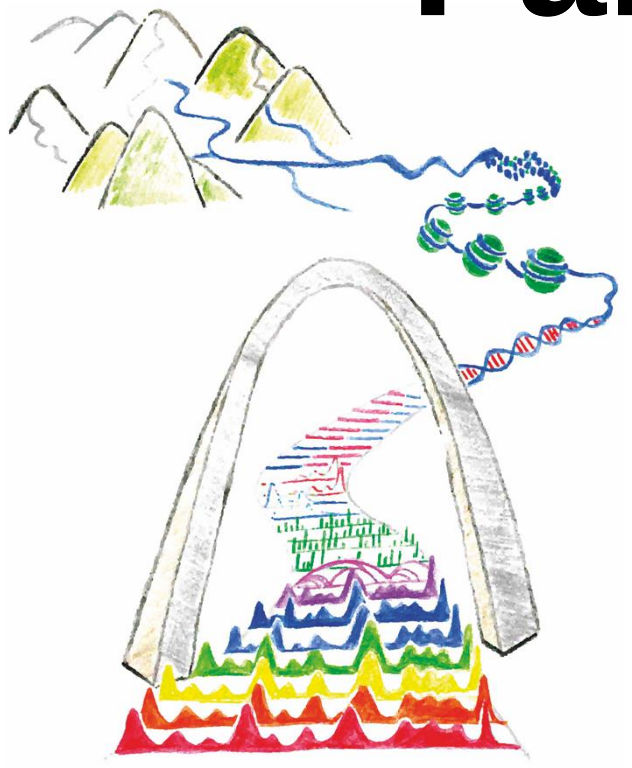
Genome and Pangenome Assembly

Juan F. Macias-Velasco, PhD

Inaugural Dean's Fellow
Department of Genetics
CIG@MGI
Washington University in St. Louis

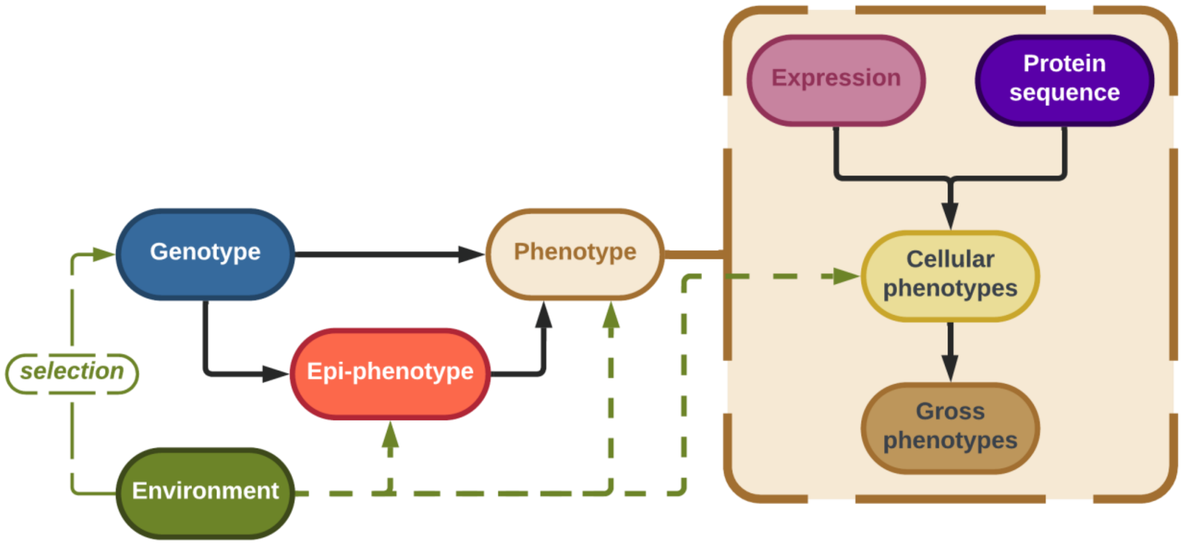


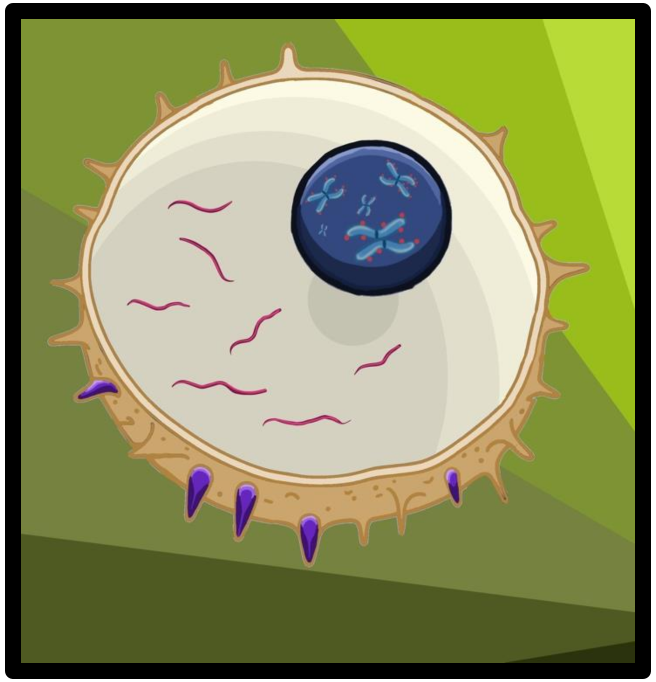
Chad Tomlinson

Eddie Belter

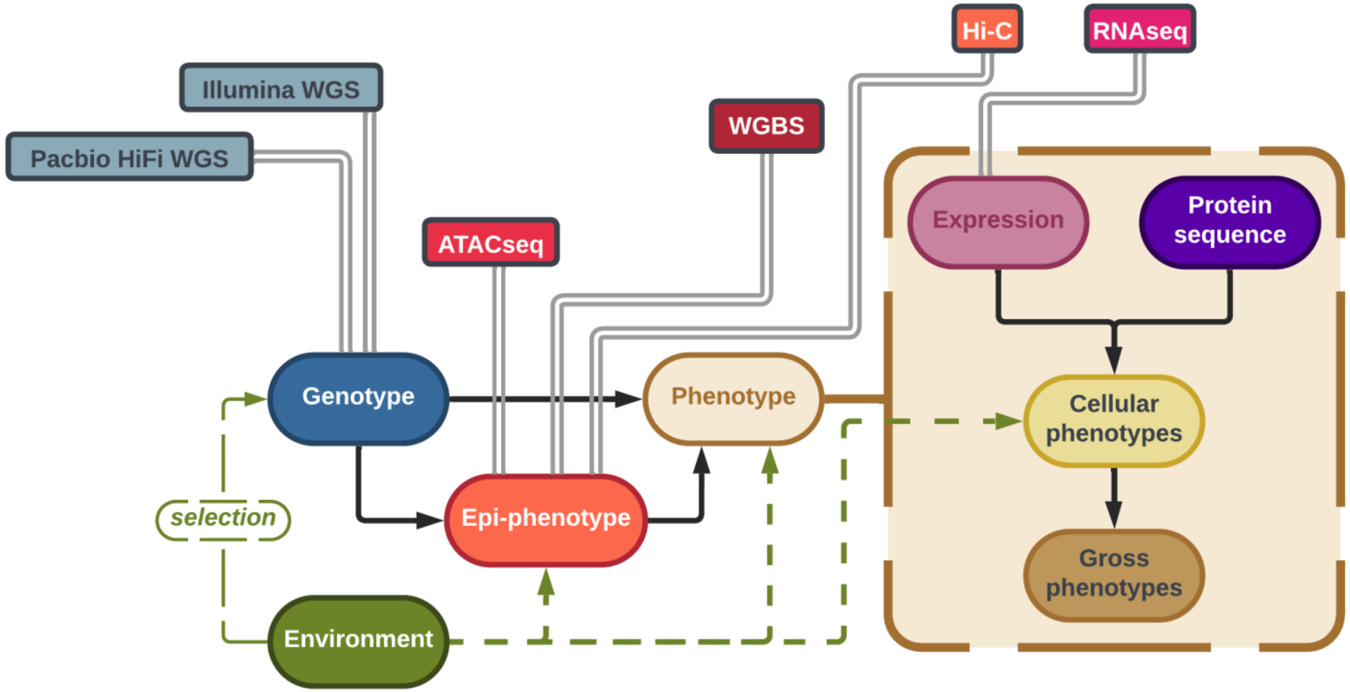
John Garza

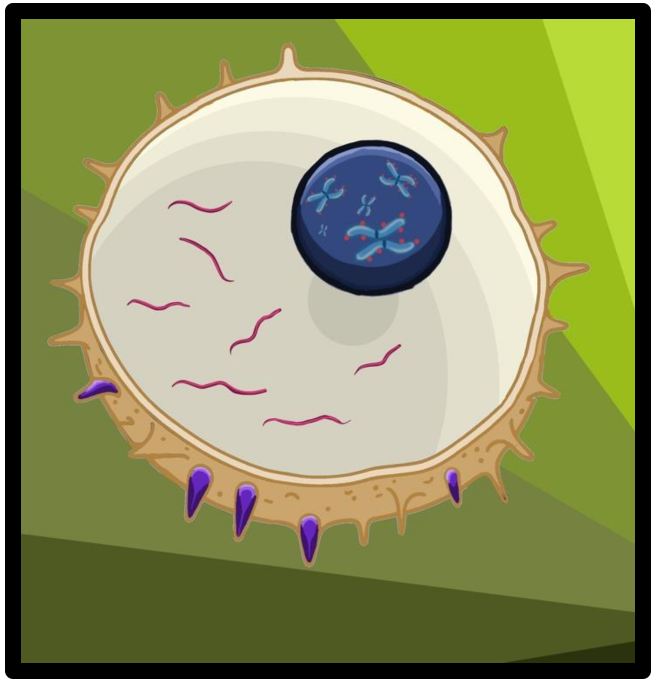
Understanding Biological systems



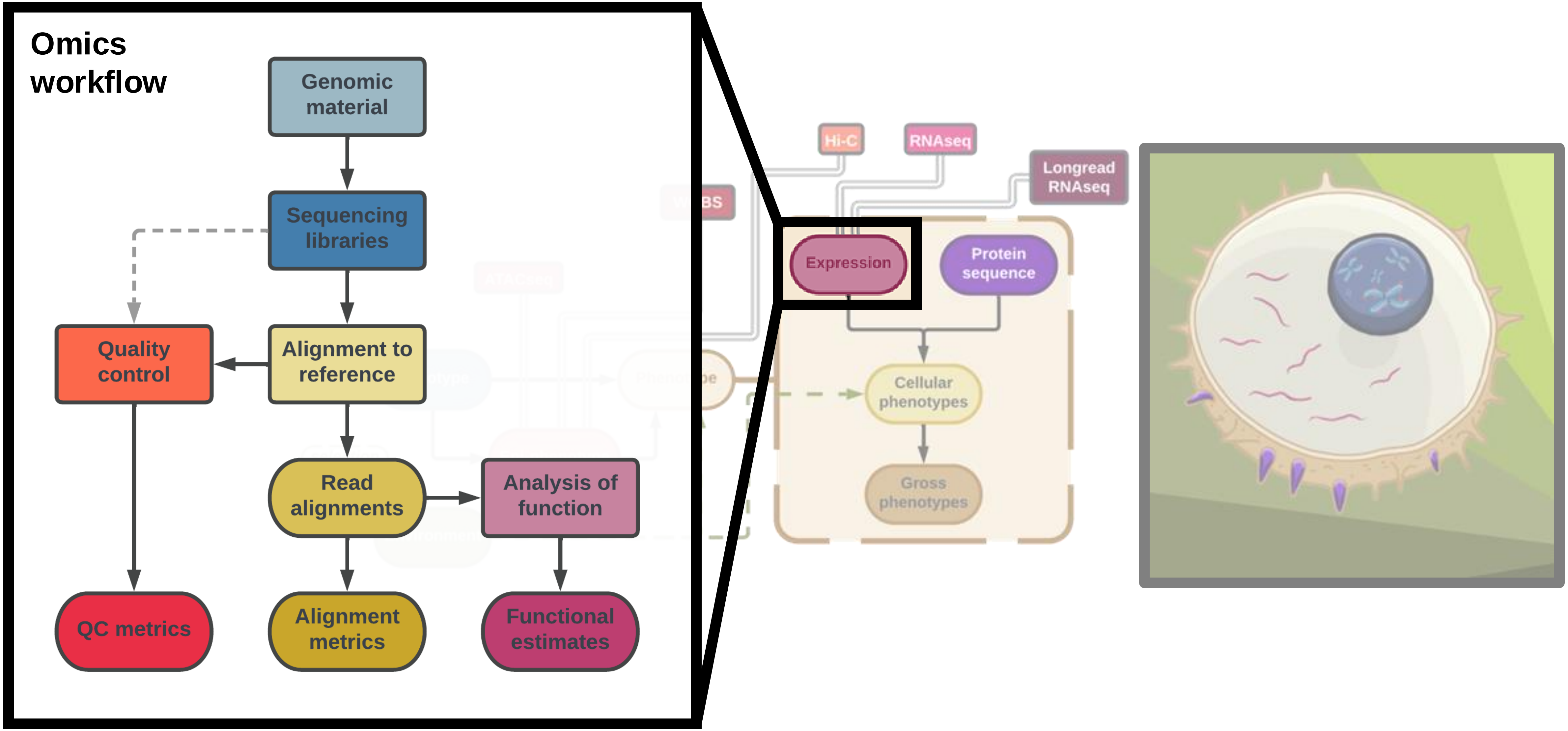


We use –omics assays to study them

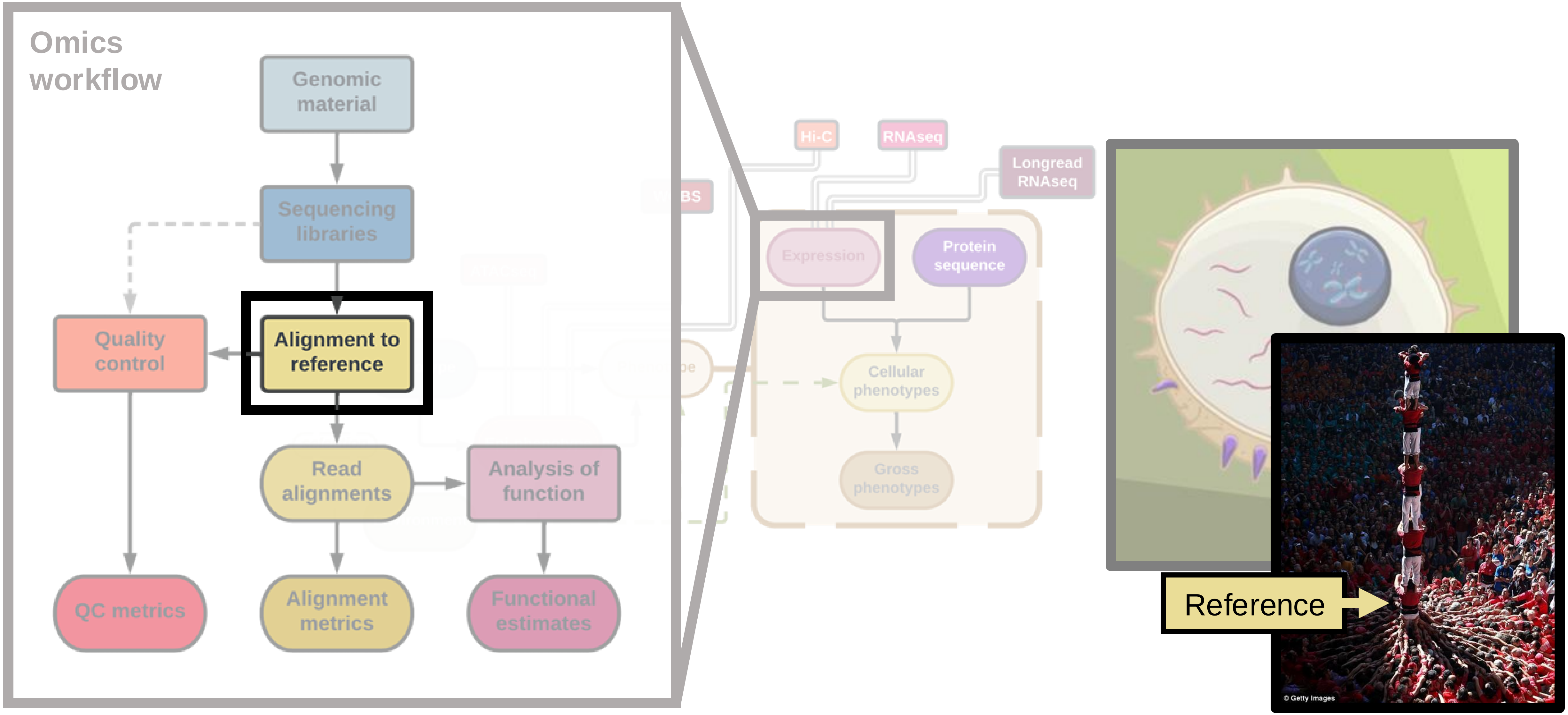




How do –omics assays work?



Probing genotype and function requires reference genome assembly

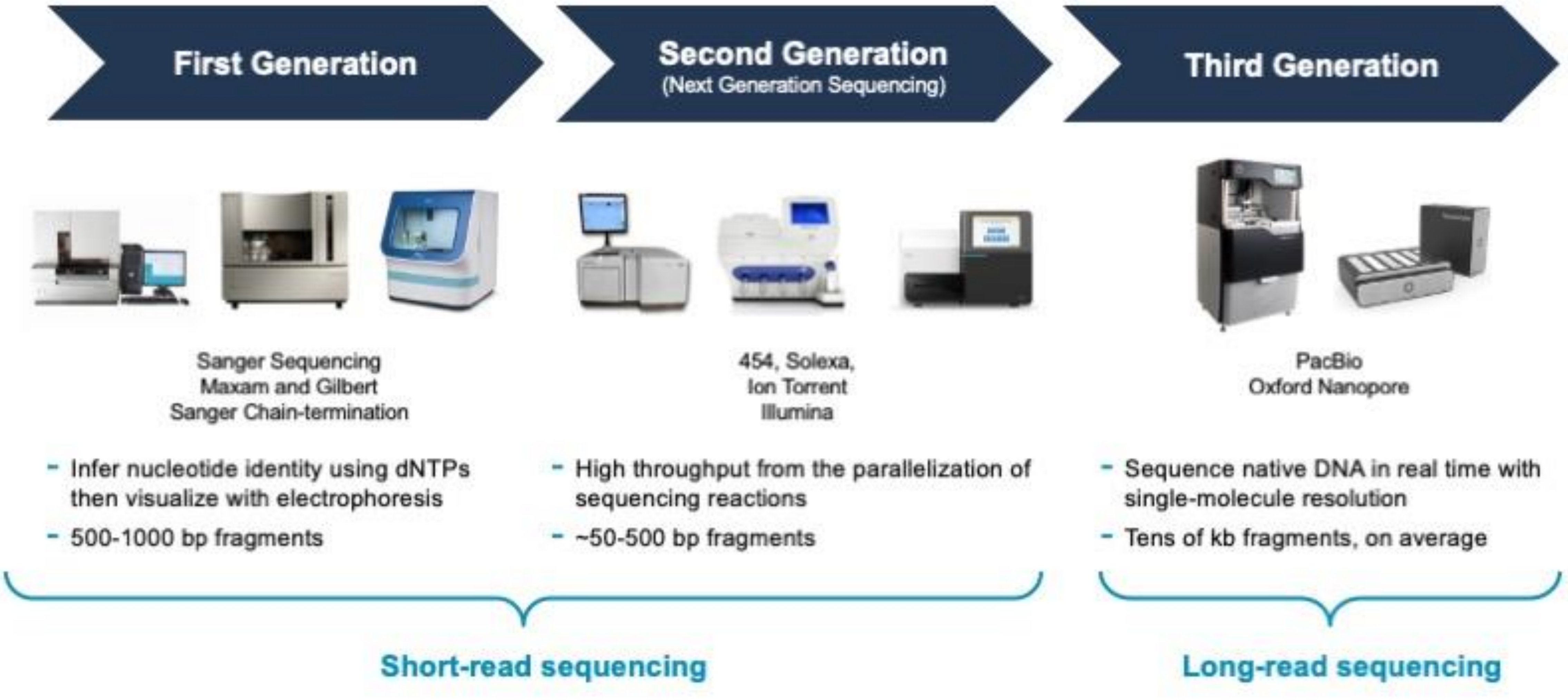


Outline

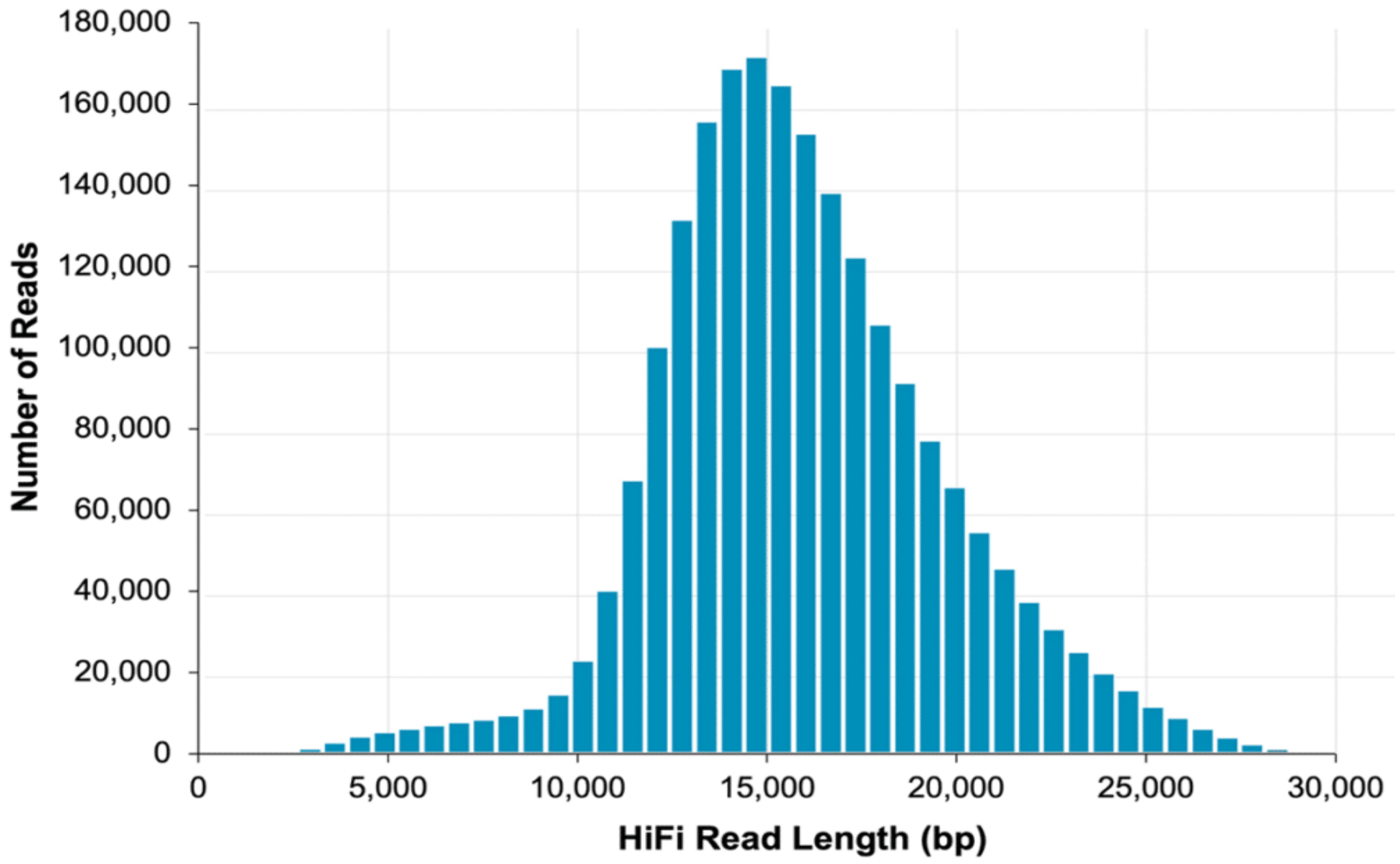
- **Linear genome assembly**
  - Definition of long read sequencing and its applications.
  - Explanation of genome assembly and how long reads improve assemblies.
  - Examples of commonly used long read genome assembly algorithms.
  - Overview of the pb-assembly (Falcon-Unzip) assembly algorithm.

- **Pangenome assembly**
  - Justification
  - Definition of genome graphs
  - Different approaches to construction
  - Utility of genome graphs

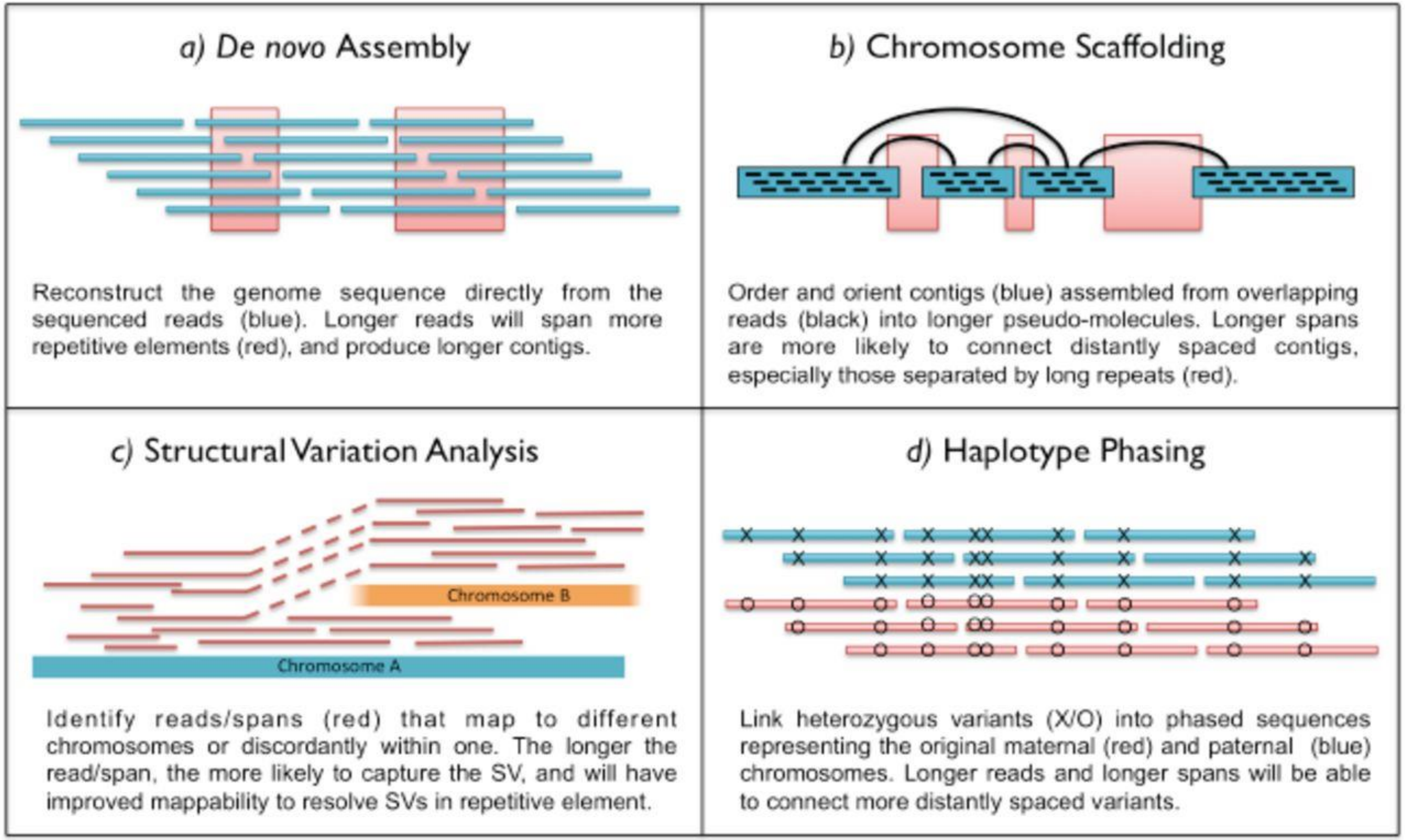
Evolution of DNA Sequencing



PacBio HiFi Read Length Distribution



Long Read Analysis Applications



Long Read Sequencing is getting cheaper

Sequencing technology	Platform	Data type	Read length (kb)		Read accuracy (%)	Throughput per flow cell (Gb)		Estimated cost per Gb (US$)	Maximum throughput per year (Gb)[a]
			N50	Maximum		Mean	Maximum		
Pacific Biosciences (PacBio)	RS II[b]	CLR	5–15	>60		0.75–1.5	2	333–933[c]	4,380
	Sequel	CLR	25–50	>100	87–92	5–10	20	98–195[d]	17,520
	Sequel II	CLR	30–60	>200		50–100	160	13–26[e]	93,440
		HiFi	10–20	>20	>99	15–30	35	43–86[e]	10,220
Oxford Nanopore Technologies (ONT)	MinION/GridION	Long	10–60	>1,000		2-20	30	50–500[f]	21,900 (MinION) 109,500 (GridION)
		Ultra-long	100–200	>1,500	87–98	0.5–2	2.5	500–2,000[f]	913 (MinION) 4,563 (GridION)
	PromethION	Long	10–60	>1,000		50–100	180	21–42[f]	3,153,600
Illumina	NextSeq 550	Single-end	0.075–0.15	0.15		16–30	>30	50–63[g]	>47,782
		Paired-end	0.075–0.15 (×2)	0.15 (×2)		32–120	>120	40–60[g]	>70,080
	NovaSeq 6000	Single-end	0.05–0.25	0.25	>99.9	65–3,000	>3,000	10–35[h]	>1,194,545
		Paired-end	0.05–0.25 (×2)	0.25 (×2)					

Goals of Genome Assembly

- **Maximize and evaluate completeness of genome sequence for an organism.**
  - As few gaps in the sequence as possible
  - As accurate of a sequence as possible

- **Assemble and organize the sequence.**
  - Into contigs
  - Into scaffolds
  - Into chromosomes

De Novo Genome Assembly Process and Outputs

- Algorithms are used to find accurate overlaps between reads. The consensus of a set of overlapping read sequences is called a contig.

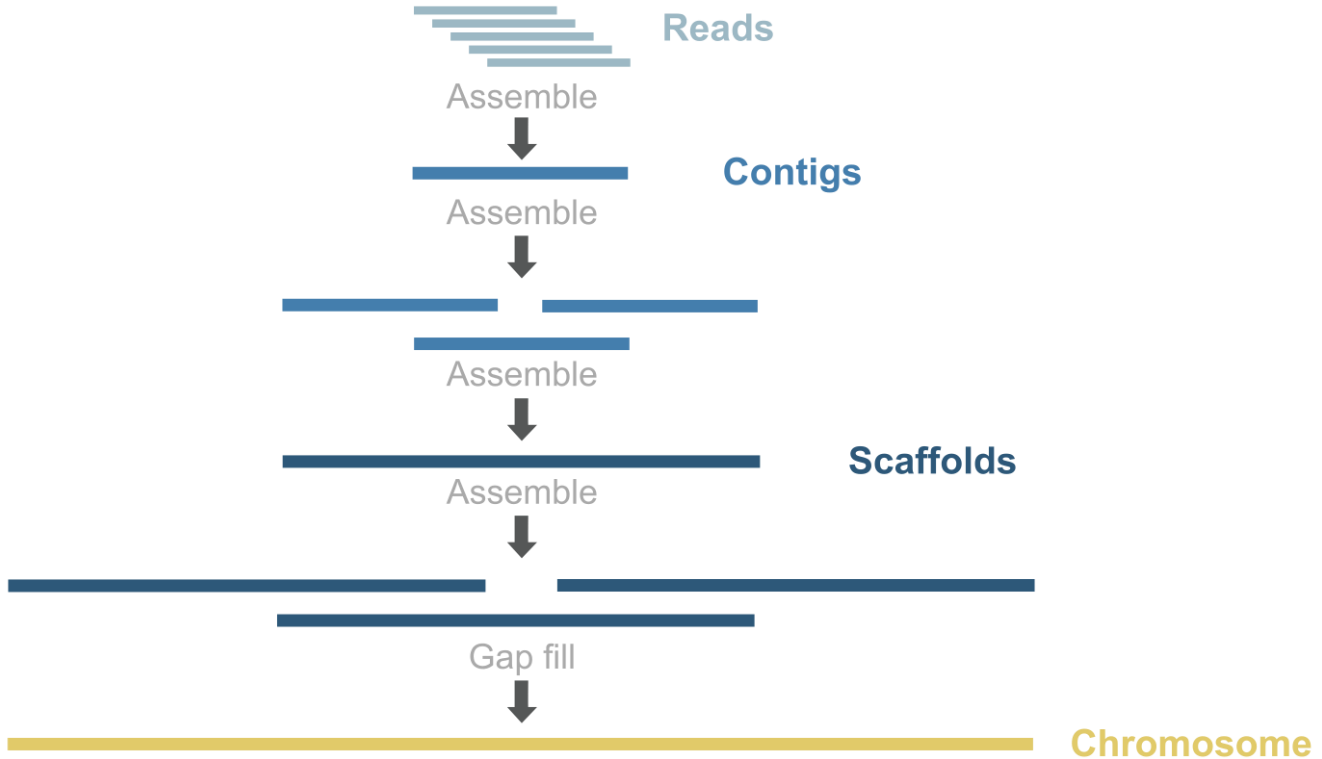
- Contigs are segments of the genome that have gaps in between. In most instances, we do not know how the contigs are to be ordered and oriented in respect to one another.

- We can use additional methods to order and orient contigs, fill in gap regions, and chain contigs together into larger sequences called scaffolds.

- Scaffolding can be an additional step performed by the assembly algorithm, but it usually involves a separate process using BioNano, Hi-C, Oxford Nanopore data. or other data types.

- Additional methods can be used to organize scaffolds into more complete chromosomal level assemblies.

- Contigs and scaffolds are typically available in FASTA and/or FASTQ formats.



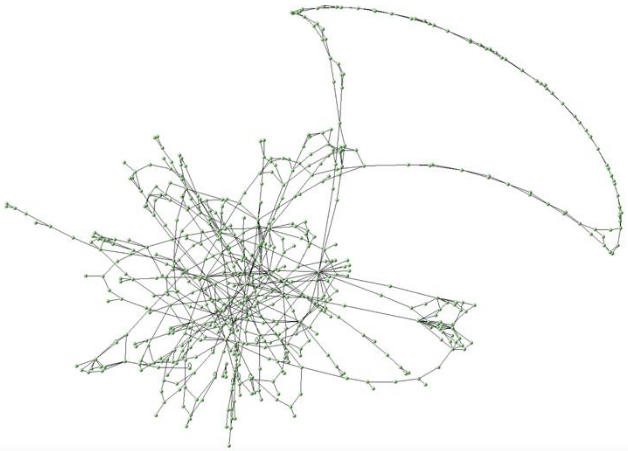
Two main assembly methods

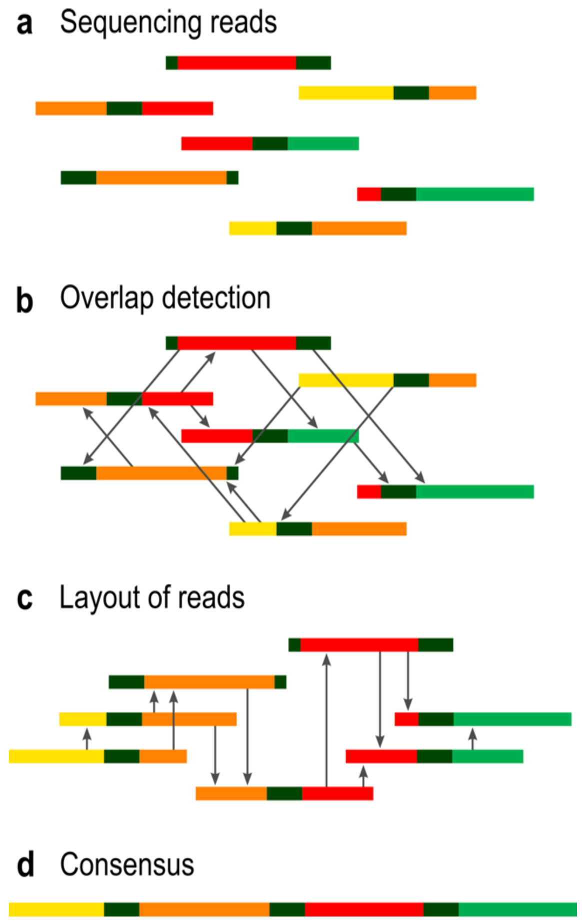
- **Overlap Layout Consensus Assemblers** – Mainly used for long reads
  - Construct overlap graph directly from reads, eliminating redundant reads; trace path for assembly.
  - Examples: pb-assembly (Falcon), Canu, Hifiasm

- **de Bruijn graph-based Assemblers** – Mainly used for short reads
  - Construct k-mer graph from the reads; original reads are discarded.
  - Trace a path through the graph to arrive at the assembly.
  - Examples: MEGAHIT, ABySS

Overlap Layout Consensus Graph

- All read vs. all read alignment and identify all possible overlaps between reads.
- The overlap relationship between reads is captured in a large assembly graph.
- The graph is refined to correct errors and simplify.
- Find the best path through the graph and traverse each node in the graph once.
- Output the consensus of the path as the assembly.





Challenges in Genome Assembly

- **Repeats**
  - Repetitive regions can make it more difficult to determine read placement leading to misassemblies, which are errors in the construction of the assembly.
  - Long reads can help this by traversing the longest repeats in the genome.

- **Sequencing Errors or Assembly Errors**
  - Base errors in sequencing reads or consensus errors introduced by assemblers can confound the assembly process.

- **Heterozygosity**: Presence of different alleles at the same loci in homologous chromosomes
  - Alleles from the same locus are more likely to be mistaken as sequences from different loci.
  - The assembler may incorporate two different contigs that actually represent the same regions of the genome.
  - This might be desirable if you are seeking haplotype separated diploid or polyploid assemblies.

- **Contamination**
  - Contamination in the sequencing data can lead to contigs of contamination in your final assembly.
  - Adapter contamination of the read data can impact assembly results.

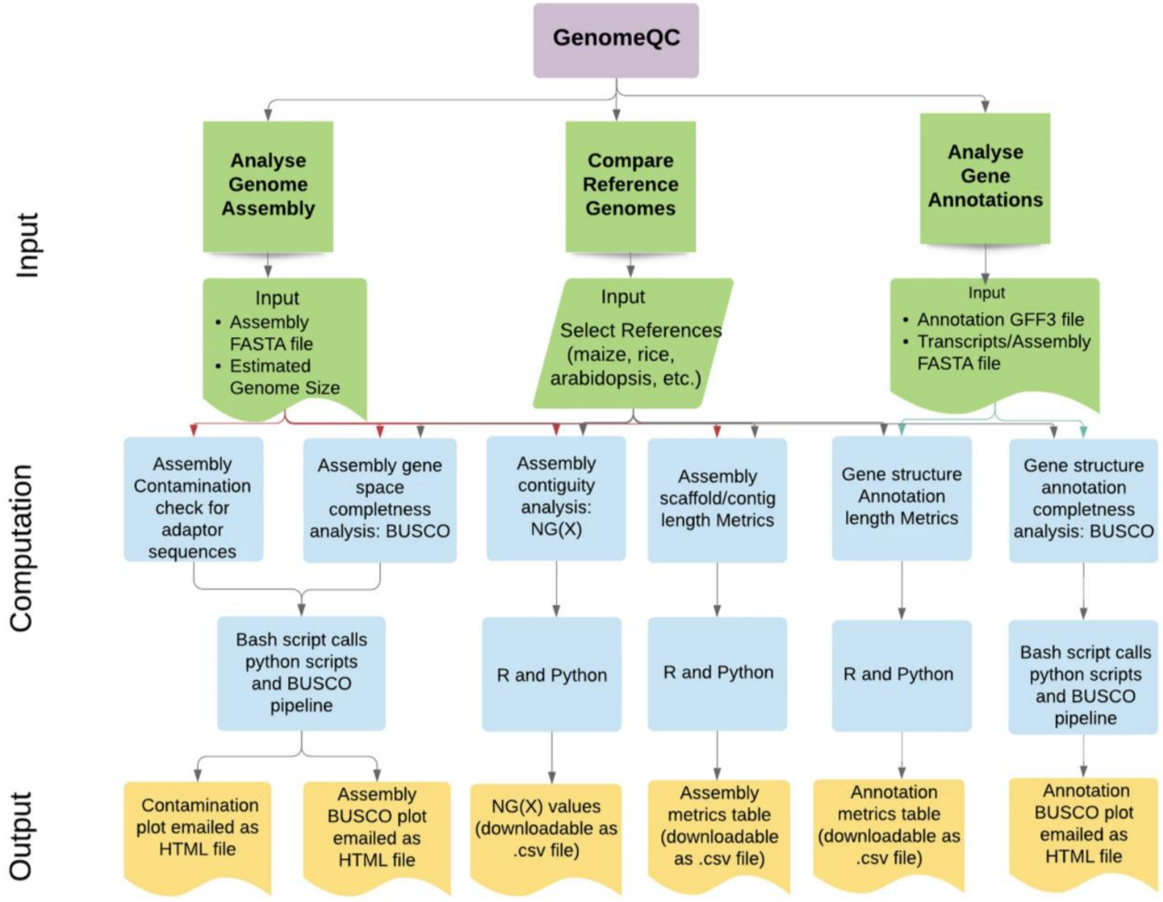
Methods to QC Genome Assemblies

- **Assembly scaffold/contig length metrics**
  -N50 length, avg. length, # contigs/scaffolds

- **Compare assembly to Reference Genome**
  -Identify misassemblies
  -Identify real SNPs and SVs (indels, translocations, duplications)

- **Evaluate against an optical map**: Ordered, genome wide high resolution restriction map
  -Identify misassemblies
  -Scaffold contig assemblies

- **Busco Analysis**: Evaluation of the assembly against a set of single-copy orthologs present in 90% of species of a particular group.
  -What percentage of core genes does your assembly contain?
  -Measure of assembly completeness.



N50 Contig/Scaffold Length

N50 size

Def: 50% of the genome is in contigs larger than N50

Example: I Mbp genome

N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:

N50 values are only meaningful to compare when base genome size is the same in all cases

Assemblers for PacBio Data

- **De novo Assemblers**
  - Pb-assembly (Falcon/Falcon-Unzip)
  - Canu/HiCanu

- **Trio Based Assemblers:** Use short reads from parents to partition child reads by maternal/paternal haplotypes prior to assembly.
  - TrioCanu
  - Hifiasm

- **Reference Assisted Assemblers:** Using the reference to assist in building the reads into contigs/scaffolds. This can bias assembly results towards the reference used.
  - RefKA

- **Hybrid Assemblers :** Use short and long reads. More effective for assembly of complex genomes (e.g. Plants)
  - Masurca
  - PBcR

PB Assembly (Falcon/Falcon-Unzip) of HiFi Data

- **Falcon Assembly:**
  -Input is PacBio HiFi data in FASTA format.

  -Output is haplotype-fused assembly in FASTA format and set of associated contigs (SVs from primary contig assembly).

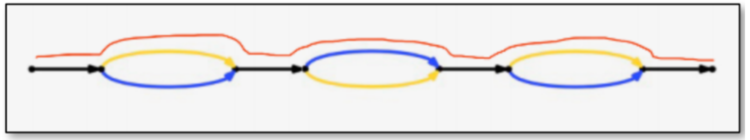
- **Falcon-Unzip:**
  -Align reads to contigs and phases the reads using heterozygous SNPs.

  -SNPs are used to separate the haplotypes into partially phased primary contigs and fully phased haplotigs.

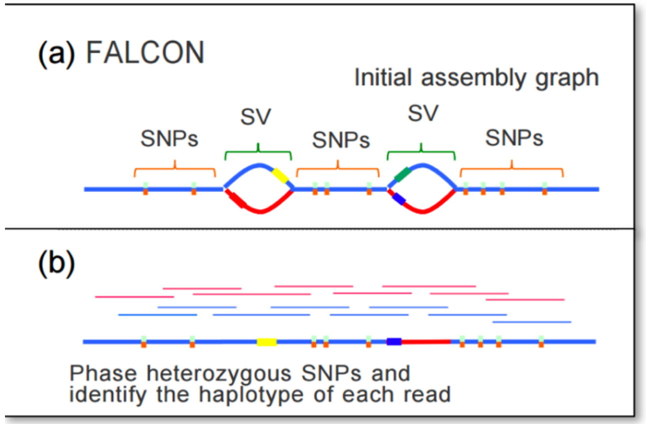
  -There are switch errors in the output between maternal and paternal haplotypes.

  -Output is a FASTA file of primary contigs and a FASTA file of haplotigs.

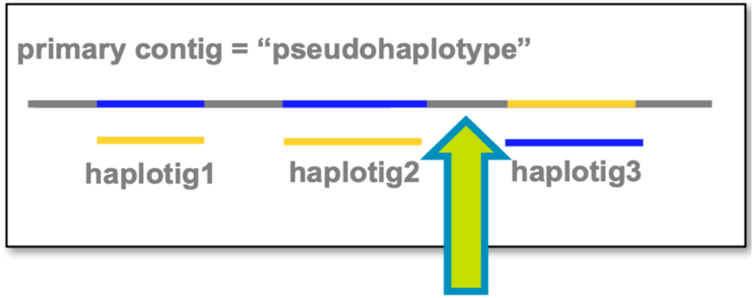
- PacBio data for diploid individual (no trio)
- Phase PacBio reads using SNPs identified in initial assembly graph
- Output phased and collapsed regions in high contiguity contigs



Weisenfeld et al. 2017



PSEUDOHAPLOTYPE AND HAPLOTIGS

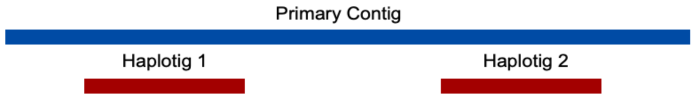


"Phase/Haplotype Switch"

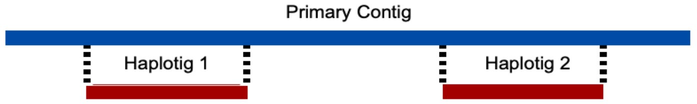
Falcon-Phase

FALCON-PHASE WORKFLOW

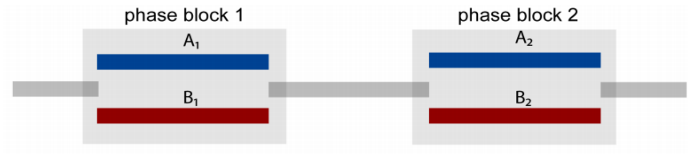
Input: FALCON-Unzip assembly & HiC data



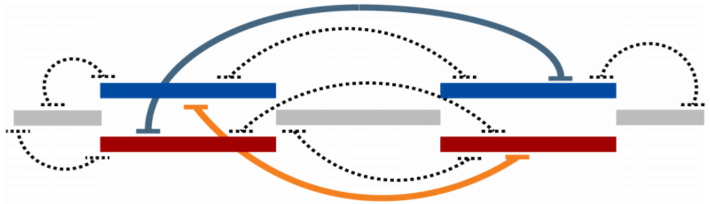
1. Identify **<u>haplotig placement</u>** on primary contigs



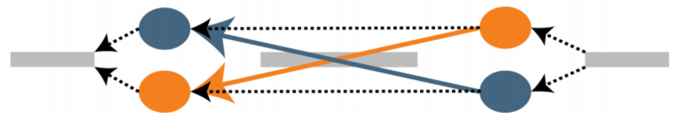
2. **<u>Mince</u>** primary contigs: separate **<u>haplotig pairs</u>** and **<u>collapsed haplotypes</u>**



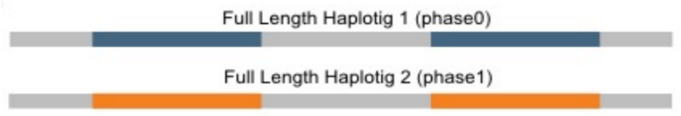
3. **<u>Map</u>** paired HiC reads to minced contigs and generate normalized **<u>contact matrix</u>**



4. **<u>Phase</u>** haplotigs along each primary contig



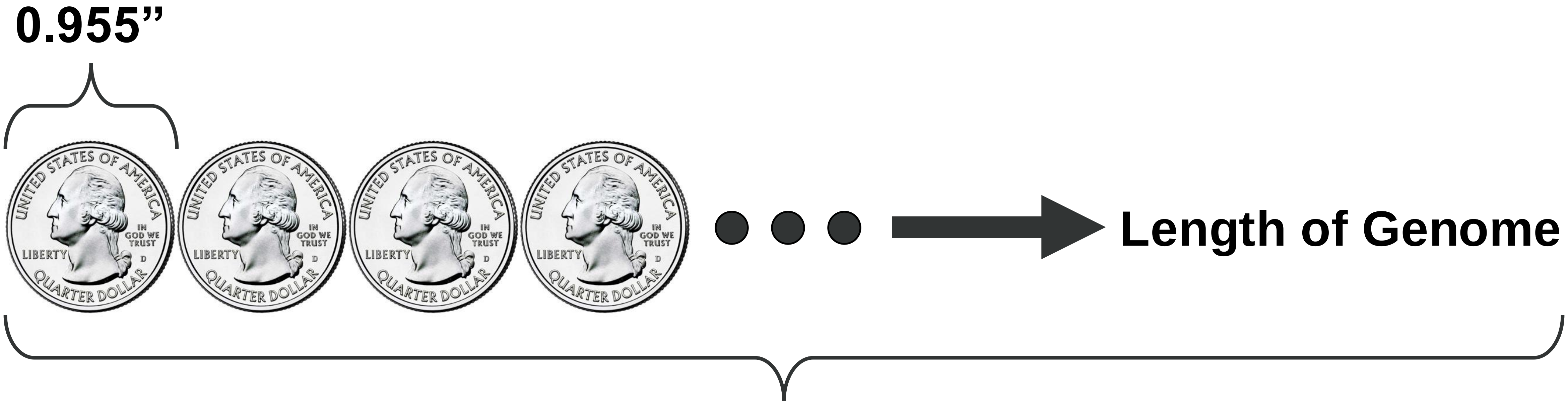
Output: phased full length haplotigs



Pangenomes

… and why you should consider using one

How big is a human genome?



How long would this be?

Immense size of genome

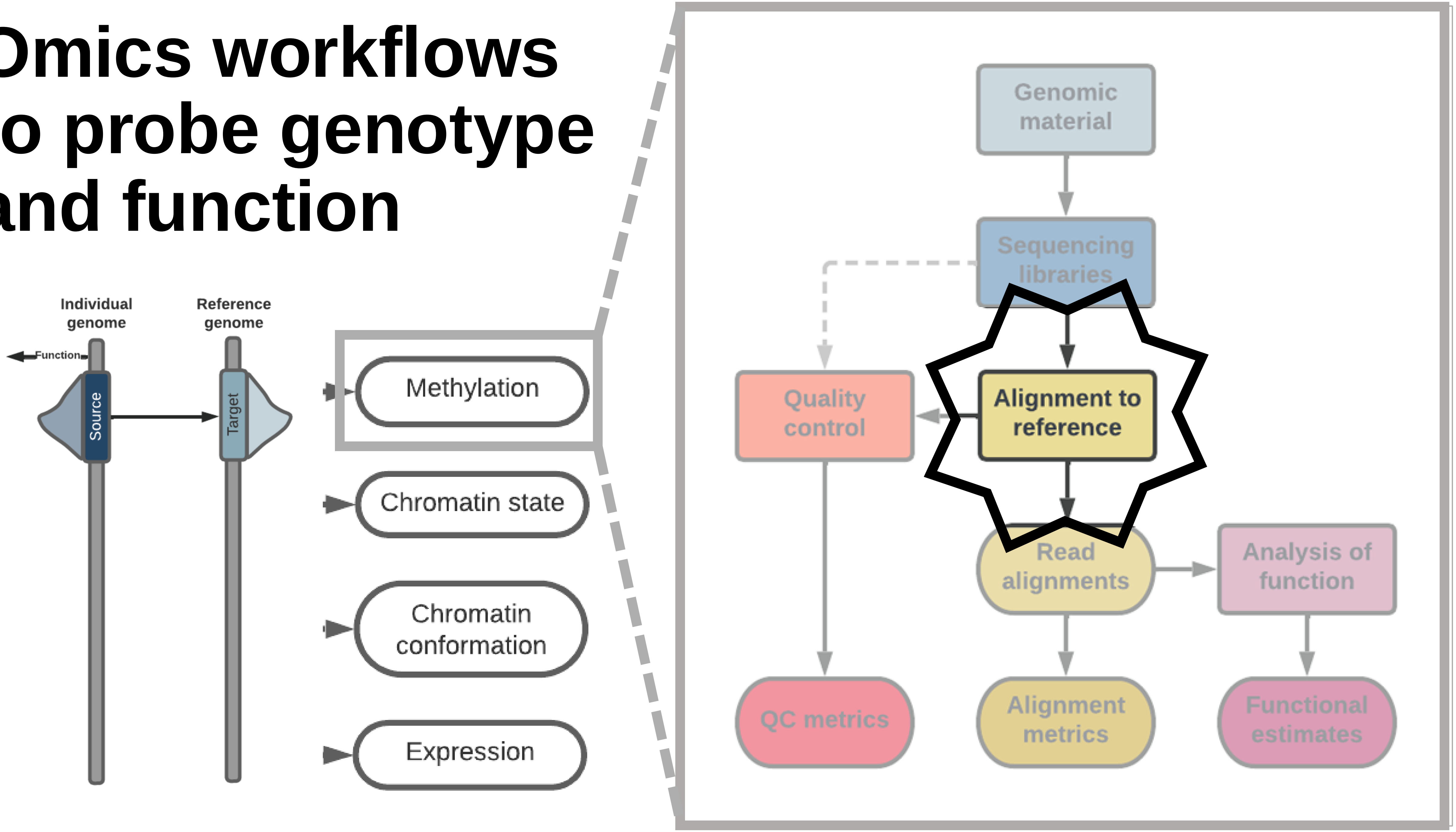
It would wrap 1.9 times around the world!!!

A single short read fragment at this
scale is about 40’
35-40’~ length of school bus

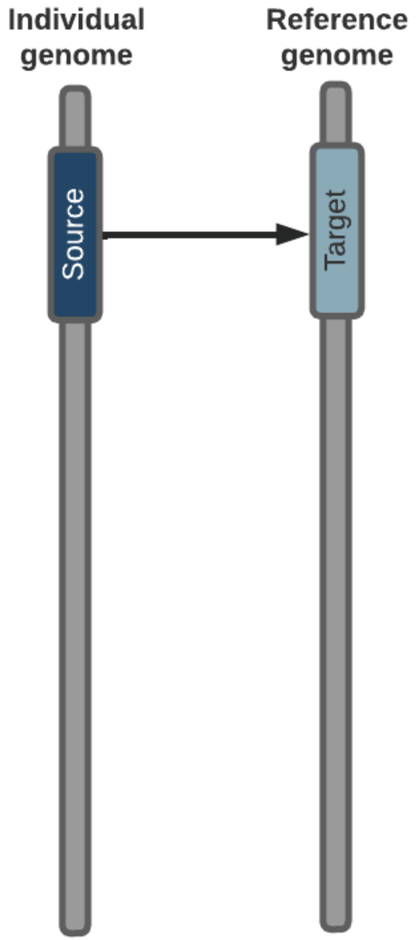
LR is about ¼ mile
36 school buses

Putting together something that long is hard!

Omics workflows to probe genotype and function

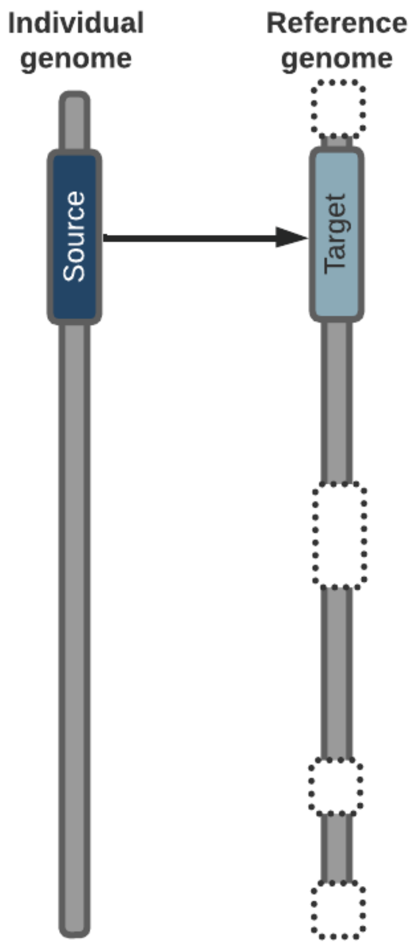


A Need to Modernize the Human Reference Genome



A Need to Modernize the Human Reference Genome

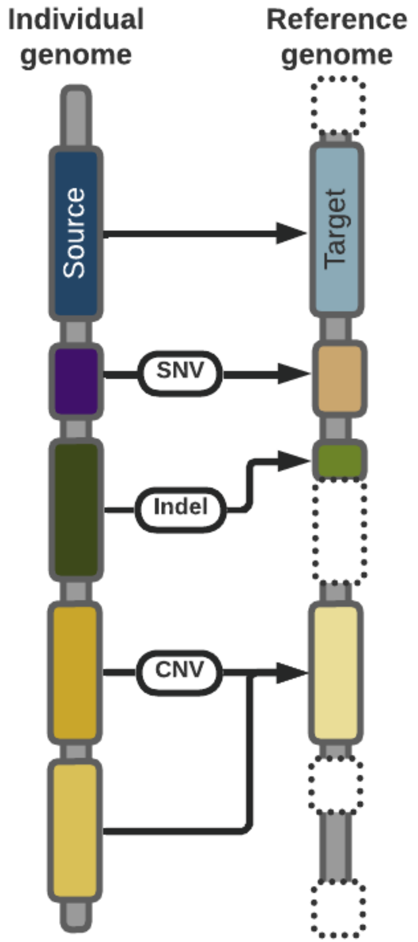
- The current reference is **not complete**



A Need to Modernize the Human Reference Genome

- The current reference is **not complete**

- The current structure is a linear haplotype**, largely representing a single individual.**

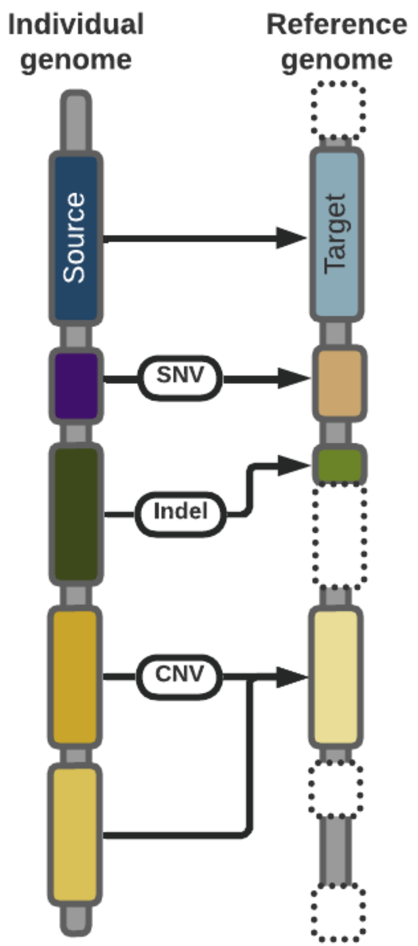


A Need to Modernize the Human Reference Genome

- The current reference is **not complete**

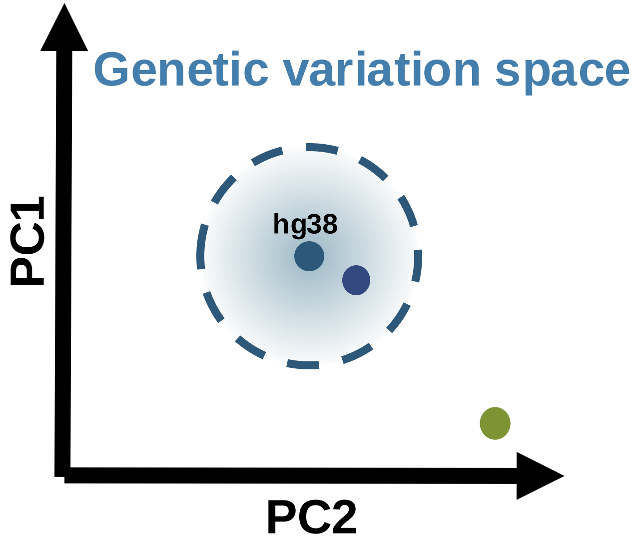
- The current structure is a linear haplotype**, largely representing a single individual.**

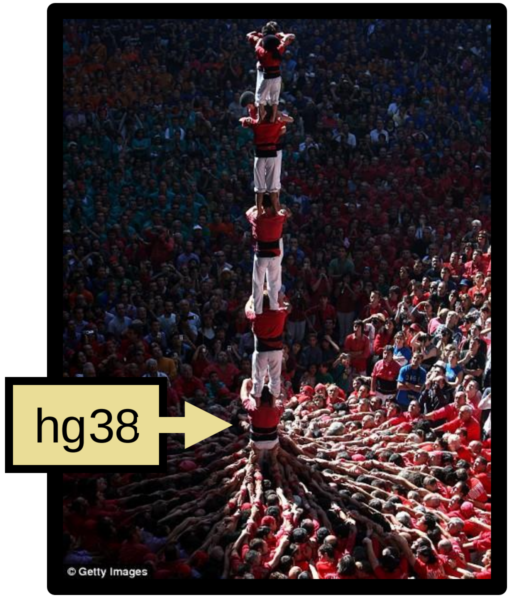
- **Mapping limitations of short reads and inherent reference biases** means *<u>we have missed more than 70% of structural variants</u>* in traditional whole-genome sequencing studies



Draft Human Pangenome

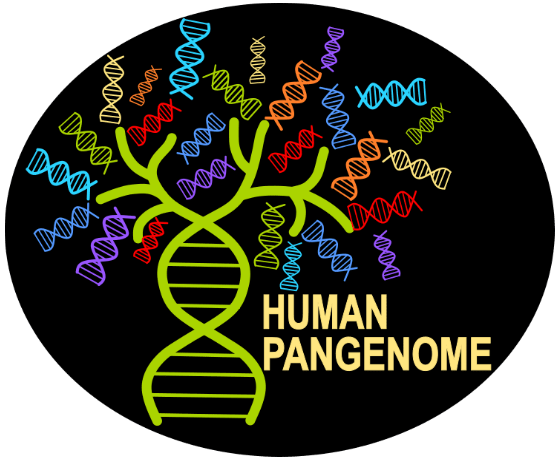
Improve upon the limitations of reference genomes

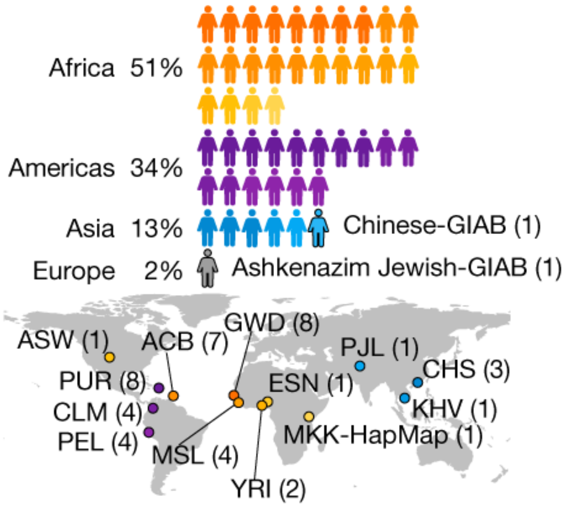


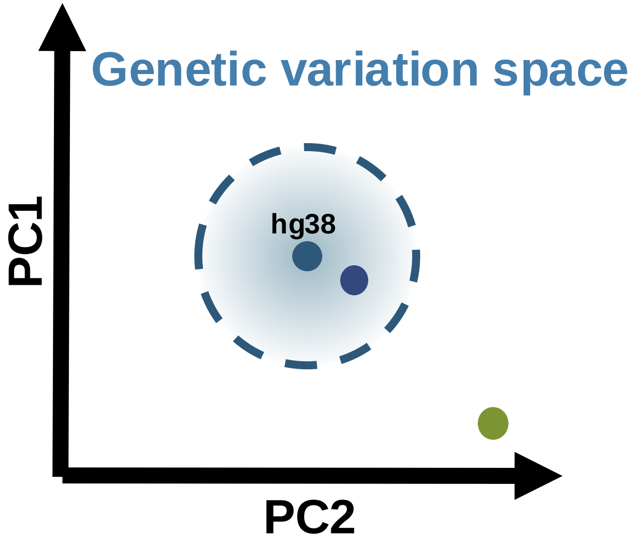


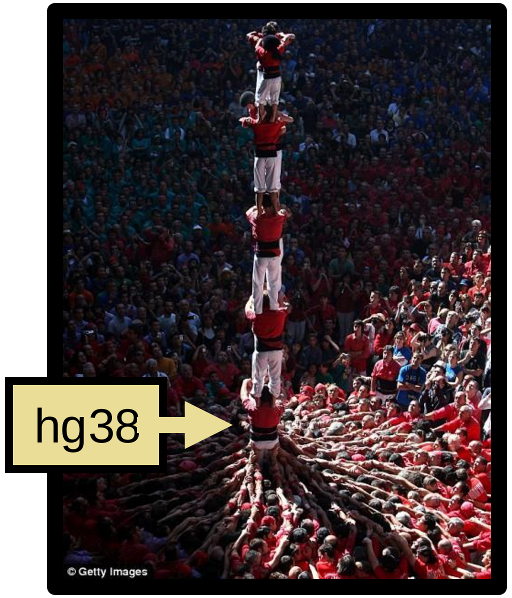
Draft Human Pangenome

Improve upon the limitations of reference genomes









Article | Open access | Published: 10 May 2023

A draft human pangenome reference

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, ... Benedict Paten

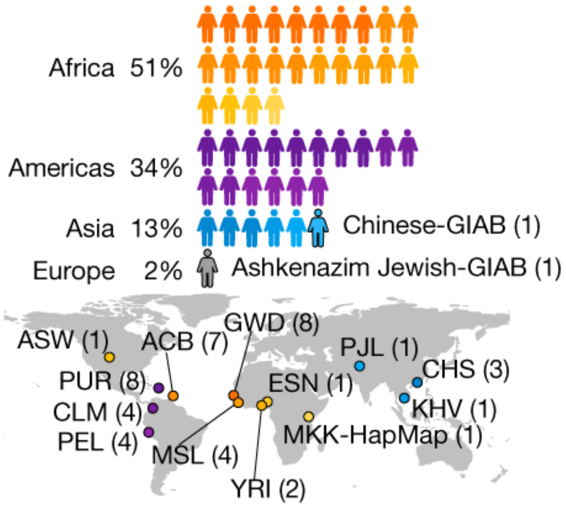
+ Show authors

Nature **617**, 312–324 (2023) | Cite this article

Draft Human Pangenome

Improve upon the limitations of reference genomes

HUMAN PANGENOME



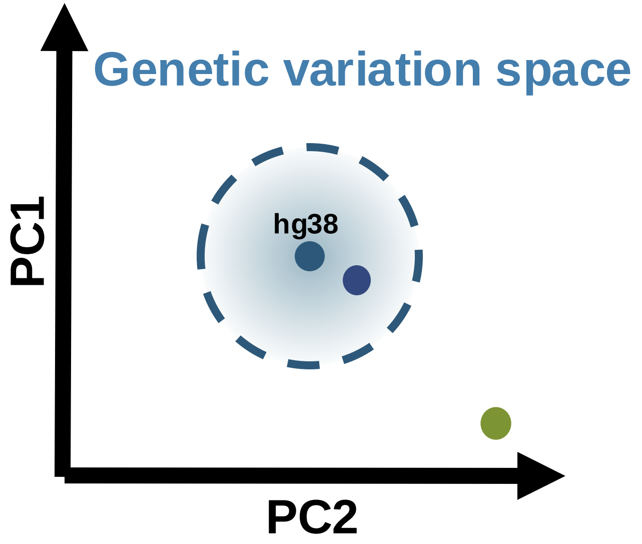
Article | Open access | Published: 10 May 2023

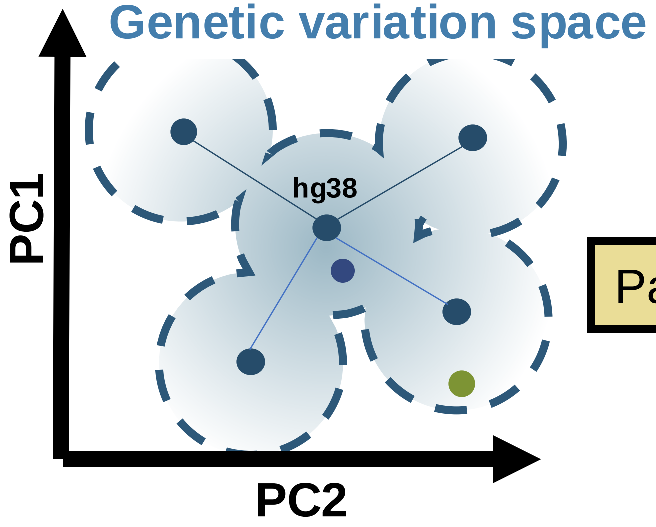
A draft human pangenome reference

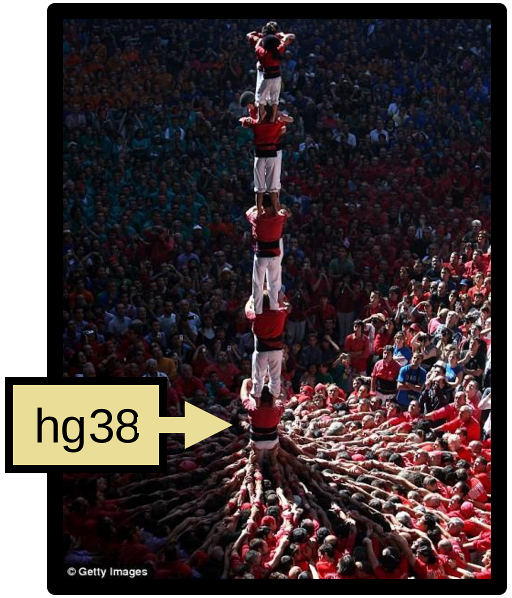
Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, ... Benedict Paten

+ Show authors

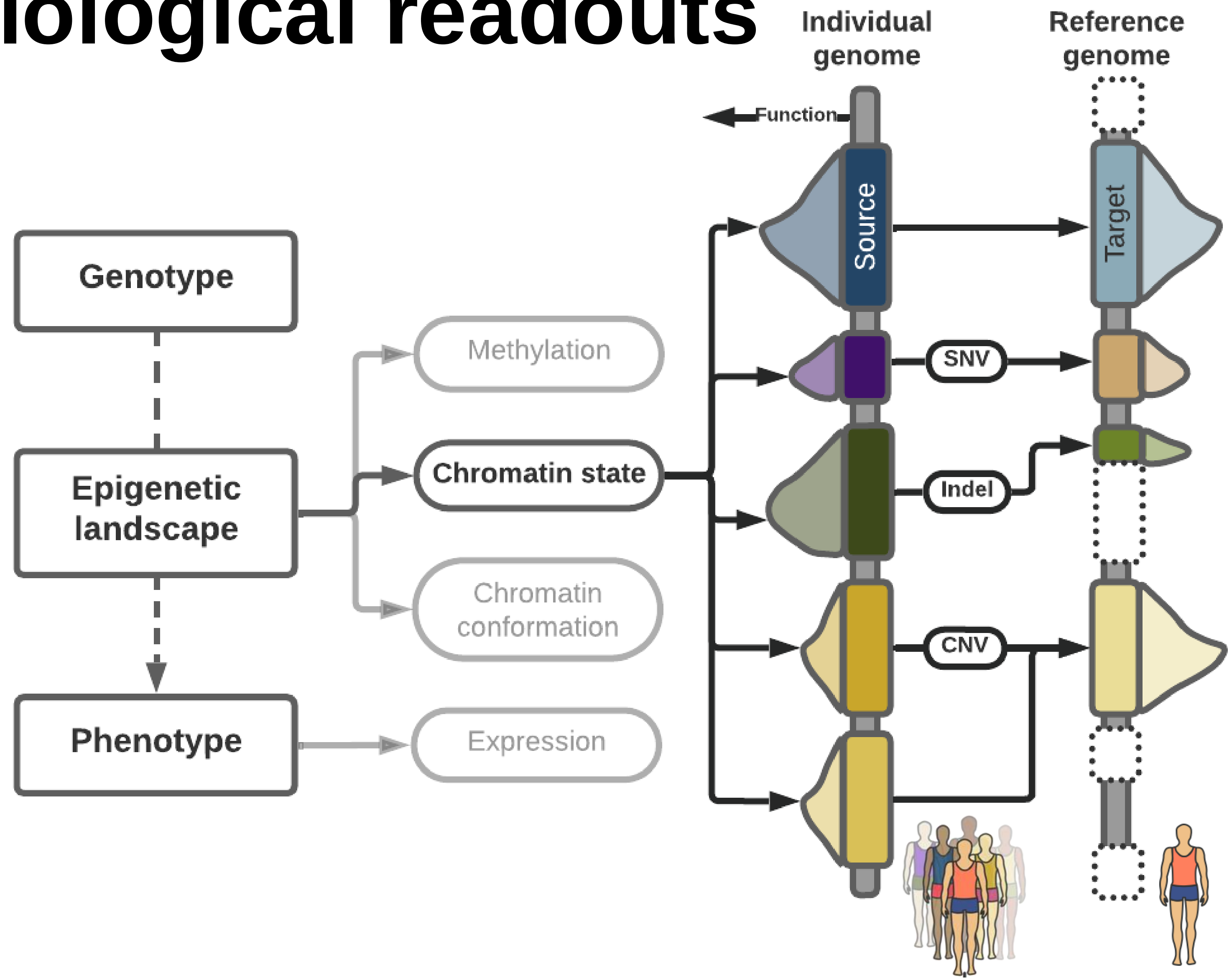
Nature **617**, 312–324 (2023) | Cite this article



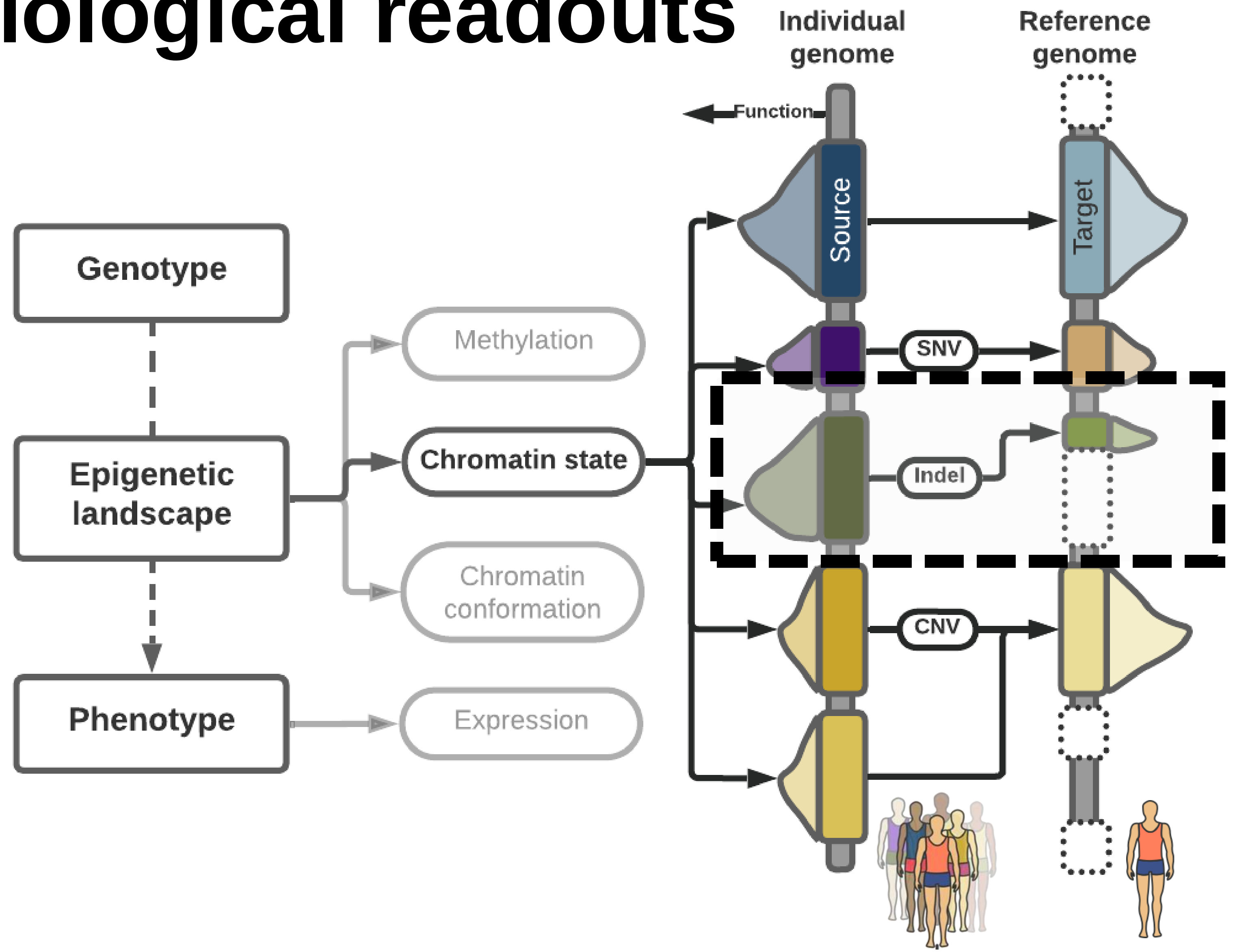




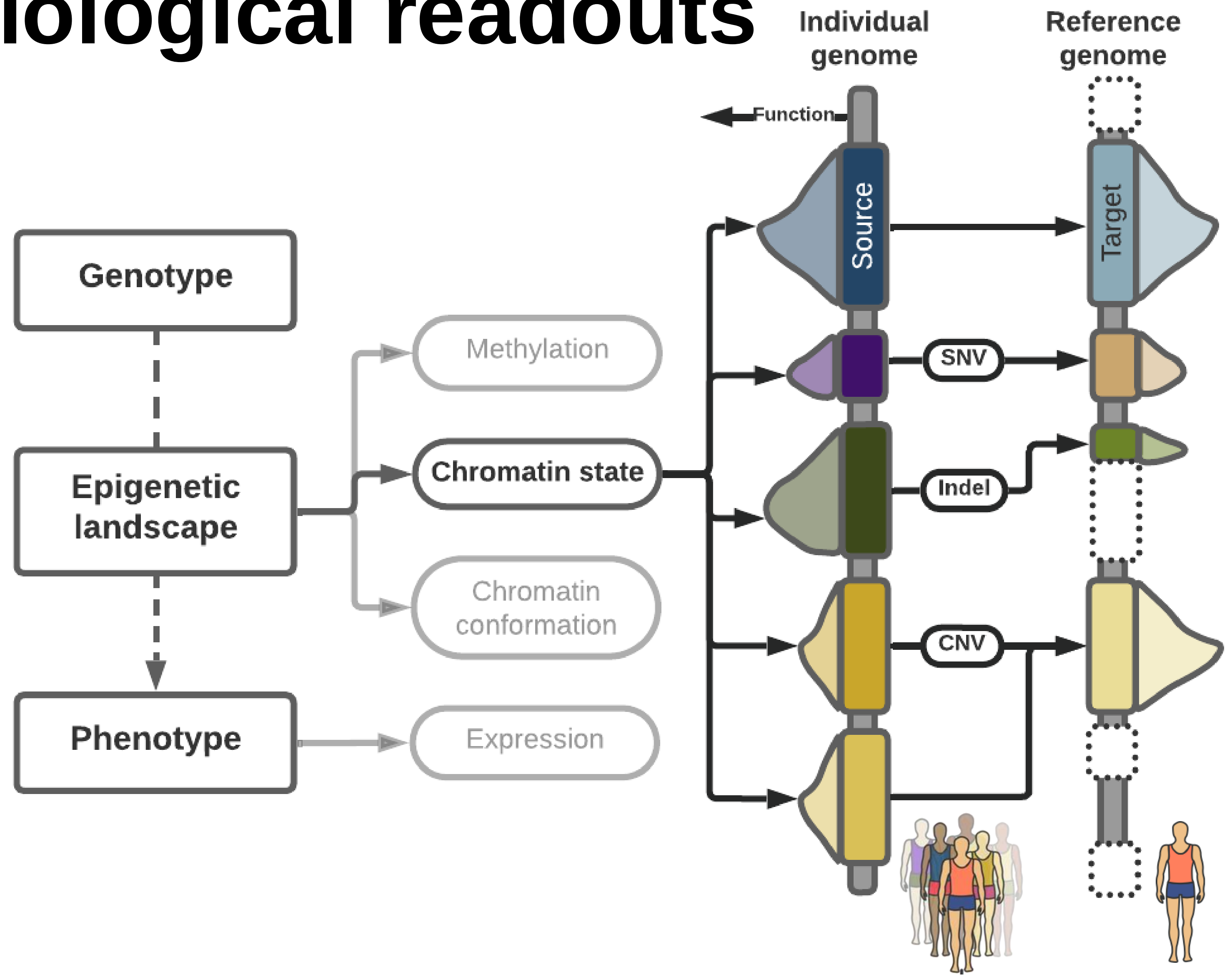
Use of under-representative reference distorts biological readouts

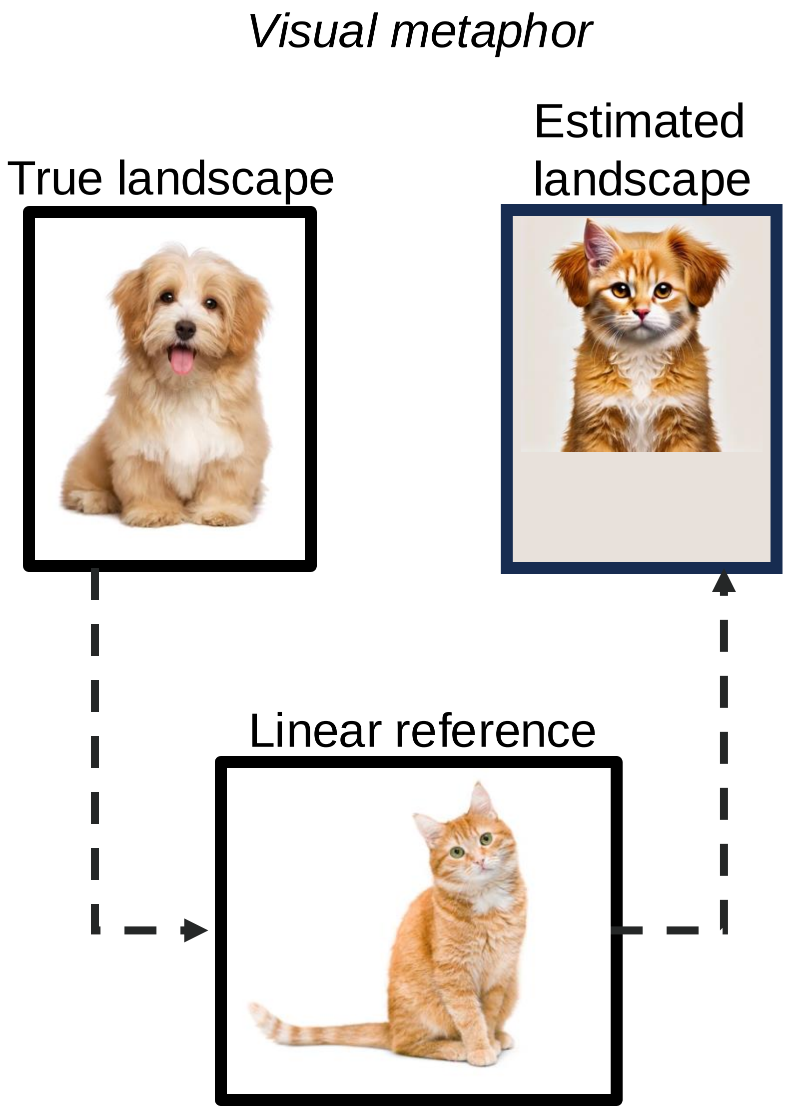


Use of under-representative reference distorts biological readouts

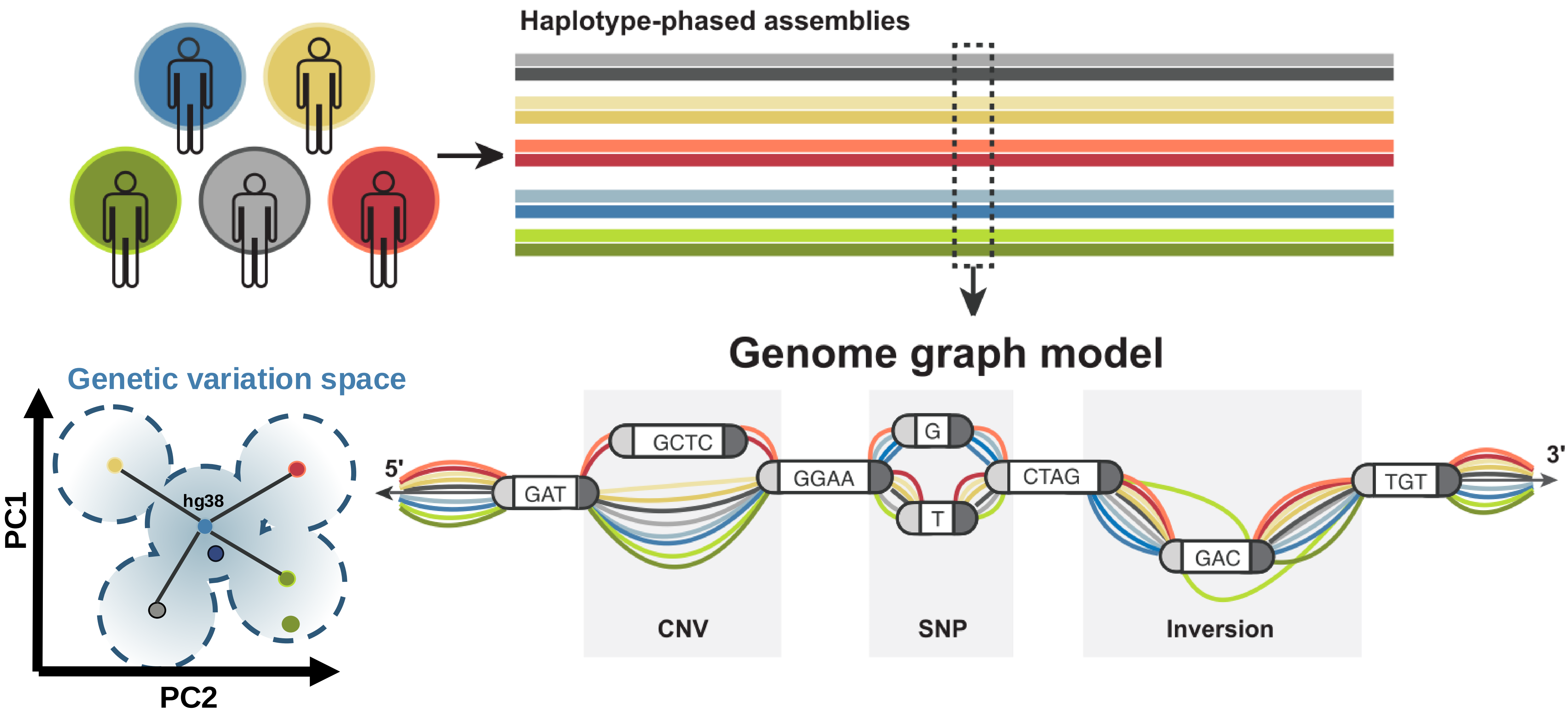


Use of under-representative reference distorts biological readouts





The Pangenome Graph



The Pangenome Graph

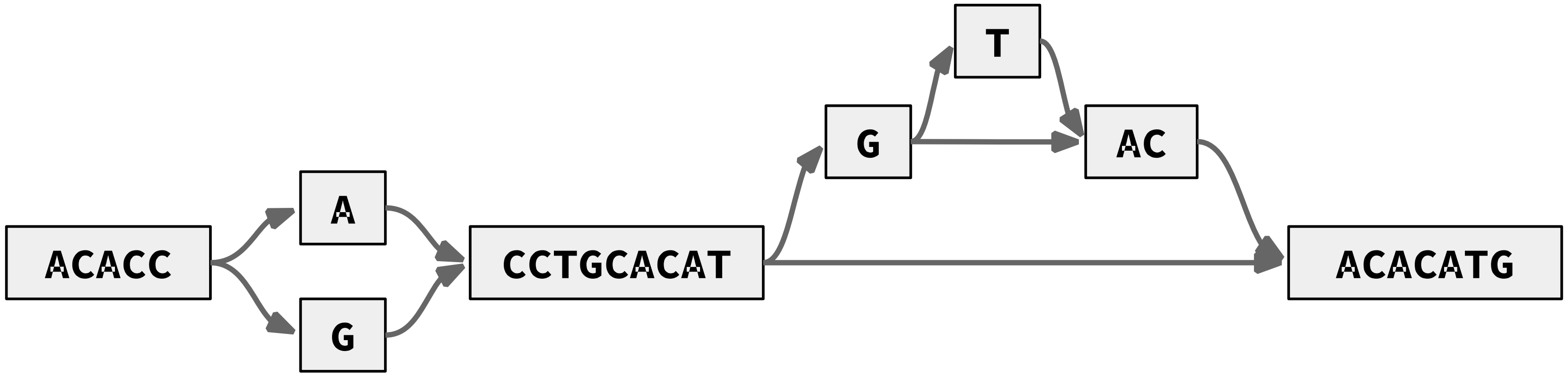
High-quality assemblies

```
ACACCACCTGCACATGACACACATG
ACACCACCTGCACATACACATG
ACACCACCTGCACATGACACACATG
ACACCACCTGCACATACACATG
ACACCACCTGCACATGACACACATG
ACACCGCCTGCACATGACACACATG
ACACCGCCTGCACATGTACACACATG
ACACCGCCTGCACATGACACACATG
```

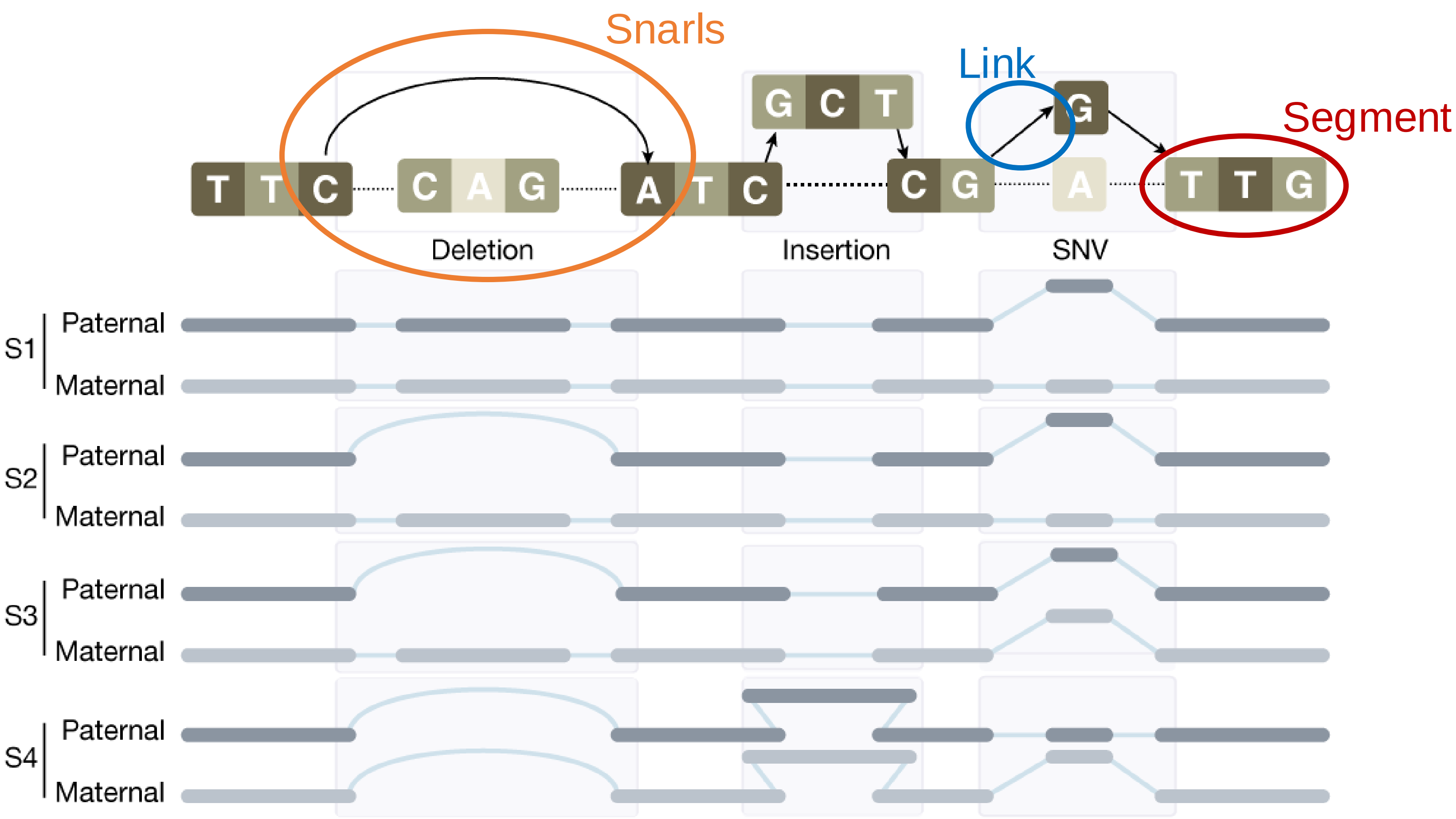
Alignment

```
ACACCACCTGCACAT GACACACATG
ACACCACCTGCACAT ---ACACATG
ACACCACCTGCACAT GACACACATG
ACACCACCTGCACAT ---ACACATG
ACACCACCTGCACAT GACACACATG
ACACCGCCTGCACAT GACACACATG
ACACCGCCTGCACAT GTACACACATG
ACACCGCCTGCACAT GACACACATG
```

Pangenome Graph



Genome graphs

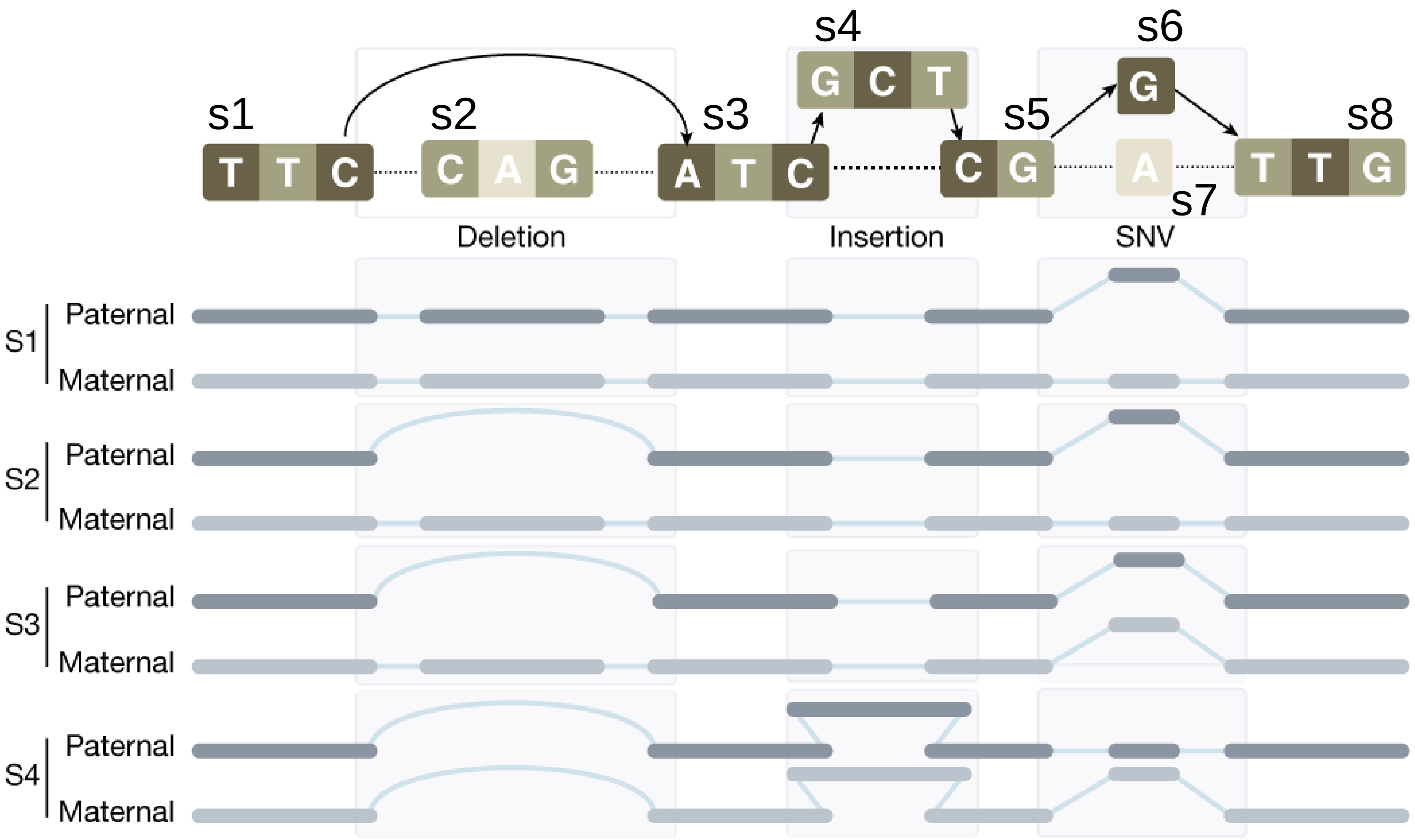


<u>Redundant vocabulary:</u>

Node = Vertex = Segment

Link = Edge

Genome graphs



s1 > s2 > s3 > s5 > s6 > s8

s1 > s2 > s3 > s5 > s7 > s8

s1 > s3 > s5 > s6 > s8

s1 > s2 > s3 > s5 > s7 > s8

Redundant vocabulary:

Node = Vertex = Segment

Link = Edge

The Pangenome Graph

Benedict paten

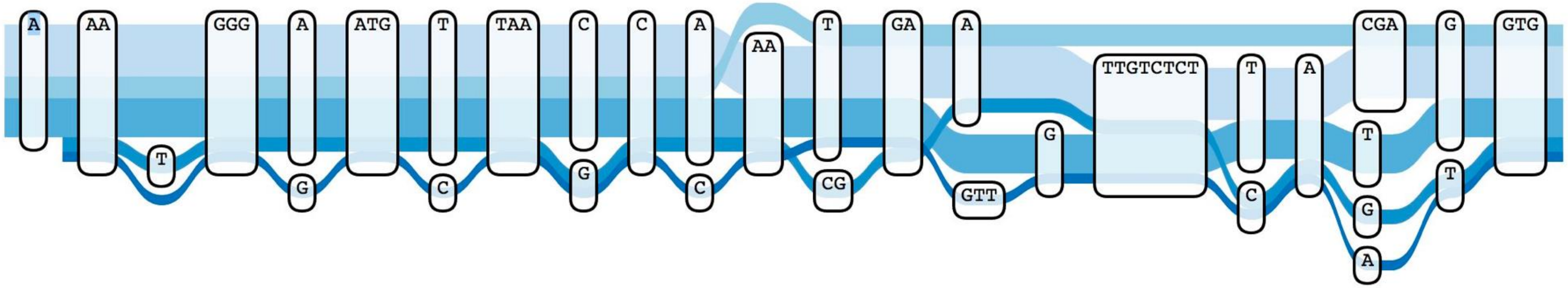
Heng Li

Tobias Marschall

Erik Garrison

Mark Chaisson

Ira Hall



Minigraph (Li et al., 2020)

Minigraph-Cactus (MC)

PanGenome Graph Builder (PGGB)

Building pangenome – approaches

- Iterative building – *affected by order of genome addition*
  - **Minigraph**
  - **Minigraph-Cactus**

- Simultaneous building
  - **PGGB**

Building pangenome – iterative

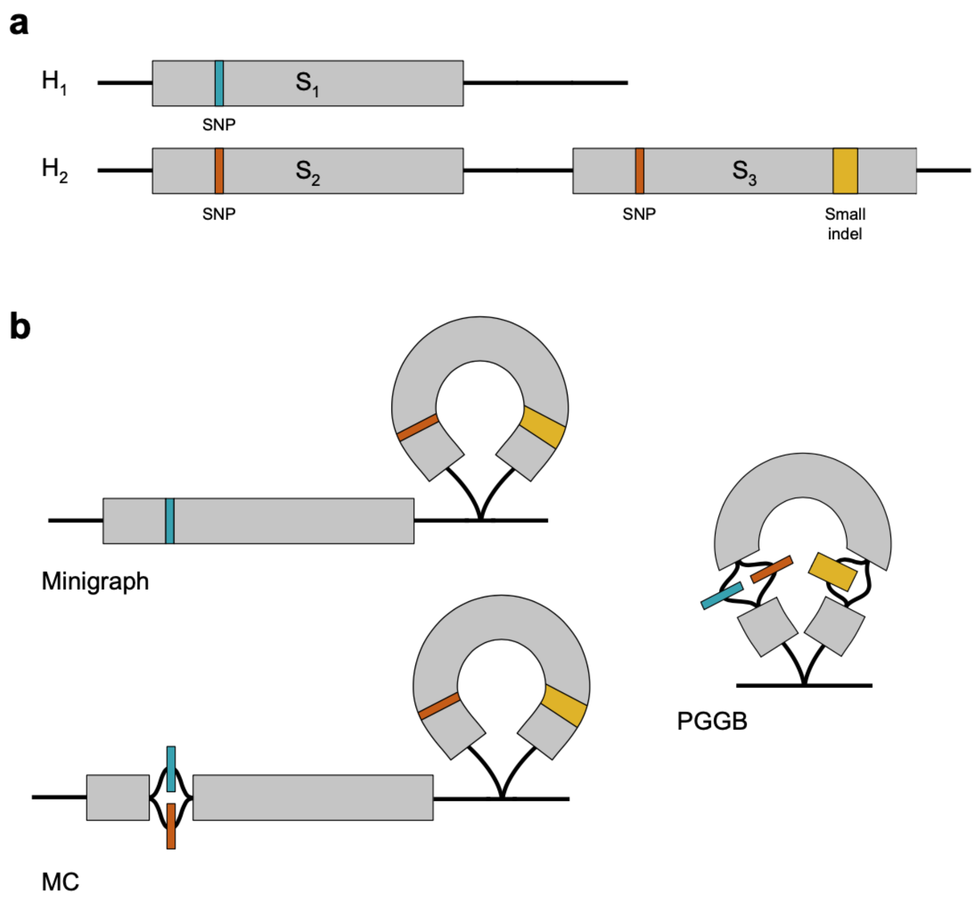
- **Minigraph**
  - Only considers structural variants >= 50bp
  - Does not consider sequence variants

- **Minigraph-Cactus** (MC)
  - 1. builds Minigraph pangenome
  - 2. performs base level sequence alignment of syntenic regions with Cactus.
  - 3. incorporates sequence variants into graph

Building pangenome – simultaneous

- **PGGB**
  - First performs everything to everything alignment
  - Compresses into a graph by partial order alignment

Building pangenome – approaches



- Iterative building
  - **Minigraph**
    - Pros: Usable, Simple
    - Cons: Only captures structural variants, biased by order
  - **Minigraph-Cactus**
    - Pros: Usable, Captures all types of variation
    - Cons: Intentionally reduced complexity, Not a lossless graph, biased by order

- Simultaneous building
  - **Pangenome Graph Builder**
    - Pros: Lossless graph, unbiased by order, lends itself best to universal coordinate system
    - Cons: Much larger in size to the point of rendering it unusable for many applications
      - *“the inclusion of heterochromatic sequences in the PGGB graph made read mapping impractically slow”*

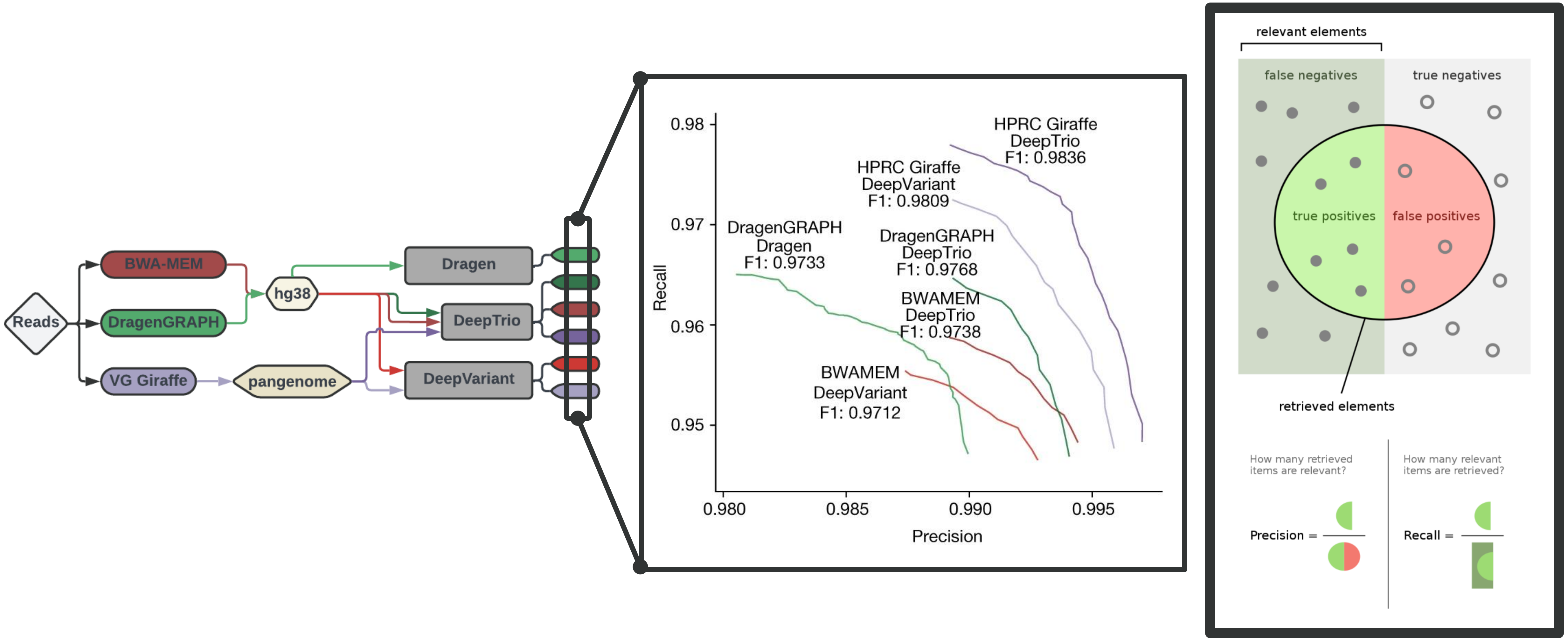
What can I use a genome graph to do?

- Calling variants – **VG Giraffe** + **DeepVariant**
- Genotyping structural variants – **Pangenie + Locityper**

- Analyze ATACseq ⟵ **VG Giraffe** + **GraphPeakCaller**
- Analyze RNAseq ⟵ **VG mpmap** + **rpvg**
- Analyze WGBS ⟵ **methylGrapher**

- Analyze HiC ⟵ *No tool exists*

Why use a genome graph?



Pangenomes

- Reduce reference biases.

- Variants that disrupt alignment are the most difficult to detect, are the most likely to impact function, and are the most likely to introduce errors in probing the biology around them

- Assembling a personalized genome and/or using a pangenome let’s you get passed reference limitations

Linear genome literature

Long-read human genome sequencing and its applications

Logsdon, G.A., Vollger, M.R. & Eichler, E.E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21,** 597–614 (2020). https://doi.org/10.1038/s41576-020-0236-x

Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome

Wenger, A.M., Peluso, P., Rowell, W.J. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37,** 1155–1162 (2019). https://doi.org/10.1038/s41587-019-0217-9

Long-read sequencing in deciphering human genetics to a greater depth

Midha, M.K., Wu, M. & Chiu, KP. Long-read sequencing in deciphering human genetics to a greater depth. *Hum Genet* **138,** 1201–1215 (2019). https://doi.org/10.1007/s00439-019-02064-y

Introduction to Genome Assembly

Bioinformatics Workbook (Online)

Andrew Severin - Author

https://bioinformaticsworkbook.org/dataAnalysis/GenomeAssembly/Intro_GenomeAssembly.html

Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing

Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13(12):1050-1054. doi:10.1038/nmeth.4035

Extended haplotype phasing of *de novo* genome assemblies with FALCON-Phase

Zev N. Kronenberg, Arang Rhie, Sergey Koren, Gregory T. Concepcion, Paul Peluso, Katherine M. Munson, Stefan Hiendleder, Olivier Fedrigo, Erich D. Jarvis, Adam M. Phillippy, Evan E. Eichler, John L. Williams, Tim P.L. Smith, Richard J. Hall, Shawn T. Sullivan, Sarah B. Kingan

bioRxiv 327064; doi: https://doi.org/10.1101/327064

GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations

Manchanda, N., Portwood, J.L., Woodhouse, M.R. *et al.* GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics* **21,** 193 (2020). https://doi.org/10.1186/s12864-020-6568-2

Genome graph literature

The Human Pangenome Project: a global resource to map genomic diversity

https://www.nature.com/articles/s41586-022-04601-8

A Draft Human Pangenome Reference

https://www.nature.com/articles/s41586-023-05896-x

Genome graphs detect human polymorphisms in active epigenomic state during influenza infection

https://www.sciencedirect.com/science/article/pii/S2666979X23000605?via%3Dihub

The design and construction of reference pangenome graphs with minigraph

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02168-z

Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus

https://www.biorxiv.org/content/10.1101/2022.10.06.511217v1

Building pangenome graphs

https://www.biorxiv.org/content/10.1101/2023.04.05.535718v1

Acknowledgements

Current lab members

Xiaoyun Xing
Noah Basri
Ju Heon Maeng
Ivy Chen
Holden Liang
Aparna Ananda
Heather Schmidt

Daofeng Li
Xiaoyu Zhuo
Kara Quid
Benpeng Miao
Jessica Harrison
Heather Lawson
Prashant Kuntala

Alan Du
Changxu Fan
Celine St. Pierre
Xuan Qu
Shihua Dong
More…

NIHR01HG007354,
R01HG007175,
U01CA200060,
U24ES026699,
U01HG009391,
U41HG010972,
U24HG012070
American Cancer Society
RSG-14-049-01-DMC
Emerson Collective Funds
Siteman Investigator
Program

Acknowledgements

Collaborative and Integrative Genomics (CIG) Group

The team:

Chad Tomlinson
Eddie Belter
Derek Albracht
John Garza
Juan Macias

THE GENOME INSTITUTE at Washington University

@JuanFMacias1

Questions?