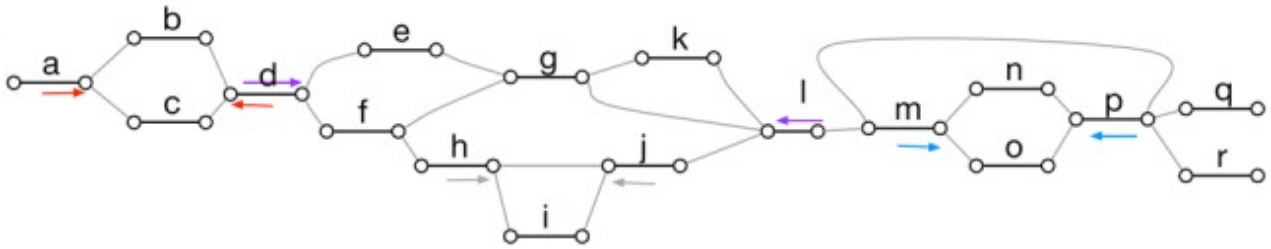
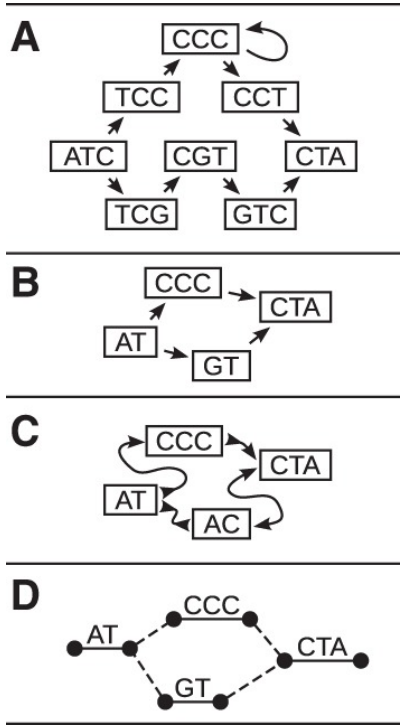
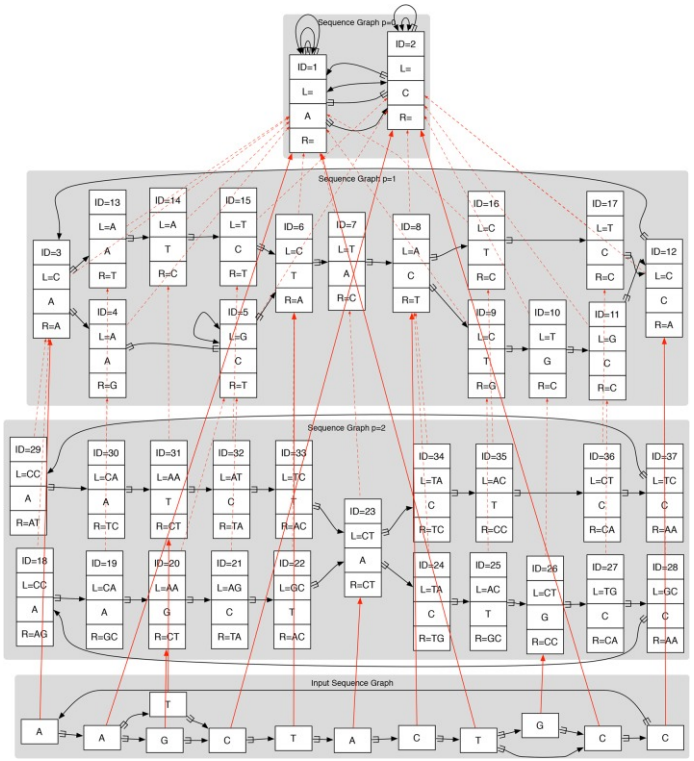
Inferring Genotypes from Skim Sequence using a Graph-Based Approach: The Practical Haplotype Graph

Peter Bradbury, Terry Casstevens, Dan Ilut, Lynn Johnson, Zachary Miller, Ramu Punna, Maria Cinta Romay, Edward Buckler

Genome Graphs Represent Diversity







Figures taken from a review by Paten et al. Genome Res. 2017 May; 27(5): 665–676. doi: 10.1101/gr.214155.116

Using a Pan-Genome

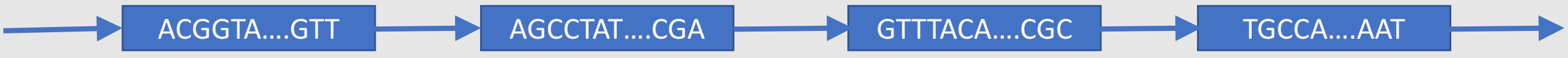
- Aligning to a pan-genome is better than aligning to a single reference
- Representing a pan-genome as a graph works
- There are a variety of approaches to graphical genomes
- Creating and using pan-genome graphs is complex

Finding a Practical Method of Using Multiple Genomes

The maize genome consists of

- **Conserved genes (and other elements)**
  - **~38,000 anchored intervals**
  - **~8-10% of the genome**
- **Non-conserved intergenic intervals (highly variable)**
- **Architecture similar across many species**

A chromosome is a sequence of haplotypes



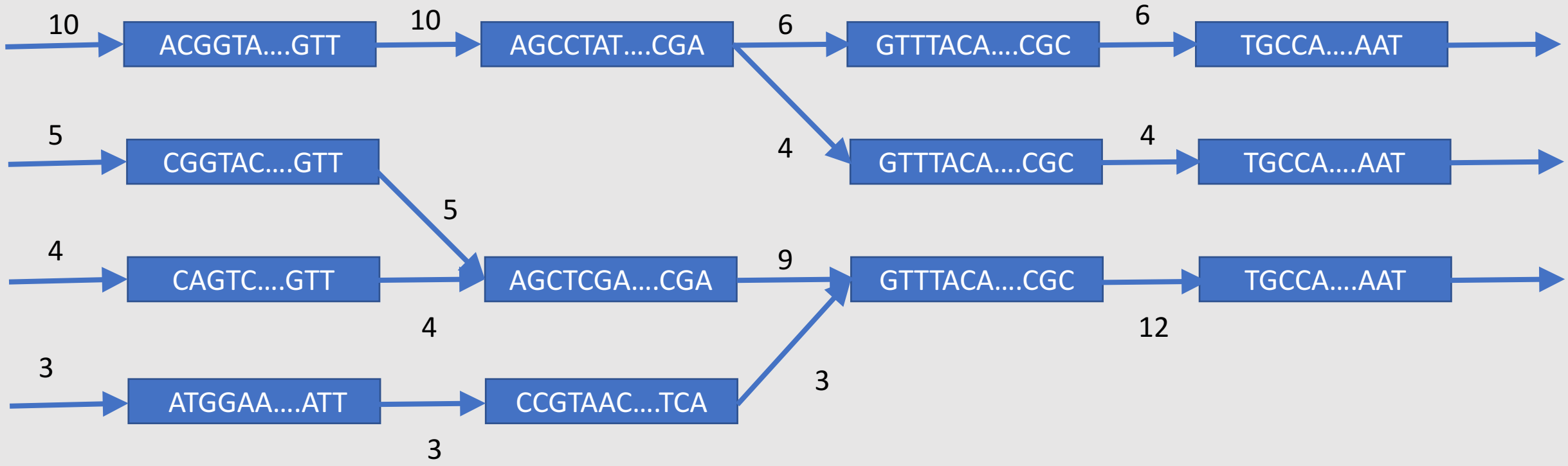
Node

A node is a haplotype

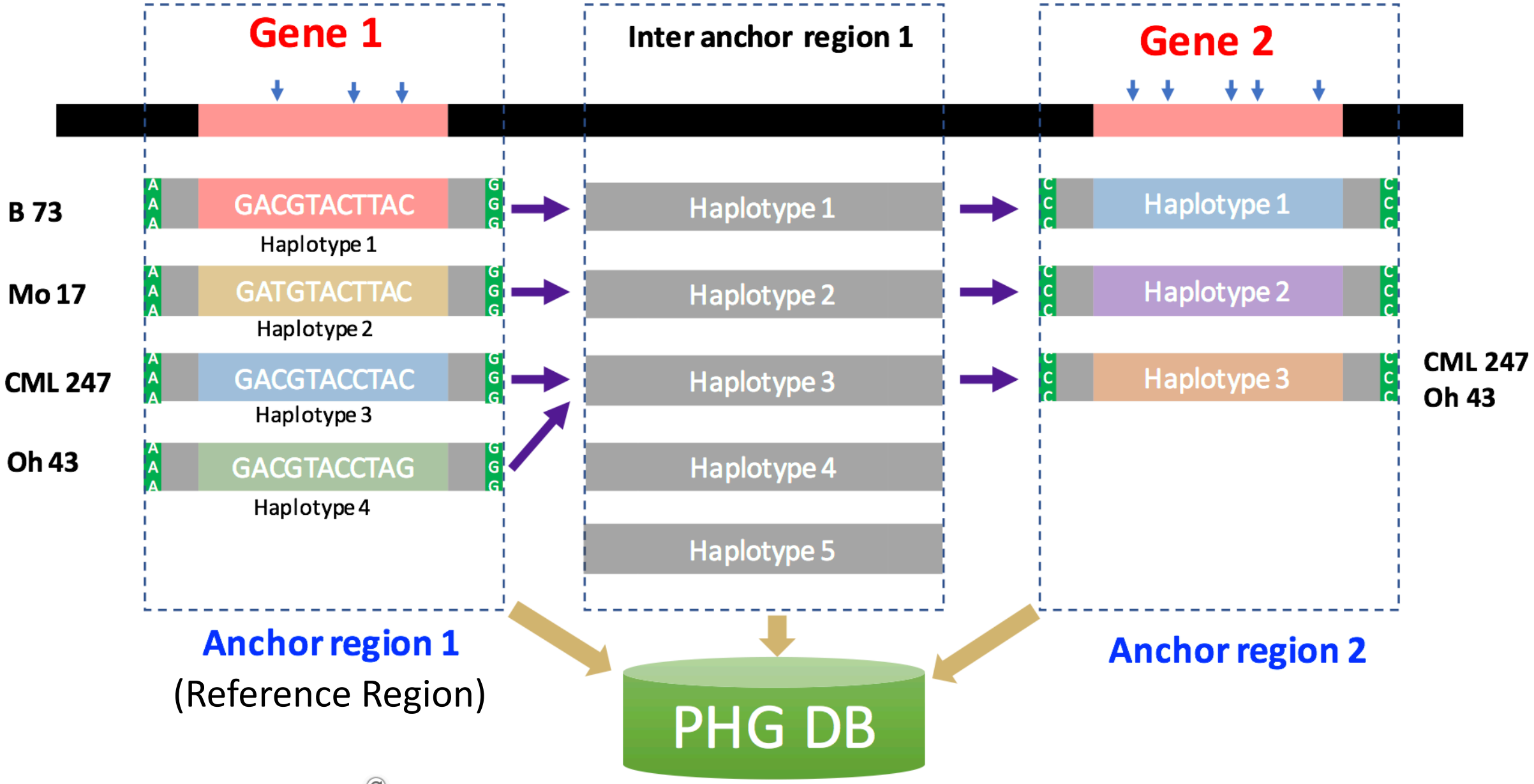
Edge

An edge is the connection between two nodes

Population of chromosomes

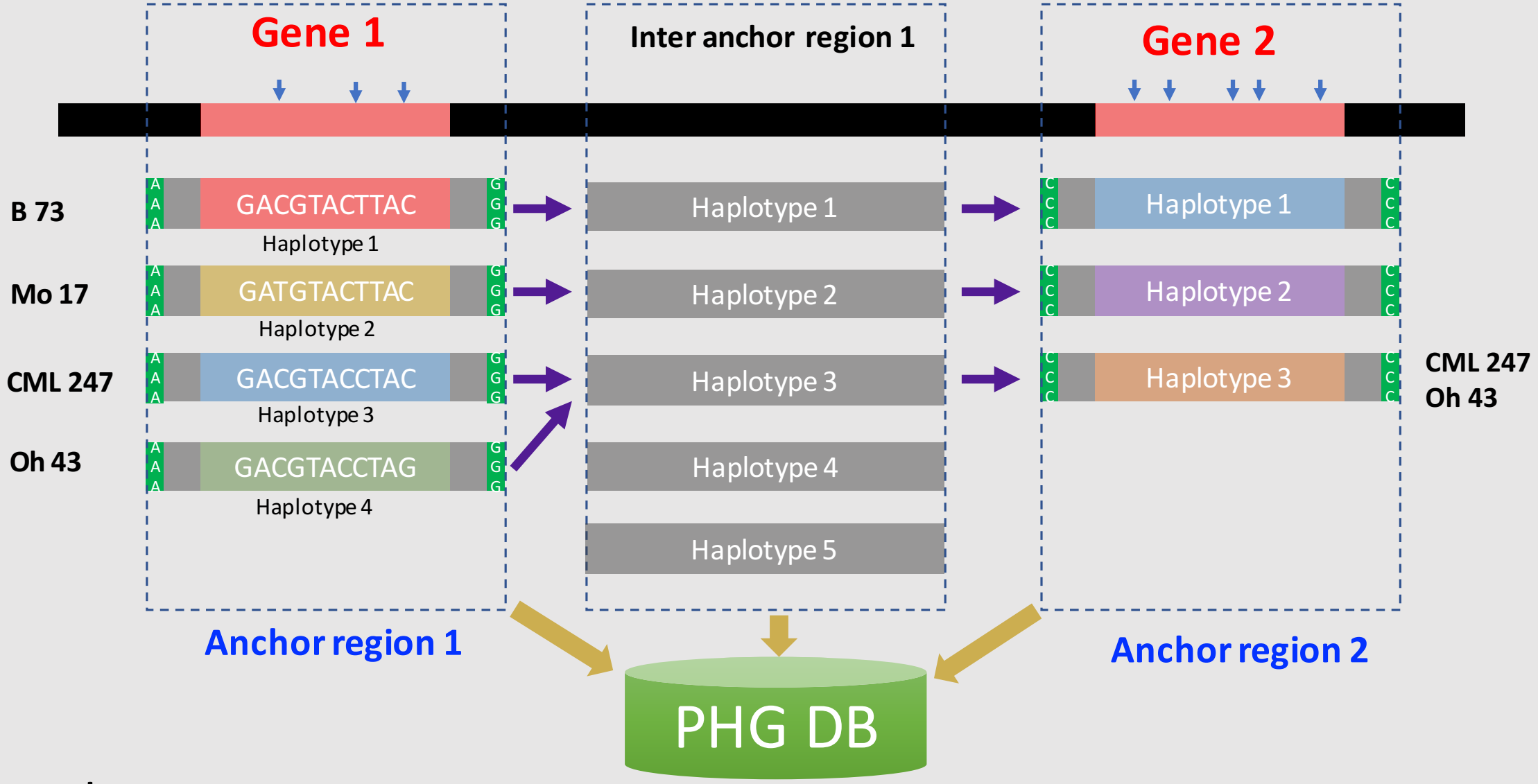


Edges are weighted by number of times observed in population


Gene 1
Inter anchor region 1
Gene 2
B 73
Mo 17
CML 247
Oh 43
GACGTACTTAC
Haplotype 1
GATGTACTTAC
Haplotype 2
GACGTACCTAC
Haplotype 3
GACGTACCTAG
Haplotype 4
Haplotype 1
Haplotype 2
Haplotype 3
Haplotype 4
Haplotype 5
Haplotype 1
Haplotype 2
Haplotype 3
CML 247
Oh 43
Anchor region 1
(Reference Region)
PHG DB
Anchor region 2

Create Consensus Haplotypes

- Cluster haplotypes within each reference range
- Reduce memory footprint
- 308 inbred line pangenome → 8 GB RAM
- Increase haplotype coverage → better quality



Key elements:

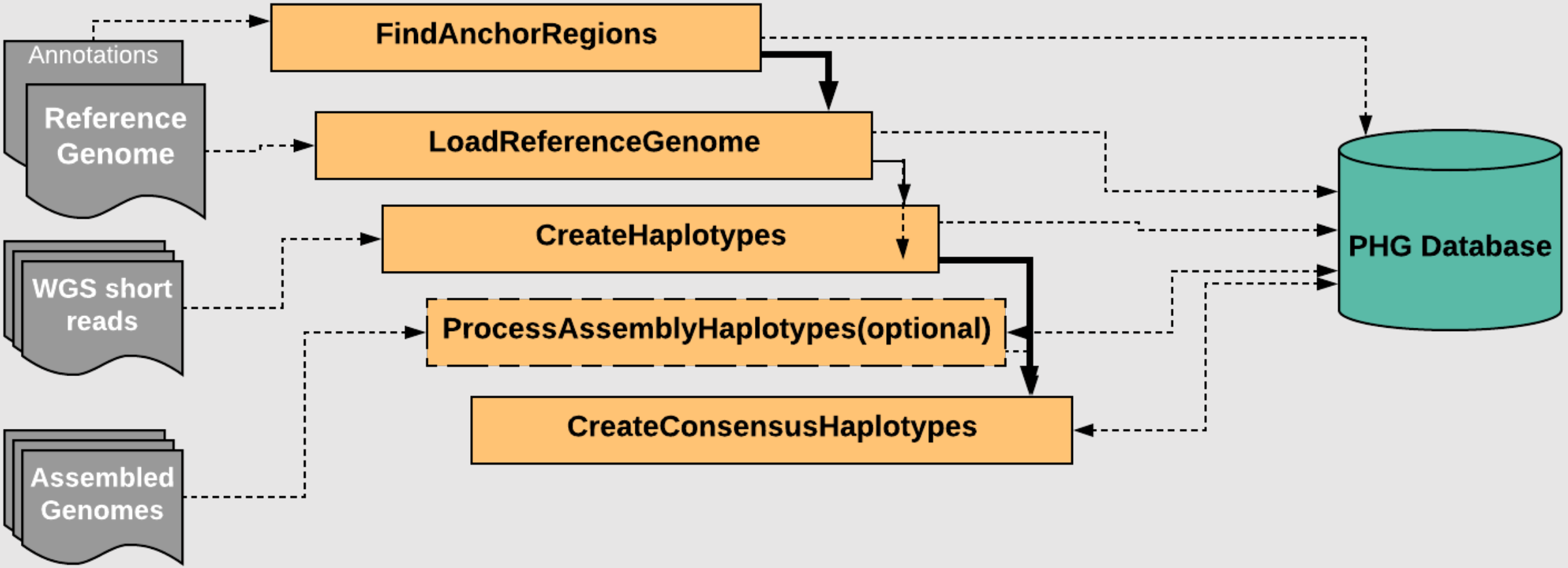
Coordinates of reference genome attached to anchor nodes

Anchor node have nearly 100% conserved start and ends

What is a Practical Haplotype Graph?

- A graph-based representation of multiple genomes
- A graph-based representation of the variation present in a population
- A computational framework
- A database
- Used to impute variants from skim sequence

Populating the PHG Database



The PHG is a computational framework

- Database
- Software
  - Populate database
  - Generate the graph in-memory from the database
  - Use the in-memory graph
- Pipeline that uses software from several sources
- Distributed as a Docker image

A Docker Image Captures The Computing Software Environment

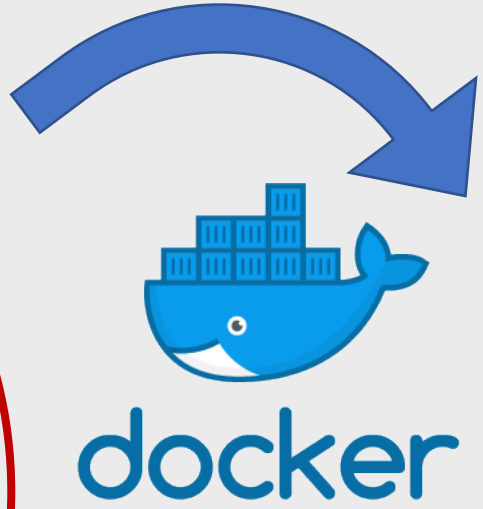
BWA

PHG

gatk

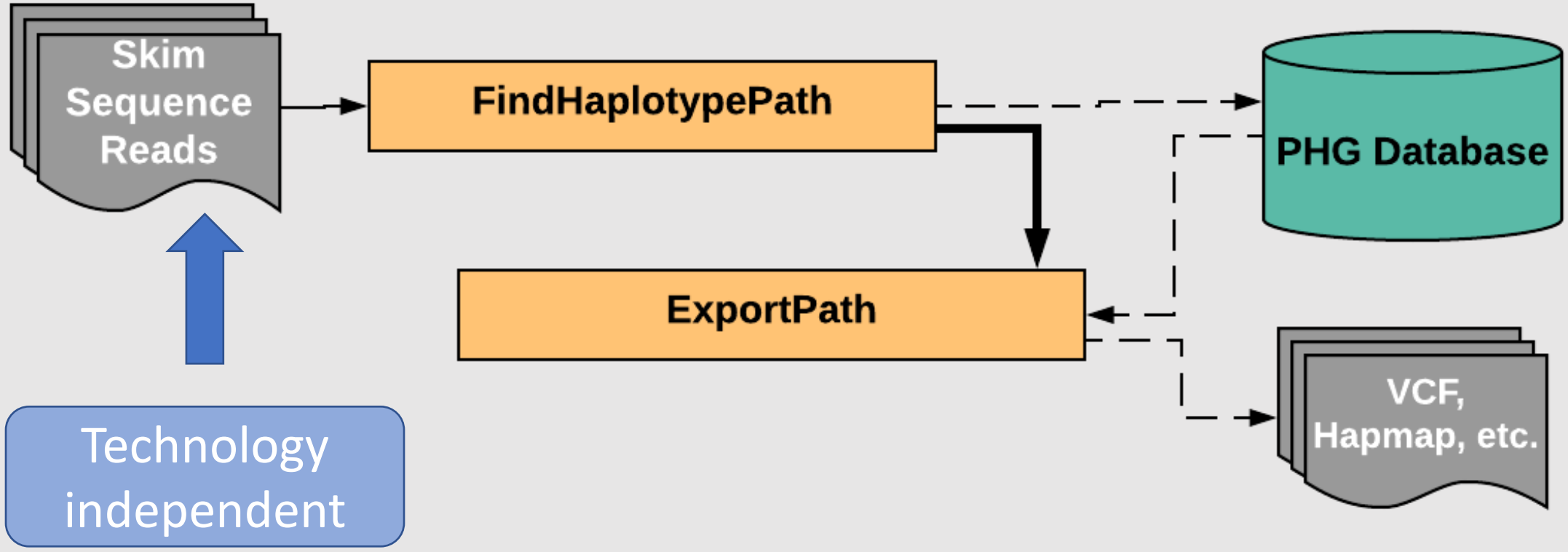
MiniMap2

SAMtools

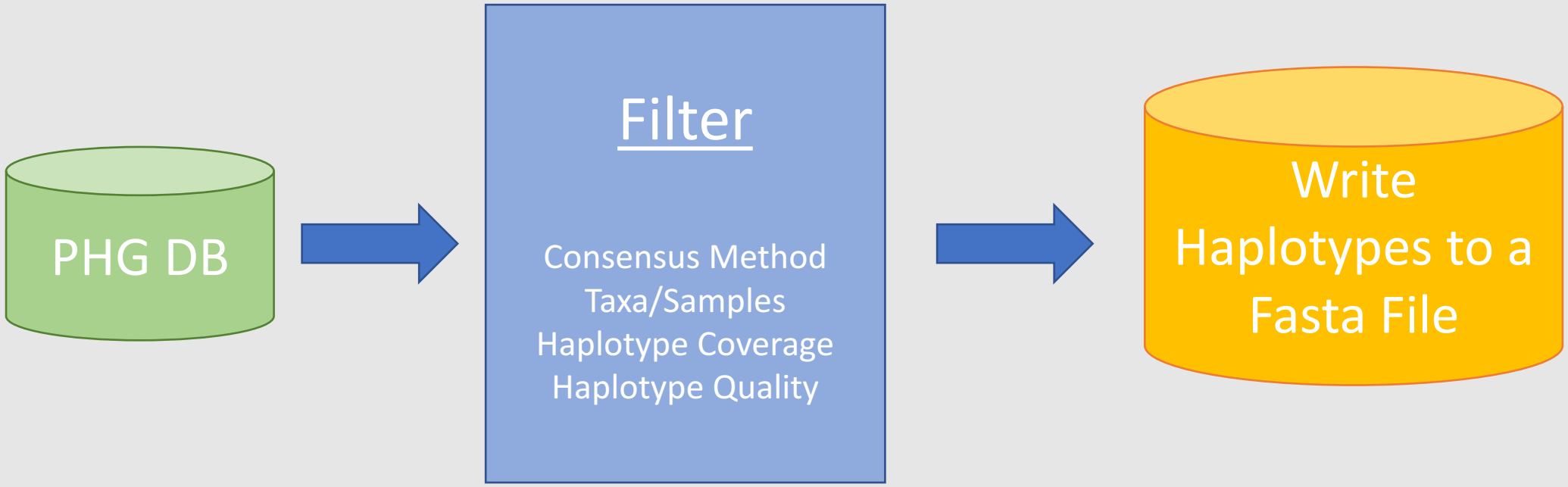


Any Supported OS

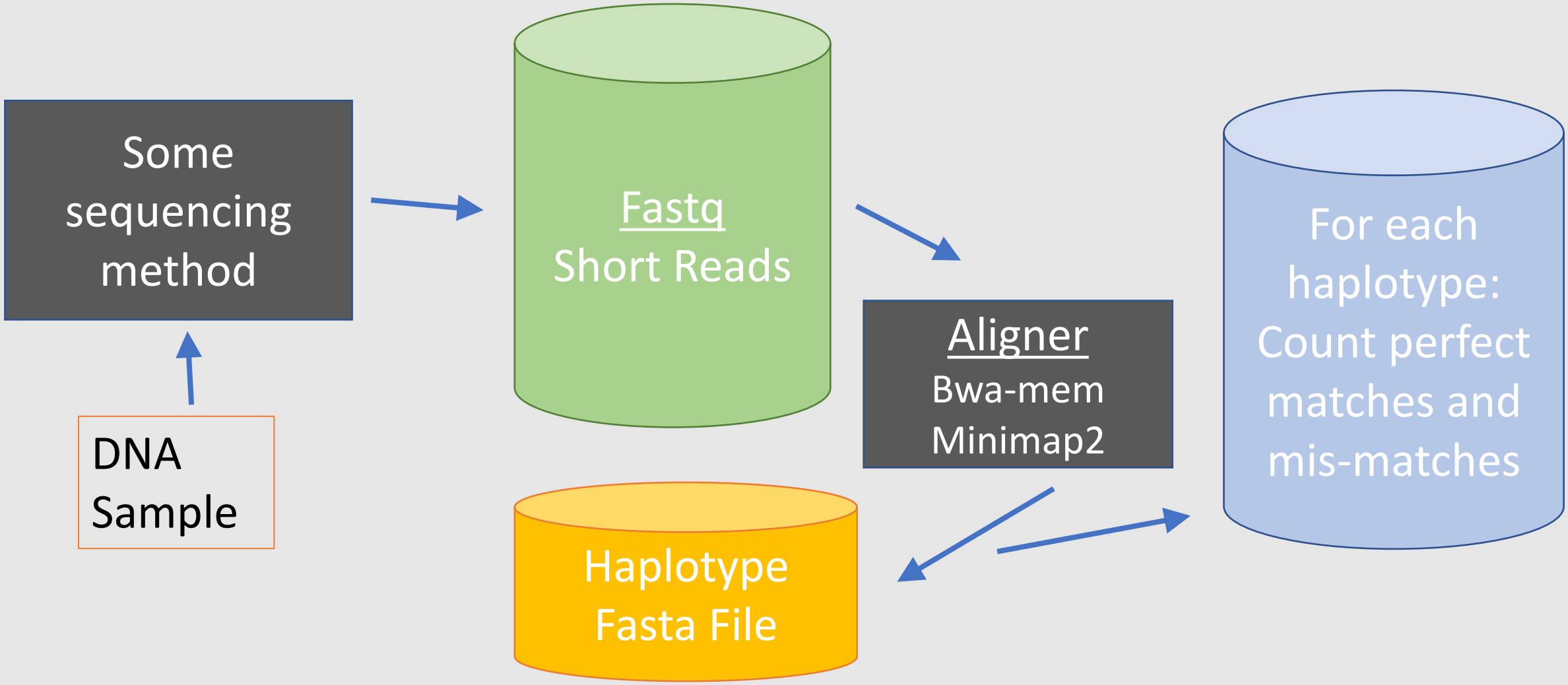
Genotyping Using a PHG



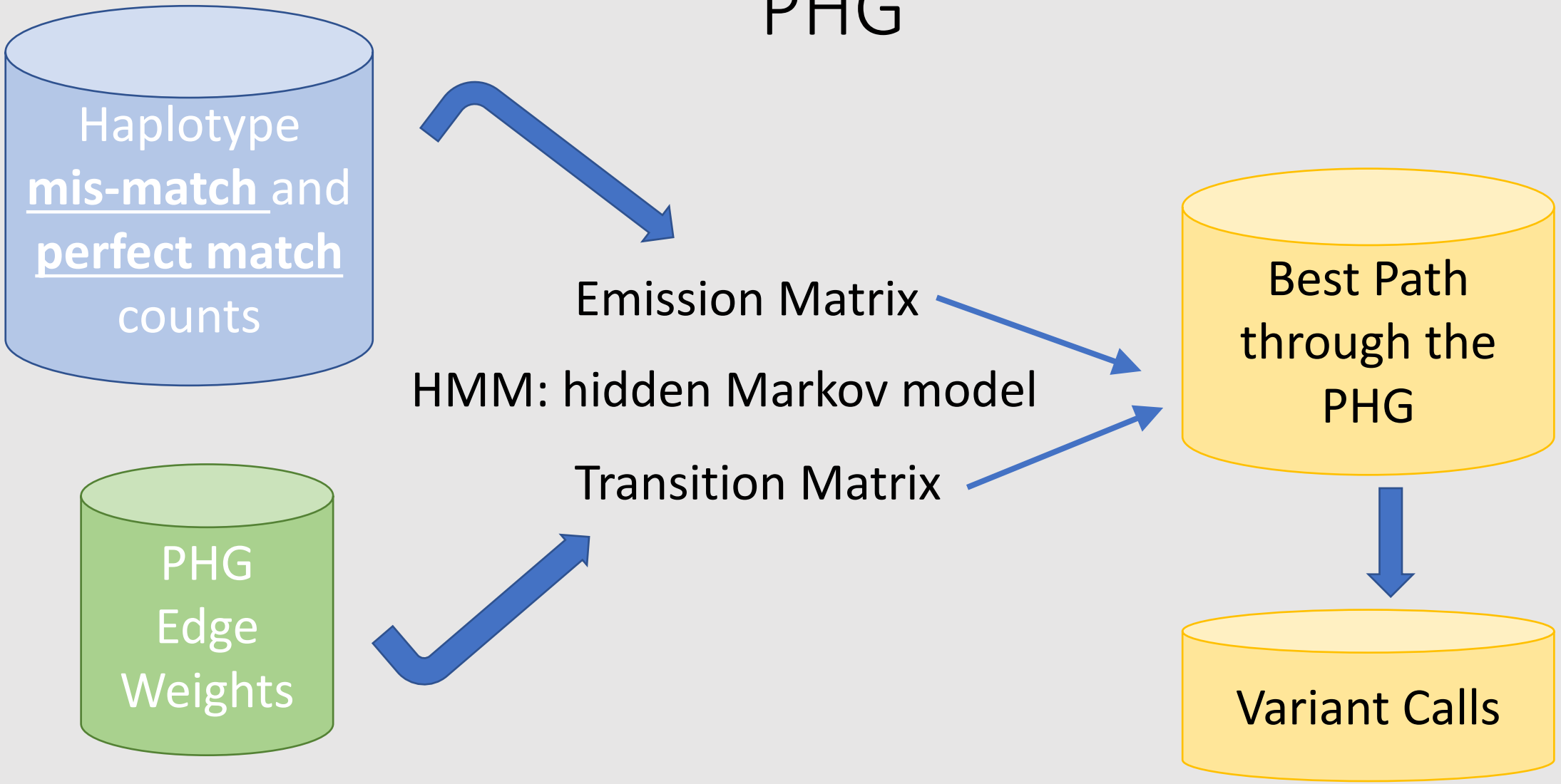
Filter and write PHG to Fasta



Align Short Sequence Reads To Haplotype Fastas



Use an HMM to find the best path through a PHG



Current status:

- Maize - 308 taxa
- Sorghum - 140 taxa (under development)
- Cassava - 348 taxa (under development)

- Tested using W22 GBS sequence
- Pathway: 85% of nodes called correctly
- Error rate calling SNPs – 2% (compared to Axiom array)

- GBS reads for 10K taxa (CIMMYT) – processed through PHG
    - 3.7 M SNPs in anchor regions
    - Used for genomic selection

Limitations

- Still under active development
- The current genotyping application targets breeding programs
  - Populations with a limited number of founders
- Testing to date has been done with inbred lines

Key Points

- PHG is a simplified pan-genome graph
- A single reference genome plus moderate coverage WGS is sufficient
- PHG software will be relatively easy to run
- Both haplotype finding and genotyping can combine sequence from a variety of technologies

Acknowledgements

Ed Buckler Lab

Lynn Johnson
Terry Casstevens
Zack Miller
Ramu Punna
Cinta Romay
Dan Ilut
Sara Miller

Michael Gore Lab

Dan Ilut

For more information

- PHG Wiki: documentation, and source code at
- https://bitbucket.org/bucklerlab/practicalhaplotypegraph
- Docker image available in March 2018

Poster 114
Lynn Johnson