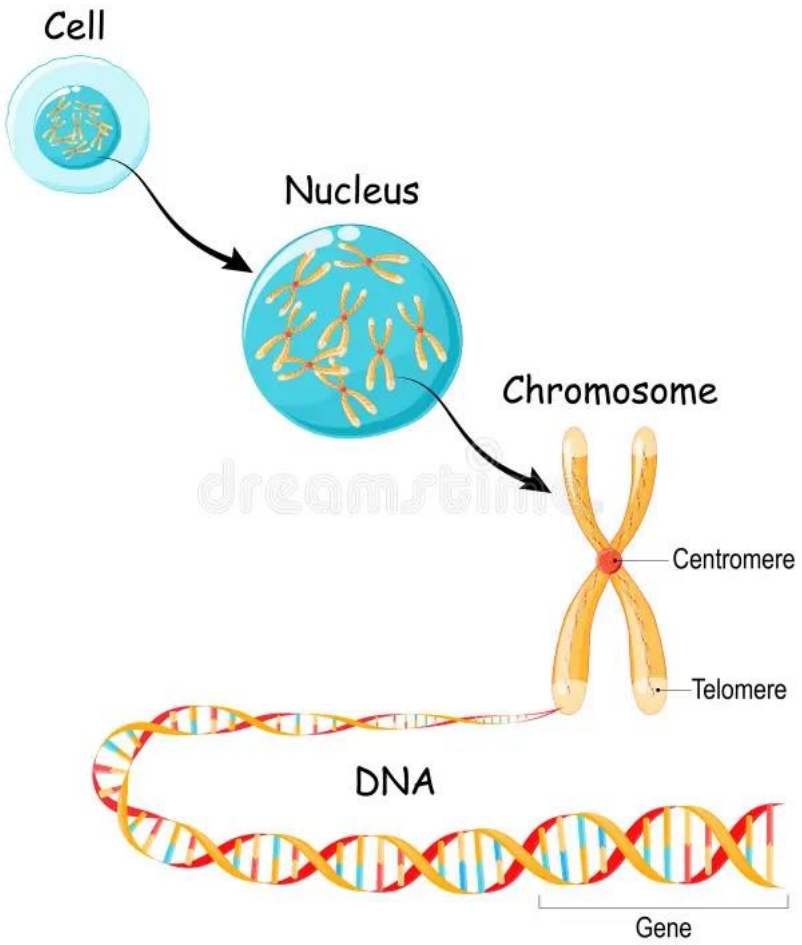
# Why genome assembly

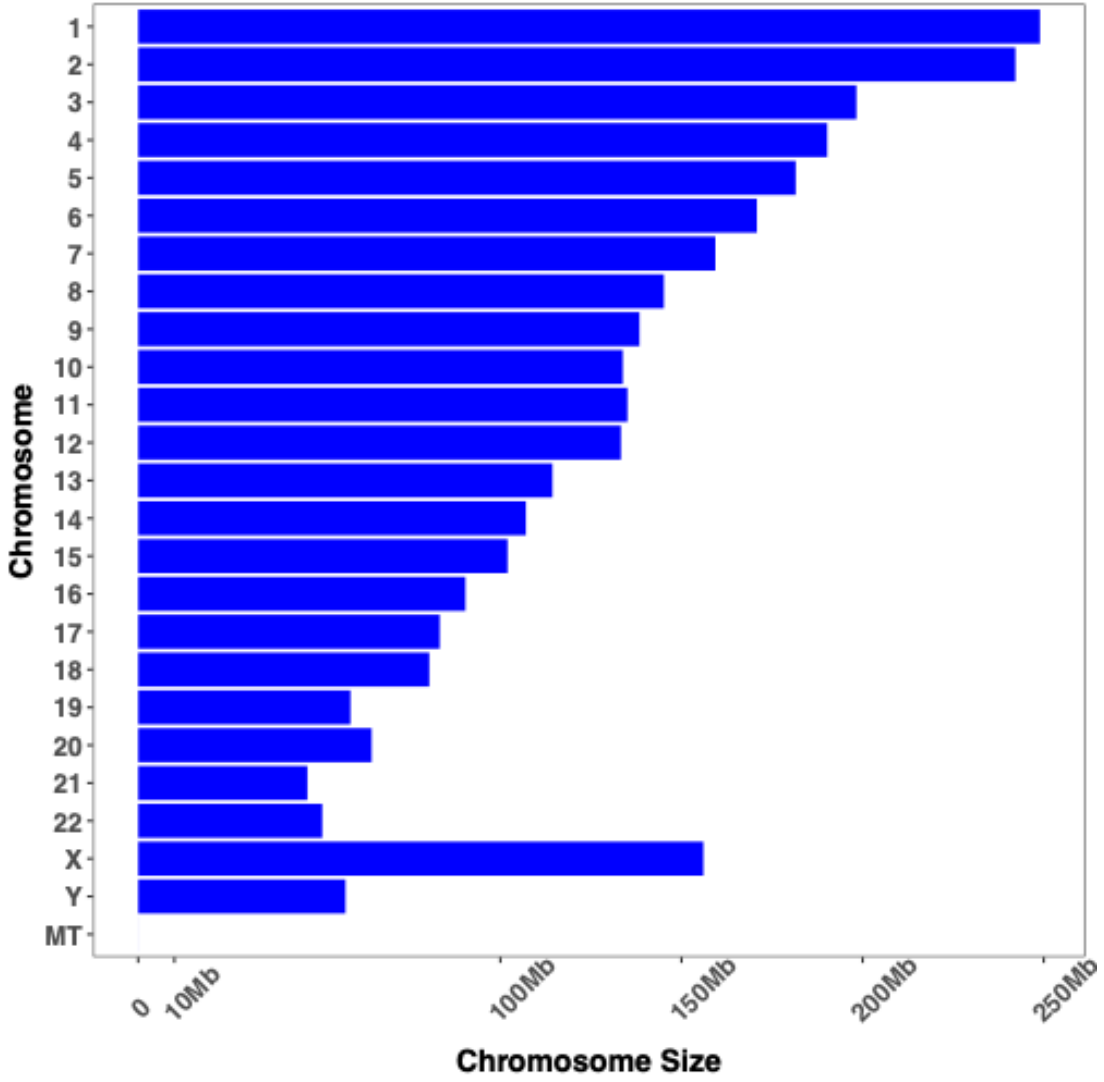


# Human reference genome

- Human Genome Project
  - 1990 – 2003
  - Cost $3Billion
- Genome Reference Consortium continued to improve the quality of the human reference. There were 79 “unresolved” gaps existing in May 2020 that accounts for ~5% of the genome.
- April 2022, Telomere-2-Telomere (T2T) consortium published a complete genome with non-Y chromosomes.
- August 2023, T2T reported the complete Y chromosome sequence

# First challenge in genome assembly

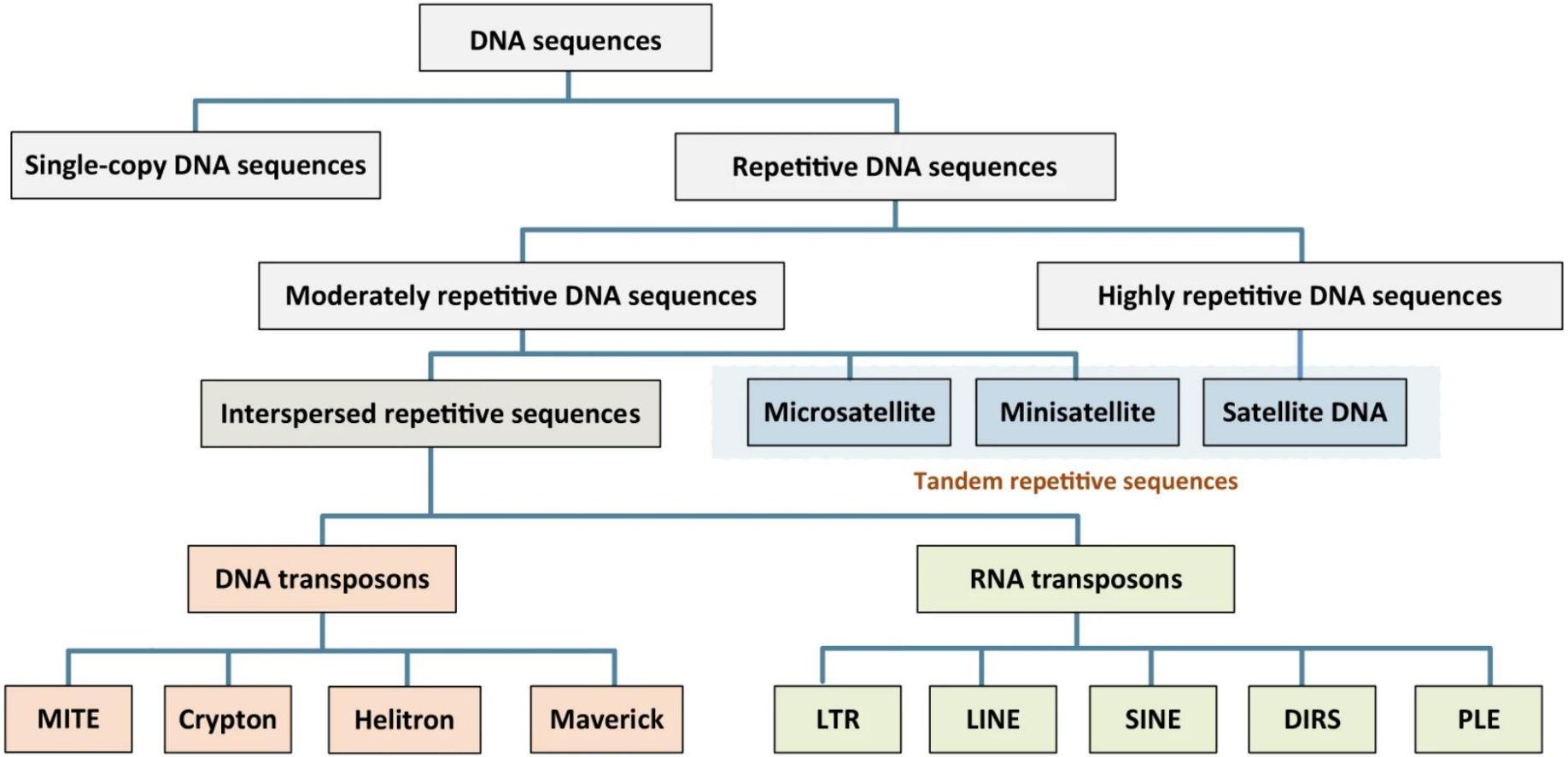


# Second challenge in genome assembly

- Repetitive sequences (repeats)
  - Patterns of nucleic acids that occur in multiple copies throughout a genome.
  - Content of repeats varies widely, ~50% in human and up to ~80% in some plant genomes.
  - Variable arrangements and sizes, with the largest repeat array in 100 Mbp.

# Second challenge in genome assembly





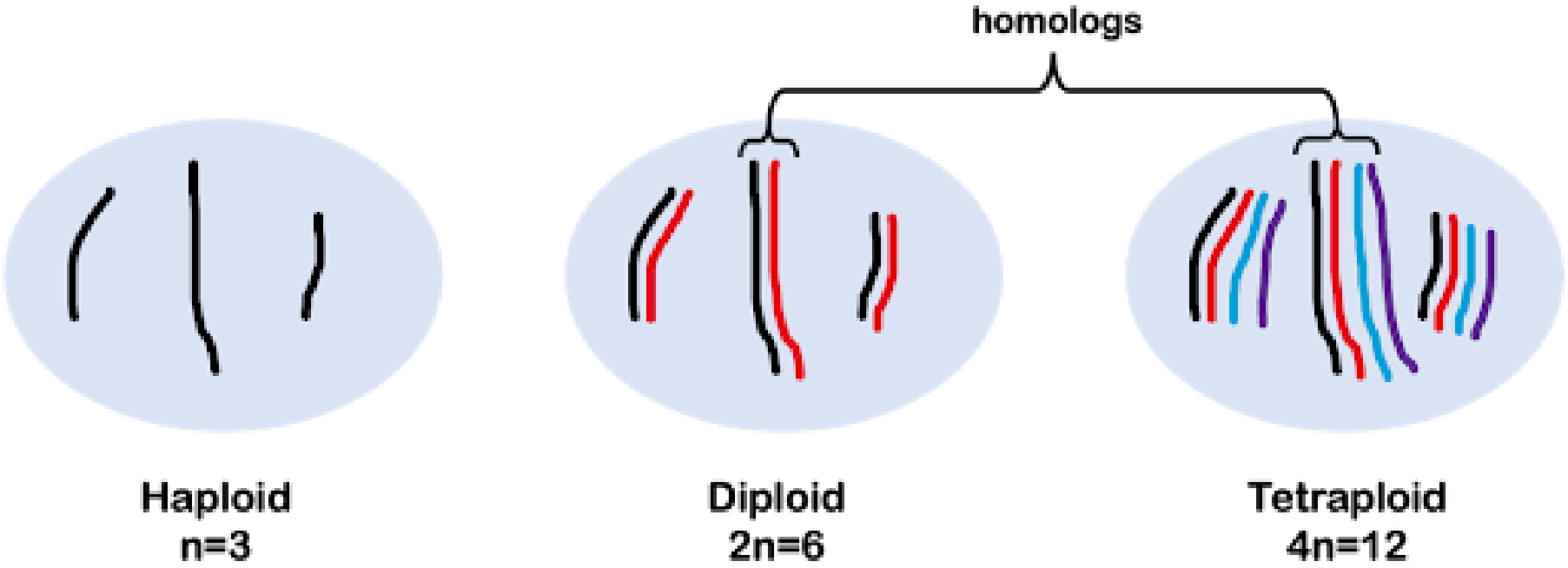

Liao, X., Zhu, W., Zhou, J. *et al*. *Commun Biol* **6**, 954 (2023). https://doi.org/10.1038/s42003-023-05322-y

# Second challenge in genome assembly

- Repeats impact essential biological processes
  - Evolution of genome sizes
  - Regulation of gene expression
  - Variation induction
  - Disease

# Third challenge in genome assembly

- Haplotype phasing






https://rotel.pressbooks.pub/genetics1/chapter/eukaryotic-genome-structure/

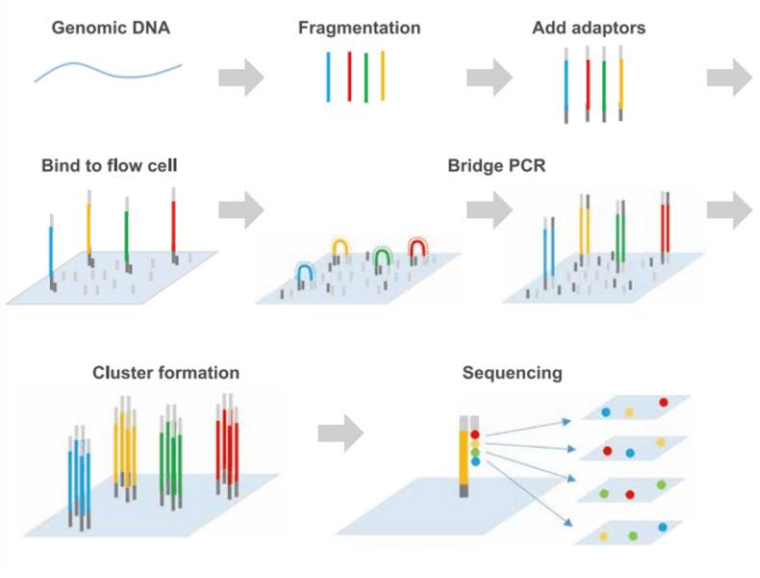
# Complementary approaches

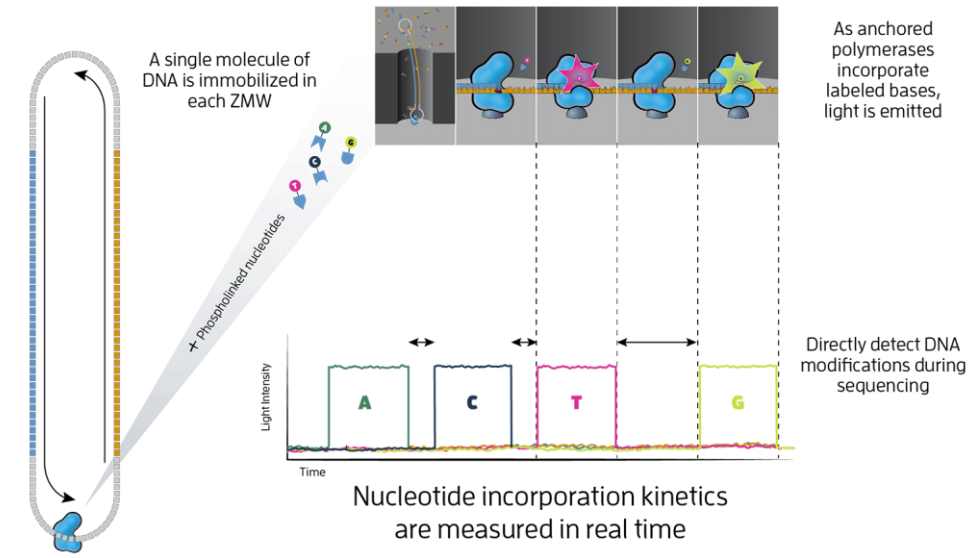
| Short-Read (Illumina, AVITI) | PacBio | PromethION Nanopore |
|---|---|---|
| Still-imaging of clusters (~1000 clonal molecules) | Movie recordings fluorescence of single molecules; HiFi: single-molecule circular consensus | Recording of electric current through a pore |
| Short reads - 2x300 bp AVITI, Miseq | HiFi reads: Up to 23 kb, N50 18 kb | Up to 100 kb, N50 25 kb<br>Ultra-long protocol up to 1Mb |
| Repeats are mostly not analyzable | spans retro elements | spans retro elements |
| High output - up to 1 Tb per lane | up to 130 Gb HiFi data per SMRT-cell | Up to 100 Gb per flowcell<br>Up to 80 million reads “cDNA-seq” |
| High accuracy ( < 0.5 %) | Raw data error rate 15 %<br>HiFi CCS data < 0.1% | Raw data error rate 2-10 % |
| Considerable base composition bias | No base composition bias | Some systematic errors |
| Very affordable | Costs 4 x higher | Costs 4 x higher |
| *De novo* assemblies result in thousands of scaffolds | “Near perfect” genome assemblies; lowest error rate | “Near perfect” genome assemblies;<br>highest contiguity |



Courtesy of Lutz Froenicke - DNA Technologies Core

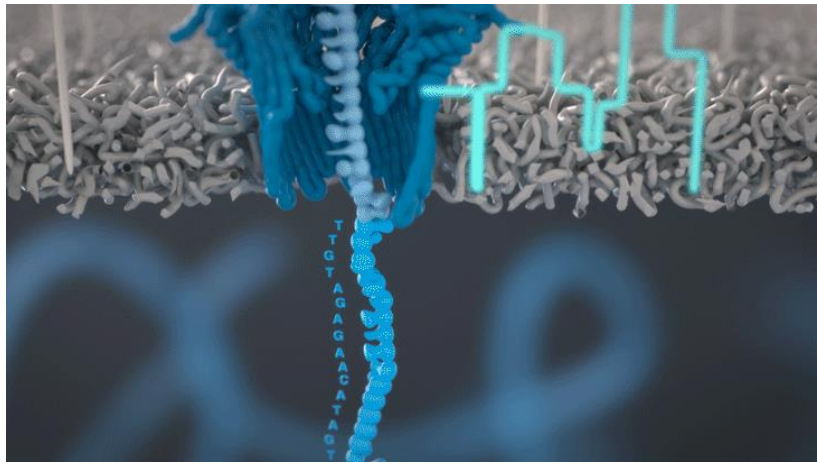
# Complementary approaches




As anchored polymerases incorporate labeled bases, light is emitted

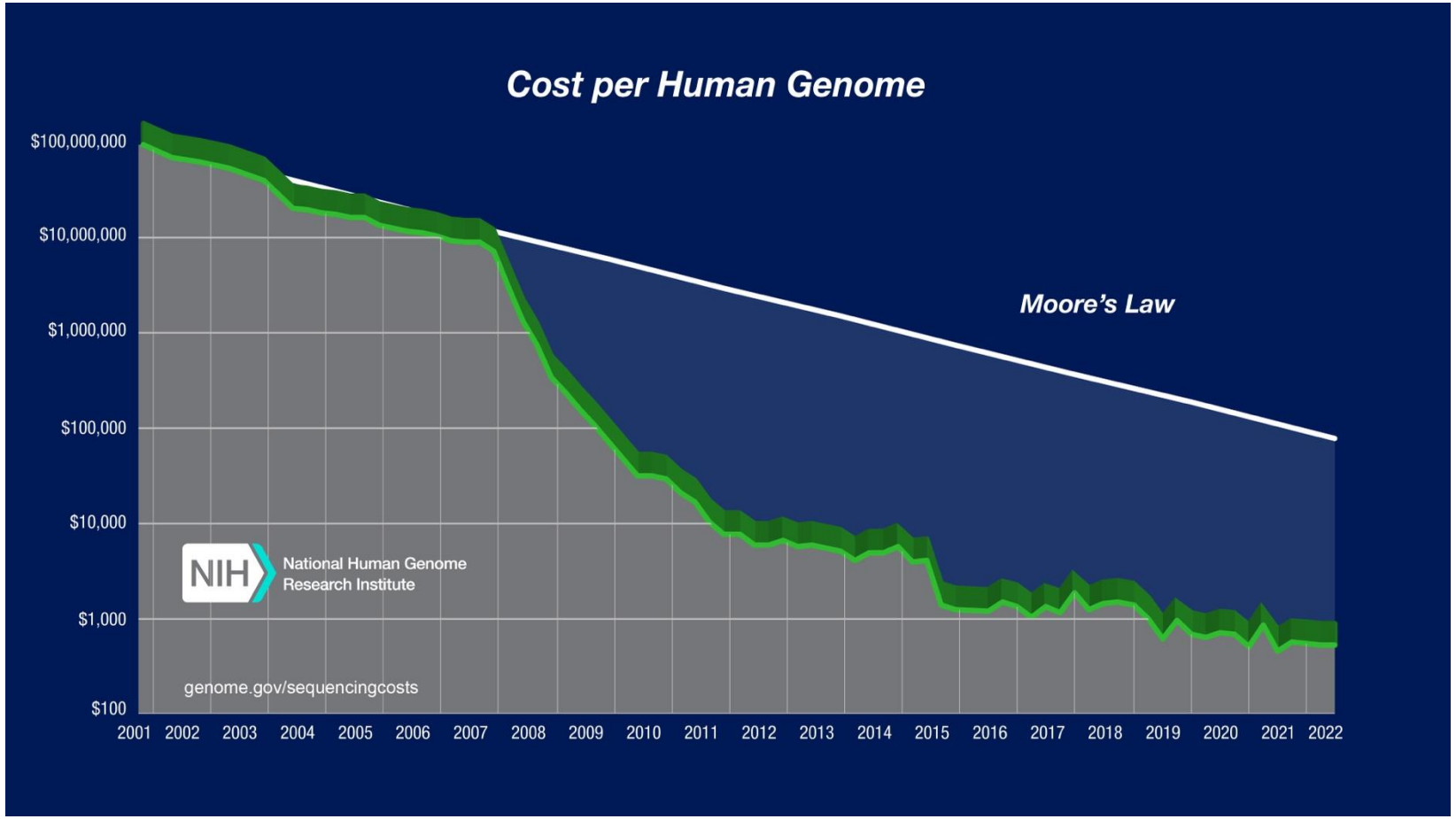
Directly detect DNA modifications during sequencing



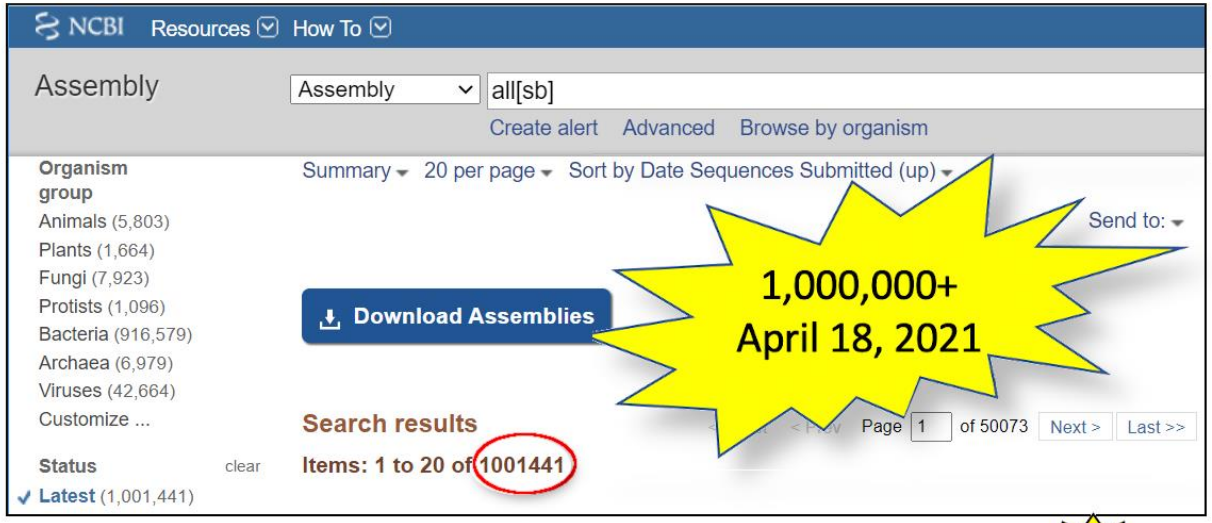
# Stages involved in assembling a genome

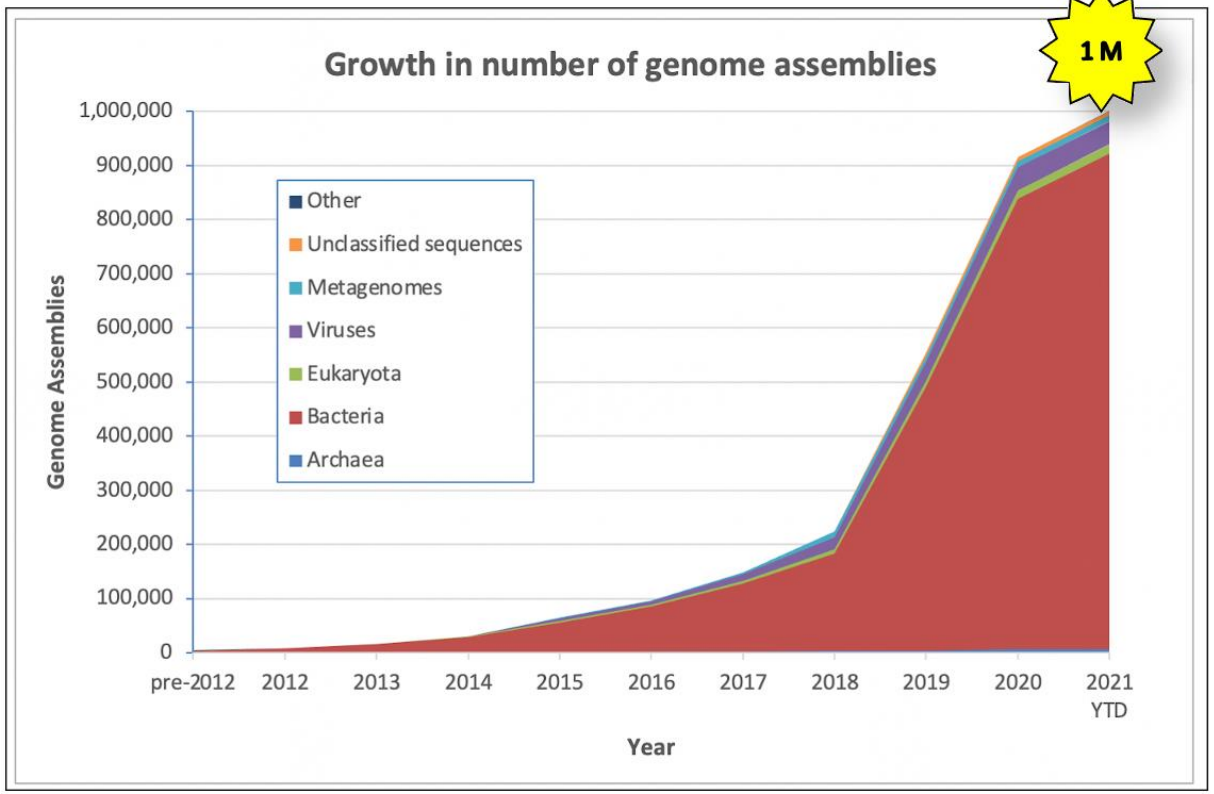
- High molecular weight DNA extract
  - Some tissues are easier to work with than others
  - Some species are easier to work with than others
- Sequencing library prep
- Sequencing
  - PacBio HiFi + ultra-long ONT/HiC
  - ONT
  - Illumina (very small genomes)
- Bioinformatically assemble the genome

# Sequencing cost keeps decreasing

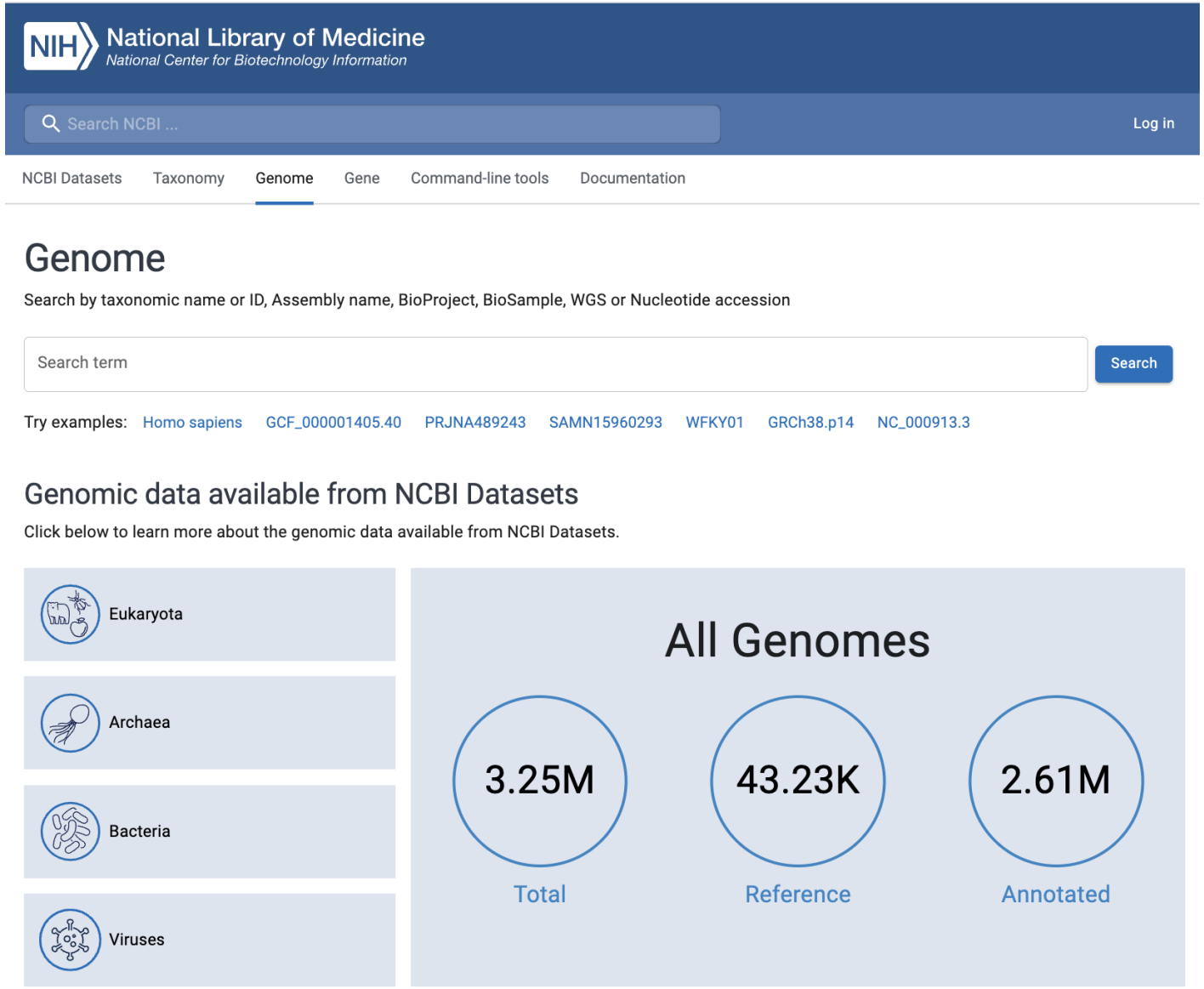


# Number of genomes assembled increasing

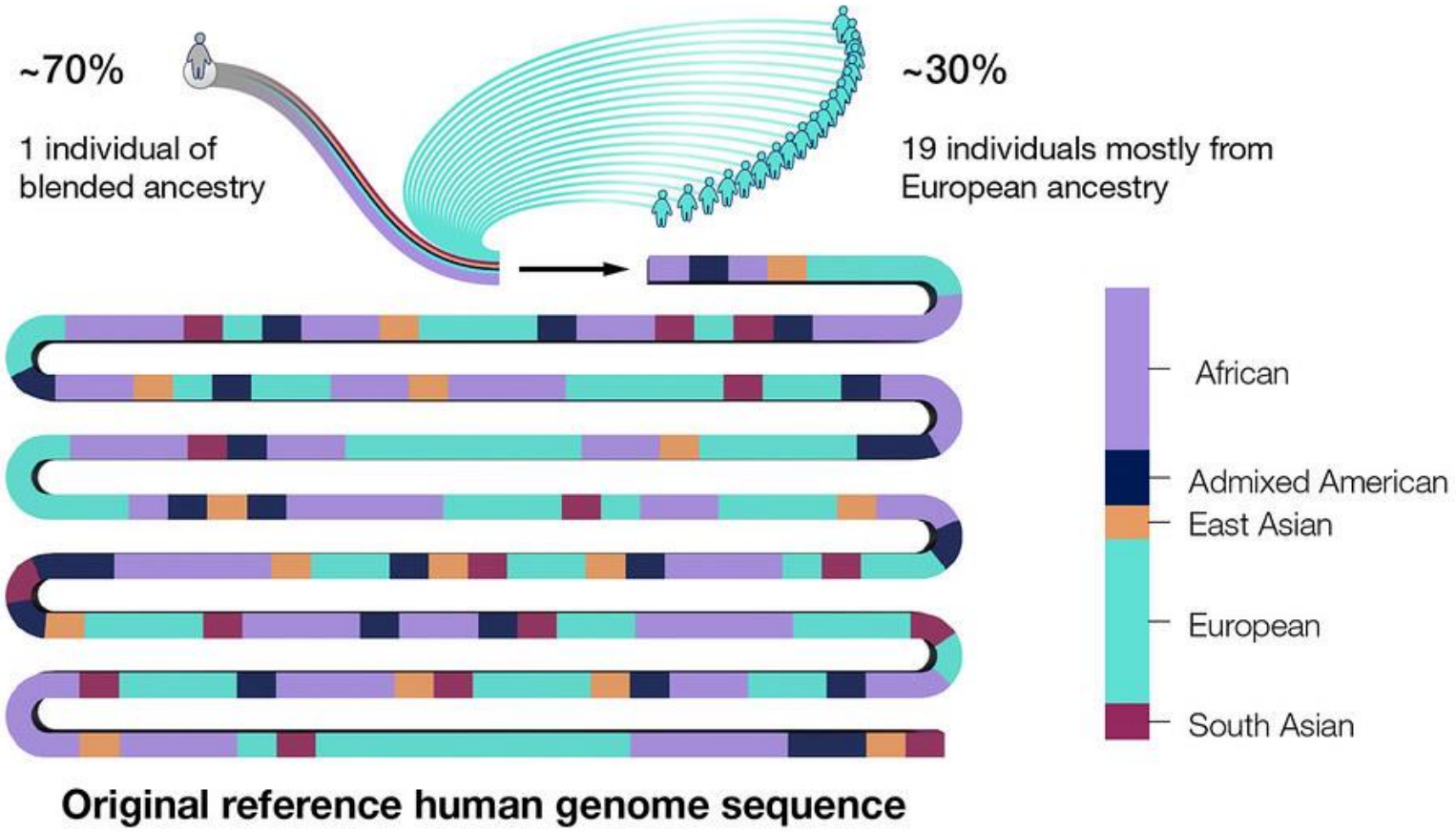


# Number of genomes assembled increasing



# Human reference genome

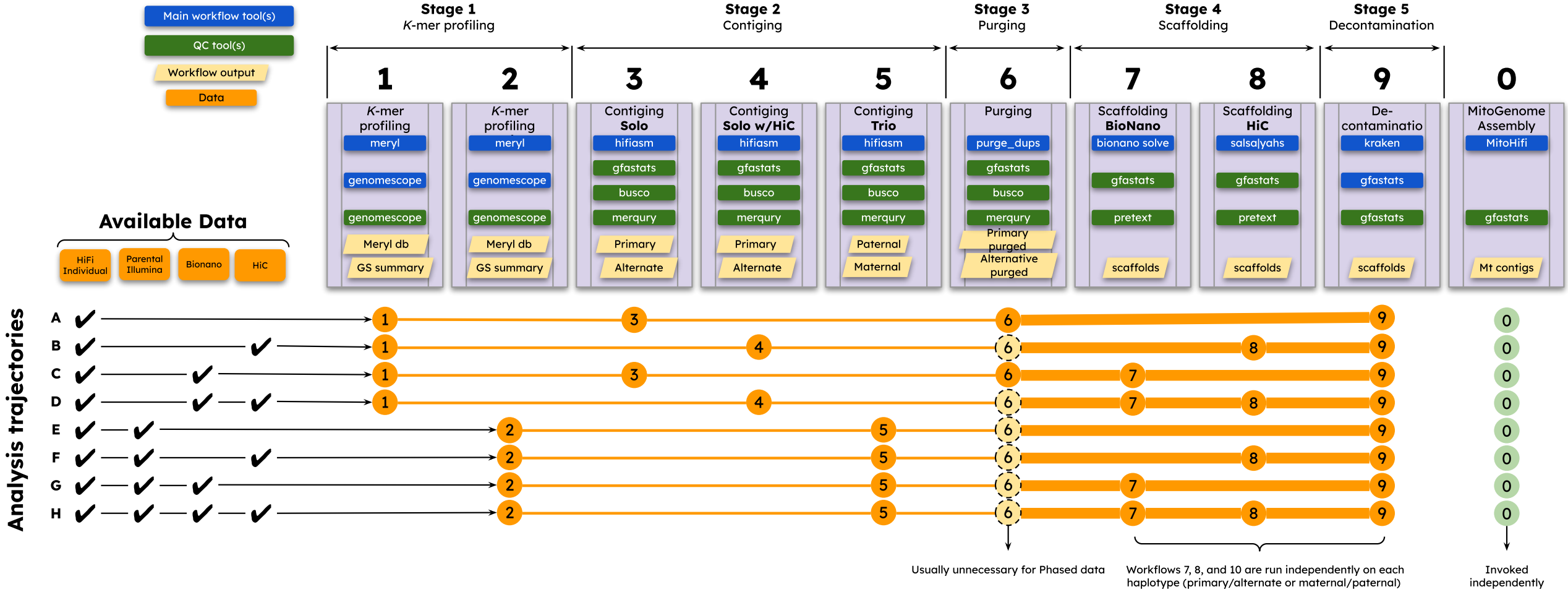


# Genome assembly efforts

- Earth BioGenome Project (EBP)
  - An initiative aiming to sequence and annotate all 1.5-1.8 Million eukaryotic species
  - A global network includes over 61 affiliated projects
  - Vertebrate Genomes Project (VGP)
- Telomere-to-Telomere (T2T) consortium
- Genome Reference Consortium (GRC)

# Generate haplotype phased telomere-to-telomere high quality genomes.

# VGP assembly pipeline




UCDAVIS
Bioinformatics Core Facility
Genome Center

https://training.galaxyproject.org/training-material/topics/assembly/tutorials/vgp_genome_assembly/tutorial.html

# VGP assembly pipeline

- Stage 1: Data QC (Kmer profiling)
- Stage 2: Assemble reads to contigs
- Stage 3: Purge duplicates
- Stage 4: Scaffold
- Stage 5: Decontamination

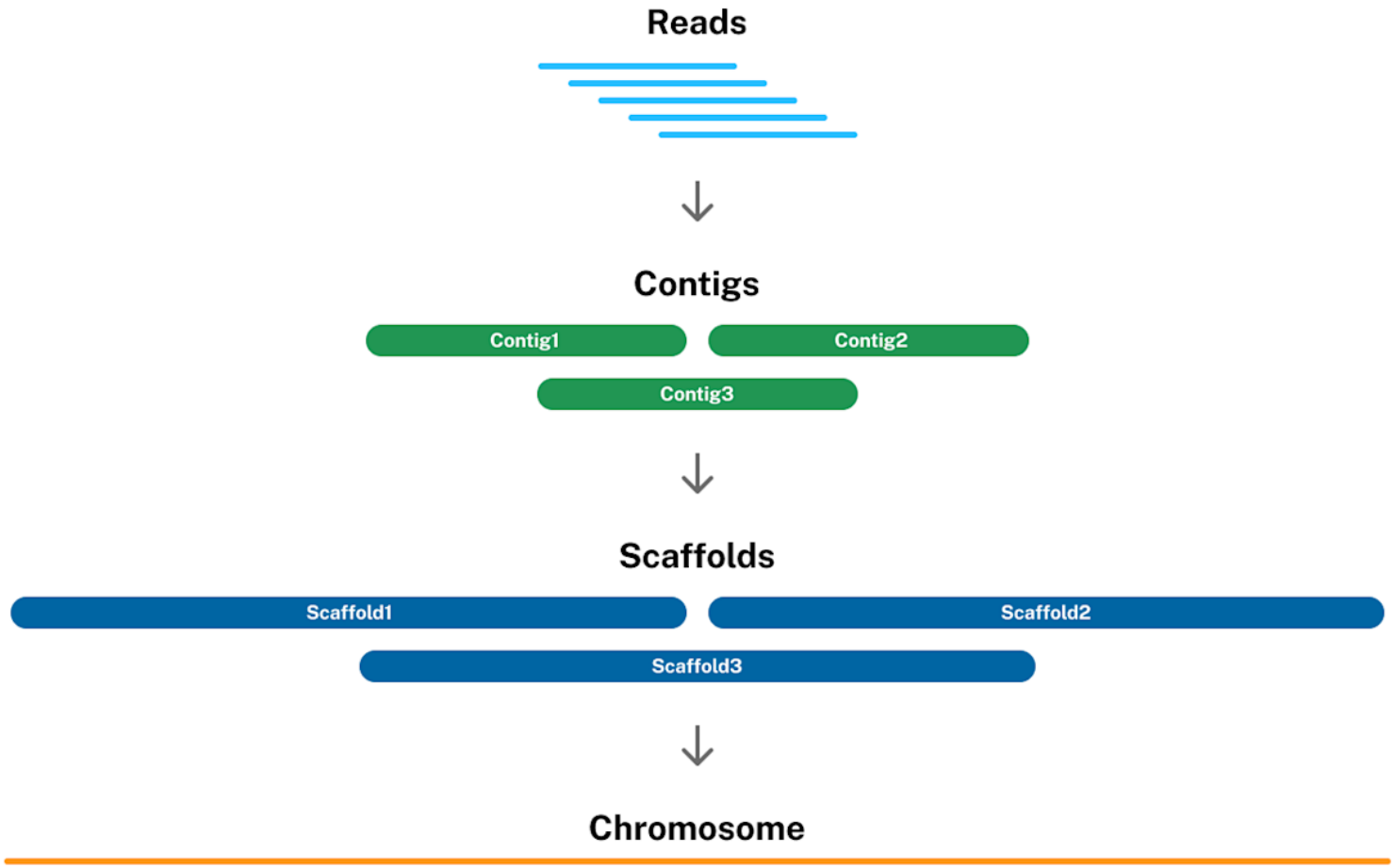
# Current approaches

- PacBio HiFi + HiC/Bionano
- PacBio HiFi + UL ONT
- ONT + UL ONT
- Trio data facilitate haplotype phasing

# VGP established assembly pipeline

- Stage 1: Data QC (Kmer profiling)
- Stage 2: Assemble reads to contigs
- Stage 3: Purge duplicates
- Stage 4: Scaffold
- Stage 5: Decontamination

# Premise of genome assembly






https://www.geneious.com/guides/assembling-dna-sequences

# Premise of genome assembly

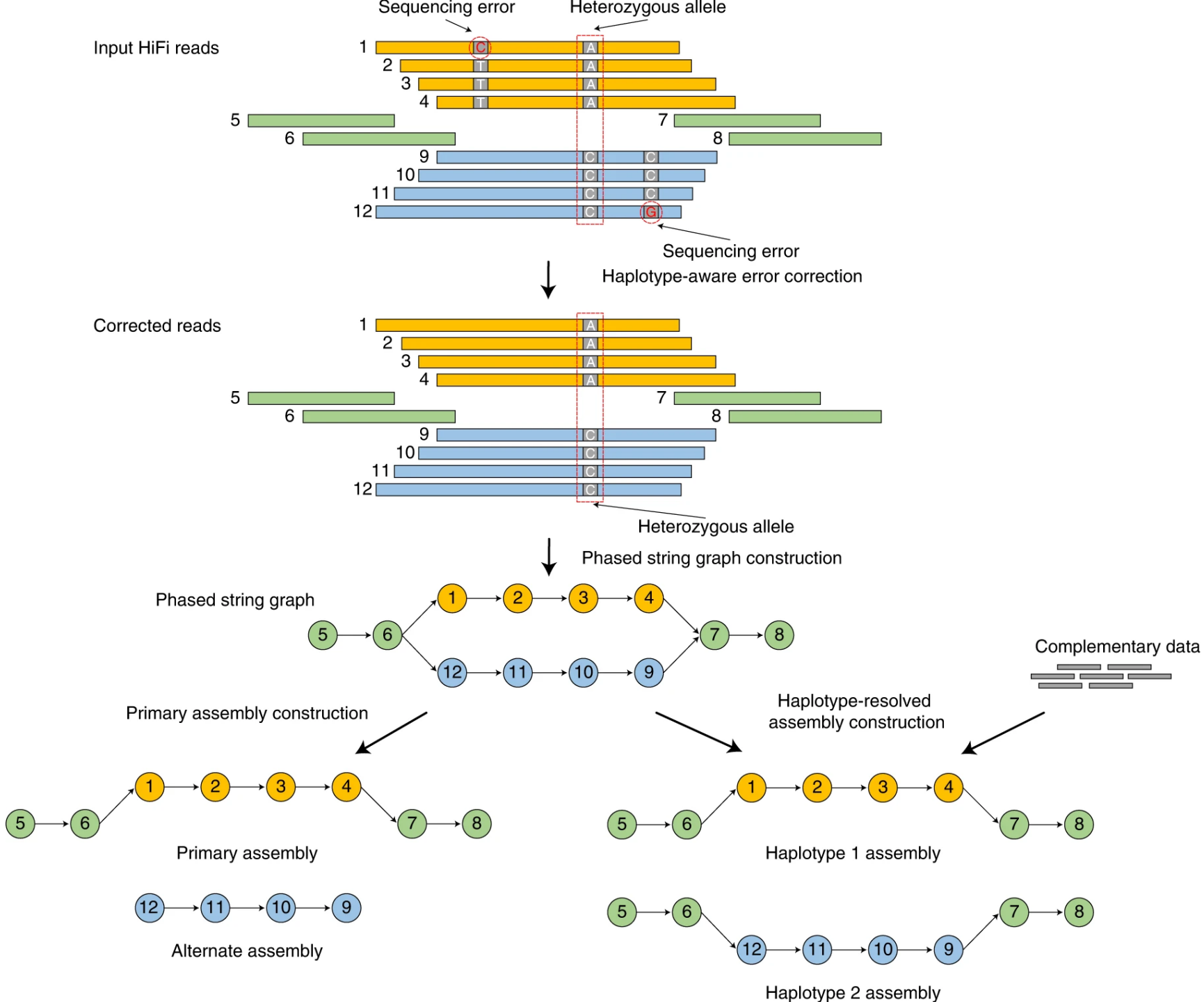
*GRAPH*
*THEORY*



https://www.geneious.com/guides/assembling-dna-sequences

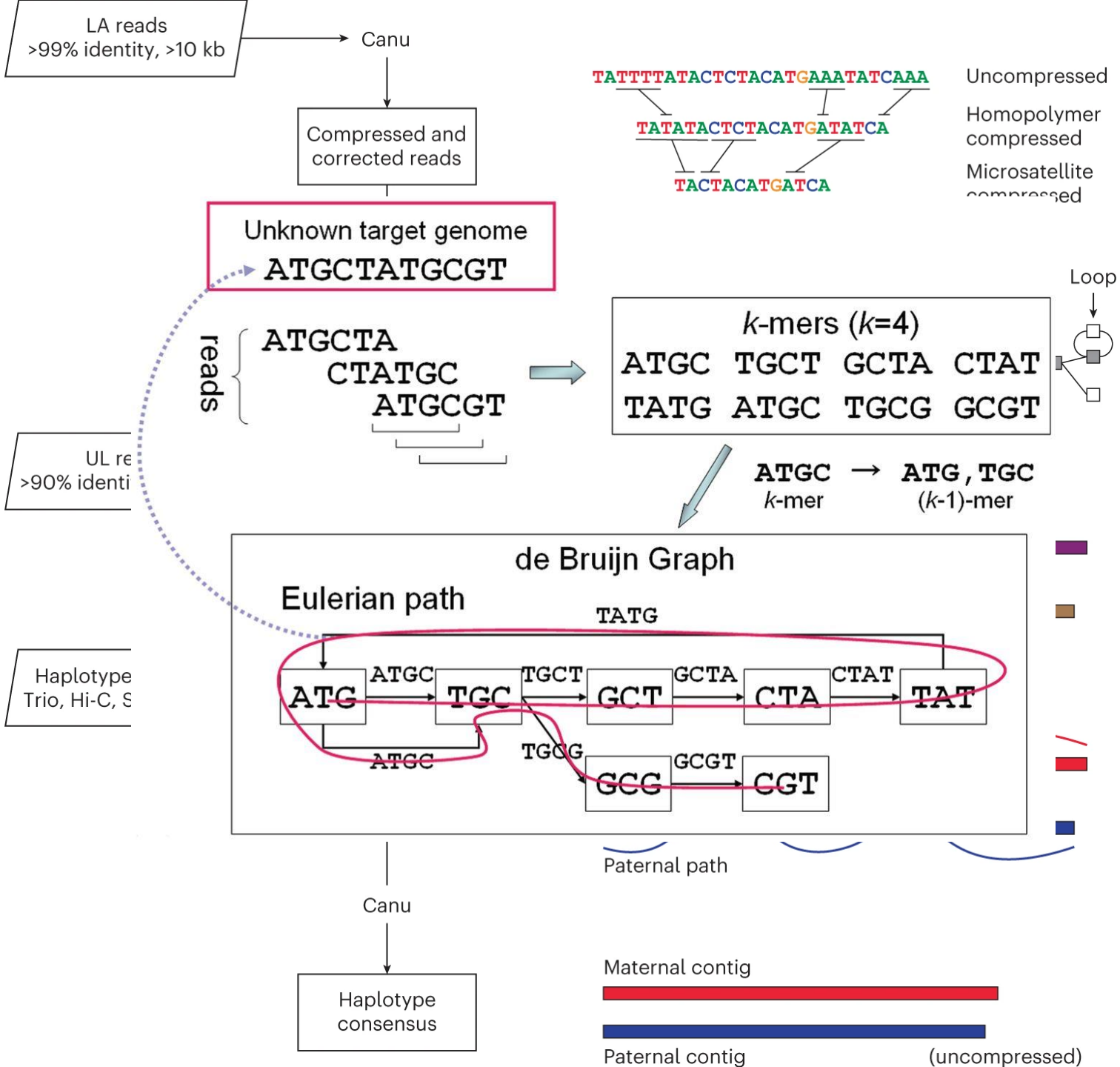
# HiFiasm

Cheng et. al, *Nat Methods*, 2021



# Verkko

Rautiainen, M. *et al. Nat Biotechnol* 2023




UCDAVIS
Bioinformatics Core Facility
Genome Center

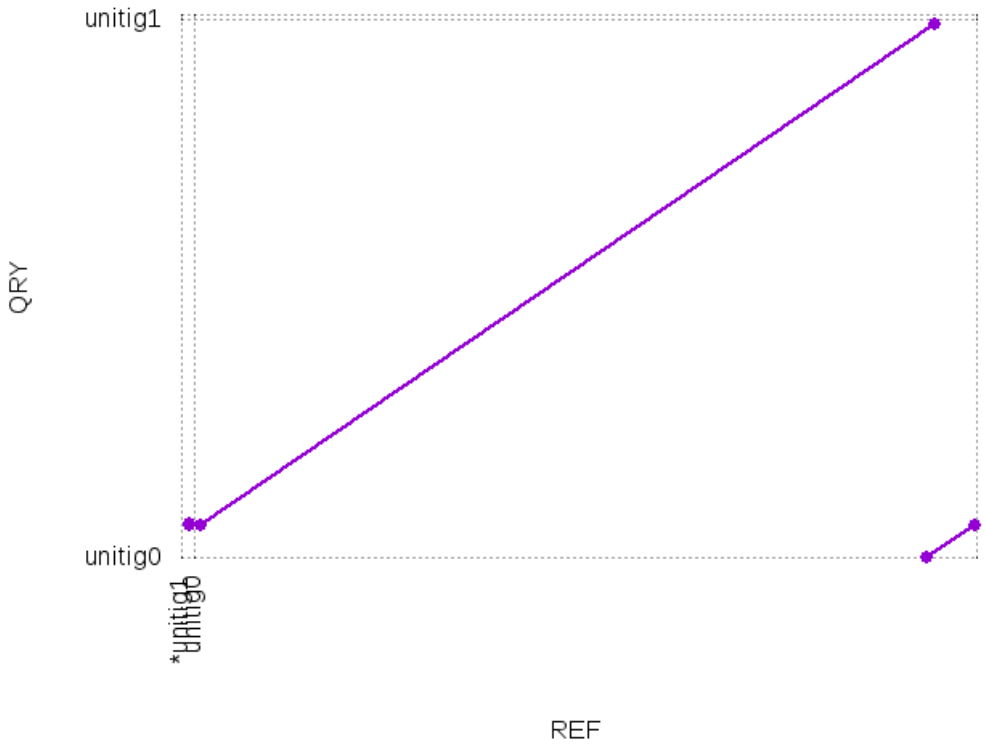
# Quality assessment

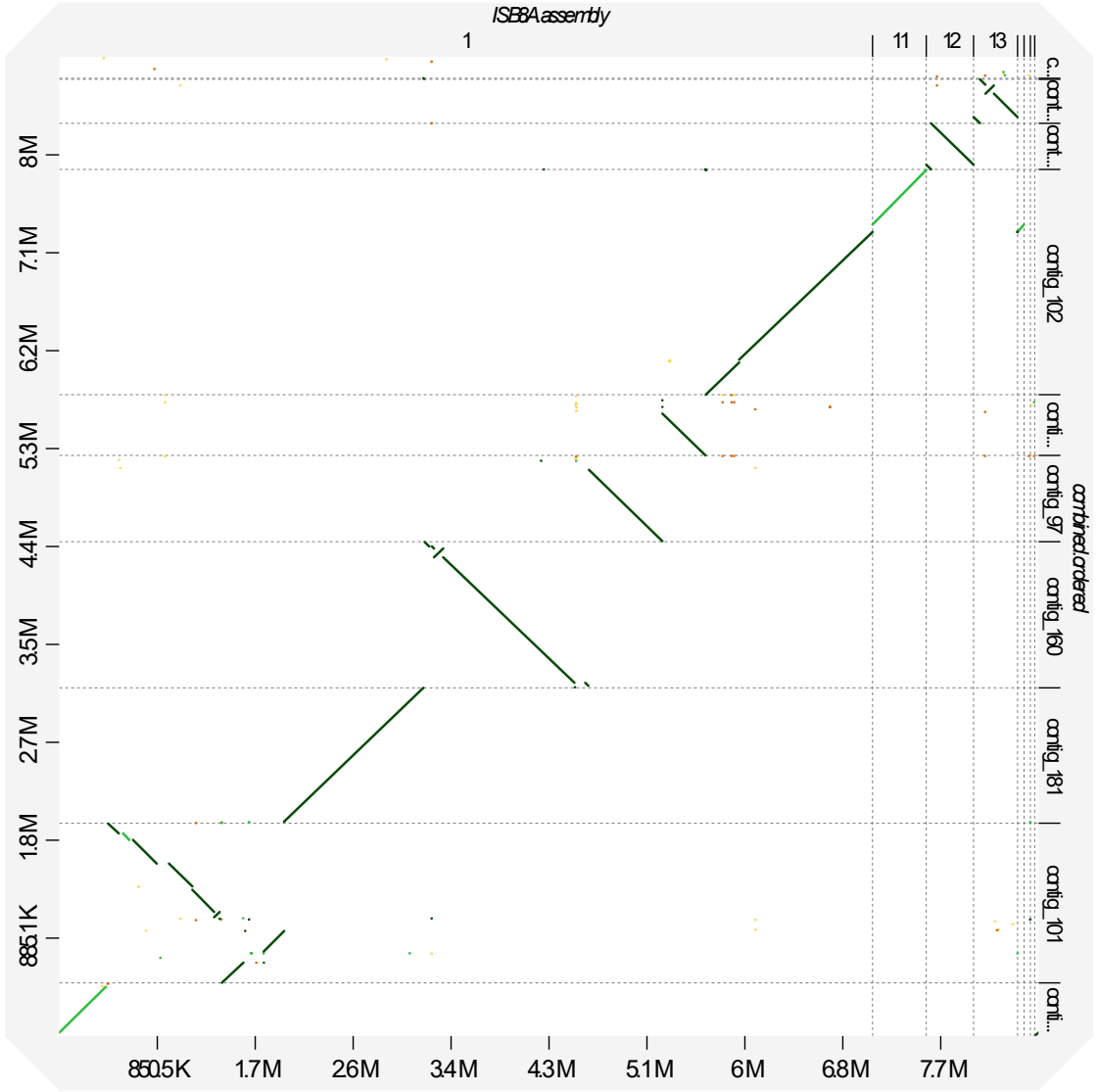
- Contiguity
  - N50
  - Gaps
- Completeness
  - BUSCO
  - Repeat structures
- Contamination

# Genome assembly

- Generate appropriate data
  - PacBio HiFi/ONT
  - HiC/ONT ultra long
  - Trio data facilitate better phasing
- Assemble
  - HiFiasm
  - Verkko
- Quality assessment

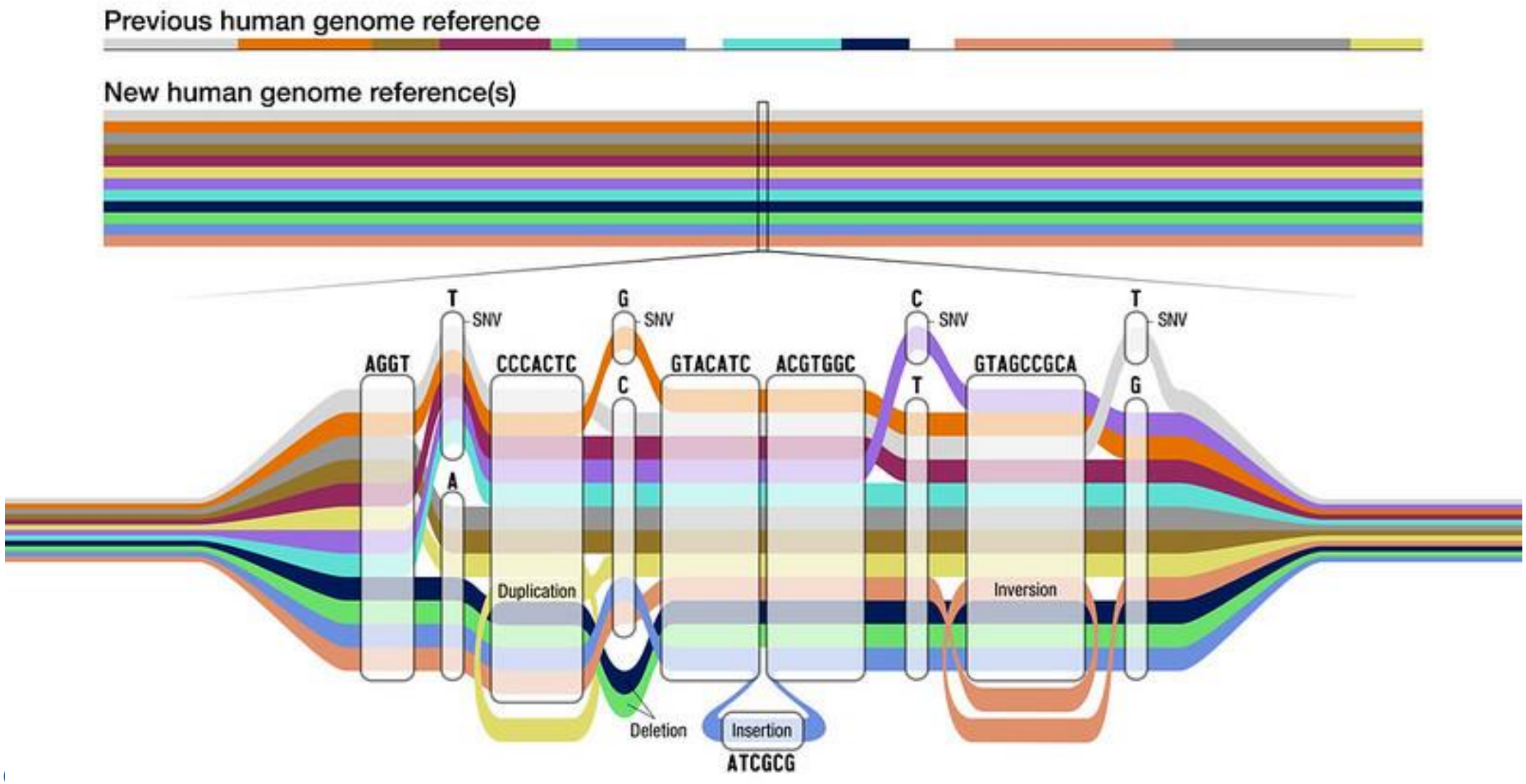
# Compare genomes




UCDAVIS
Bioinformatics Core Facility
Genome Center

# Pangenome era




Darryl Leja, NHGRI, https://www.nih.gov/news-events/news-releases/scientists-release-new-human-pangenome-reference

# Thank you!