What is the Practical Haplotype Graph?

Why do we need it?

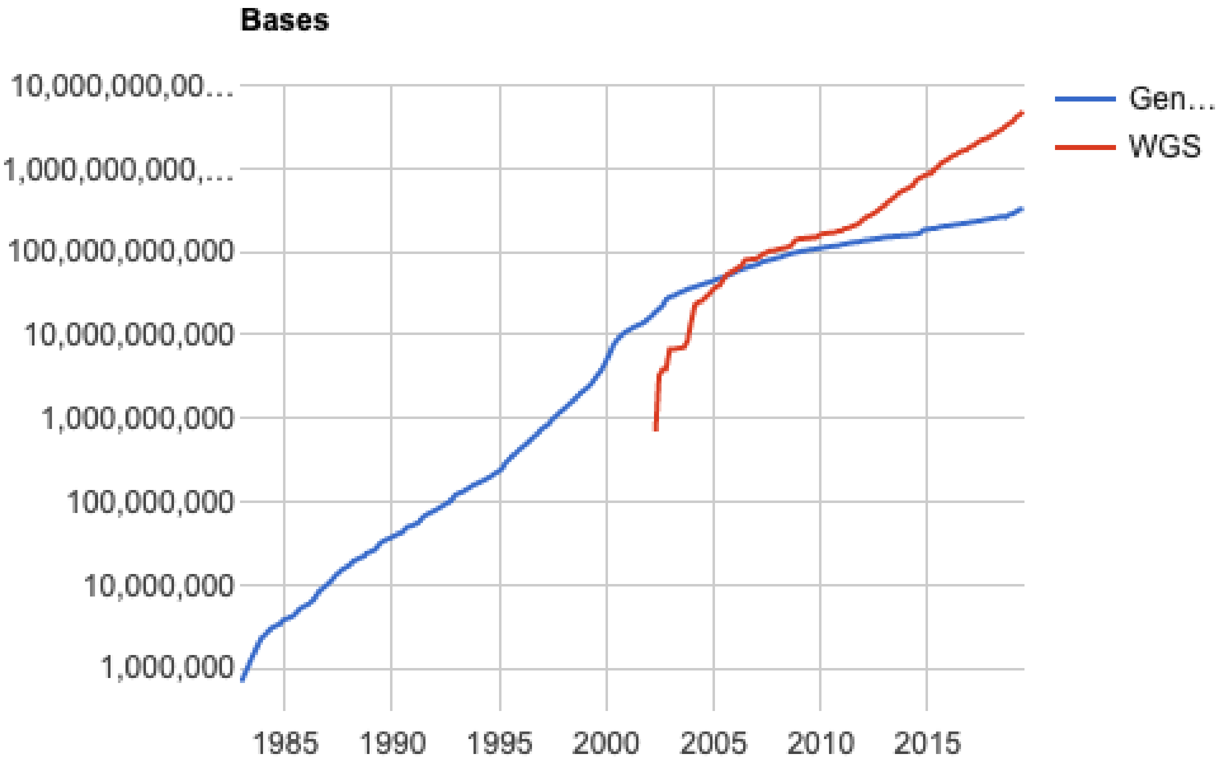
Sarah Jensen, Ed Buckler, Jean-Luc Jannink

Cornell University

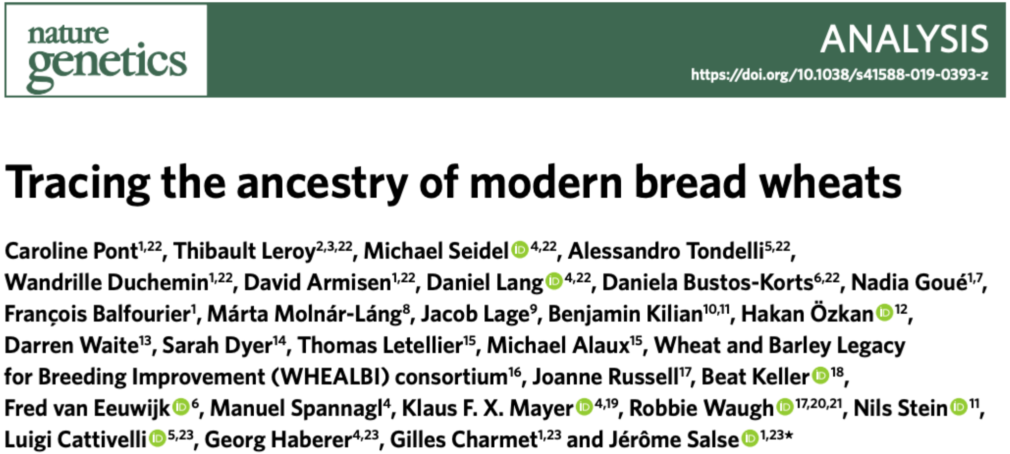
WheatCAP PHG training, July 2019

Since 1982, the number of bases in GenBank has ***doubled*** every 18 months

The numbers are literally off the charts



Even in wheat, many individuals are being sequenced



nature genetics

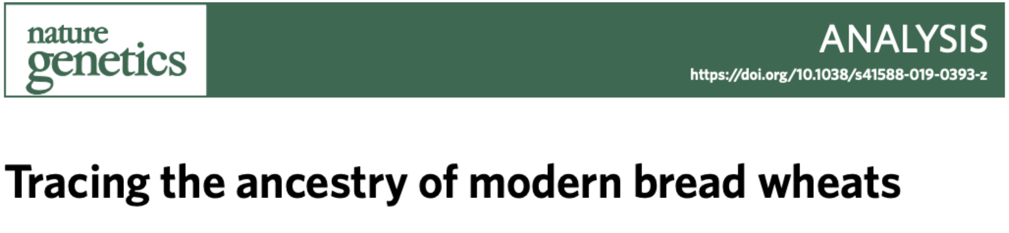
ANALYSIS

https://doi.org/10.1038/s41588-019-0393-z

Tracing the ancestry of modern bread wheats

Caroline Pont[1,22], Thibault Leroy[2,3,22], Michael Seidel[4,22], Alessandro Tondelli[5,22], Wandrille Duchemin[1,22], David Armisen[1,22], Daniel Lang[4,22], Daniela Bustos-Korts[6,22], Nadia Goué[1,7], François Balfourier[1], Márta Molnár-Láng[8], Jacob Lage[9], Benjamin Kilian[10,11], Hakan Özkan[12], Darren Waite[13], Sarah Dyer[14], Thomas Letellier[15], Michael Alaux[15], Wheat and Barley Legacy for Breeding Improvement (WHEALBI) consortium[16], Joanne Russell[17], Beat Keller[18], Fred van Eeuwijk[6], Manuel Spannagl[4], Klaus F. X. Mayer[4,19], Robbie Waugh[17,20,21], Nils Stein[11], Luigi Cattivelli[5,23], Georg Haberer[4,23], Gilles Charmet[1,23] and Jérôme Salse[1,23*]

Even in wheat, many individuals are being sequenced



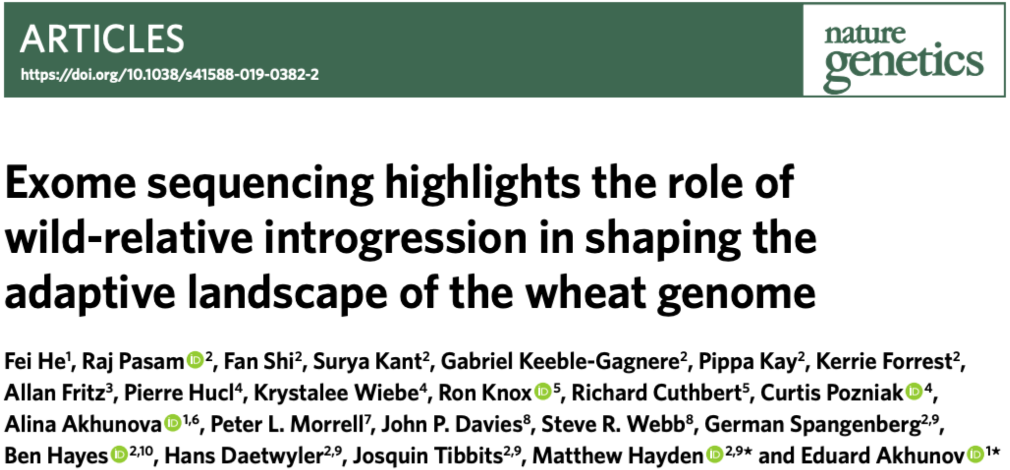
nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-019-0393-z

Tracing the ancestry of modern bread wheats

Caroline Pont[1,22], Thibault Leroy[2,3,22], Michael Seidel[4,22], Alessandr
Wandrille Duchemin[1,22], David Armisen[1,22], Daniel Lang[4,22], Daniel
François Balfourier[1], Márta Molnár-Láng[8], Jacob Lage[9], Benjamin Kil
Darren Waite[13], Sarah Dyer[14], Thomas Letellier[15], Michael Alaux[15], W
for Breeding Improvement (WHEALBI) consortium[16], Joanne Russel
Fred van Eeuwijk[6], Manuel Spannagl[4], Klaus F. X. Mayer[4,19], Robl
Luigi Cattivelli[5,23], Georg Haberer[4,23], Gilles Charmet[1,23] and Jérôm



ARTICLES

https://doi.org/10.1038/s41588-019-0382-2

nature genetics

Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome

Fei He[1], Raj Pasam[2], Fan Shi[2], Surya Kant[2], Gabriel Keeble-Gagnere[2], Pippa Kay[2], Kerrie Forrest[2], Allan Fritz[3], Pierre Hucl[4], Krystalee Wiebe[4], Ron Knox[5], Richard Cuthbert[5], Curtis Pozniak[4], Alina Akhunova[1,6], Peter L. Morrell[7], John P. Davies[8], Steve R. Webb[8], German Spangenberg[2,9], Ben Hayes[2,10], Hans Daetwyler[2,9], Josquin Tibbits[2,9], Matthew Hayden[2,9]* and Eduard Akhunov[1]*

Even in wheat, many individuals are being sequenced

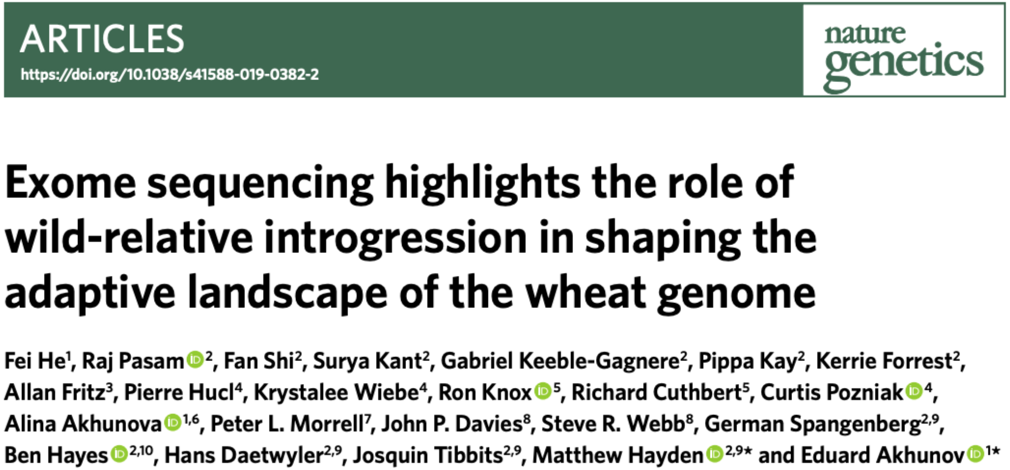
nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-019-0393-z

Tracing the ancestry of modern bread wheats

Caroline Pont[1,22], Thibault Leroy[2,3,22], Michael Seidel[4,22], Alessandr
Wandrille Duchemin[1,22], David Armisen[1,22], Daniel Lang[4,22], Daniela
François Balfourier[1], Márta Molnár-Láng[8], Jacob Lage[9], Benjamin Kil
Darren Waite[13], Sarah Dyer[14], Thomas Letellier[15], Michael Alaux[15], W
for Breeding Improvement (WHEALBI) consortium[16], Joanne Russel
Fred van Eeuwijk[6], Manuel Spannagl[4], Klaus F. X. Mayer[4,19], Robl
Luigi Cattivelli[5,23], Georg Haberer[4,23], Gilles Charmet[1,23] and Jérôm



ARTICLES

https://doi.org/10.1038/s41588-019-0382-2

nature genetics

Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome

Fei He[1], Raj Pasam[2], Fan Shi[2], Surya Kant[2], Gabriel Keeble-Gagnere[2], Pippa Kay[2], Kerrie Forrest[2], Allan Fritz[3], Pierre Hucl[4], Krystalee Wiebe[4], Ron Knox[5], Richard Cuthbert[5], Curtis Pozniak[4], Alina Akhunova[1,6], Peter L. Morrell[7], John P. Davies[8], Steve R. Webb[8], German Spangenberg[2,9], Ben Hayes[2,10], Hans Daetwyler[2,9], Josquin Tibbits[2,9], Matthew Hayden[2,9]* and Eduard Akhunov[1]*

Fruitful use of this data depends on summarizing it effectively

The PHG is

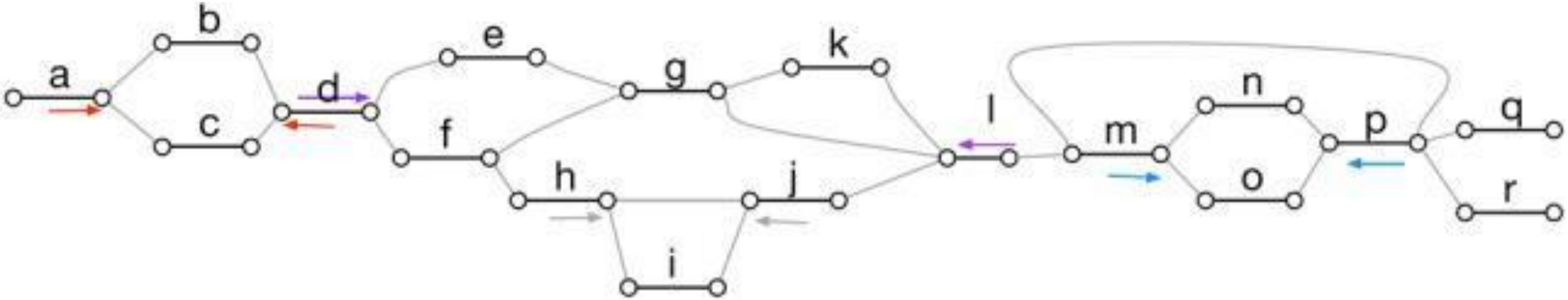
- A proposal for how to represent pan-genomes
- Software to do so
- Implementation of a primary use case: imputation of whole genome sequence from skim sequence

Outline

- The proposal
  - It’s rationale from a structural genomics / population genetics perspective
- Outline of the approach to implement the proposal
- Presentation of the use case imputation from skim sequence

The pan-genome captures genomic variants across individuals in a species

- Haplotype graphs represent diversity

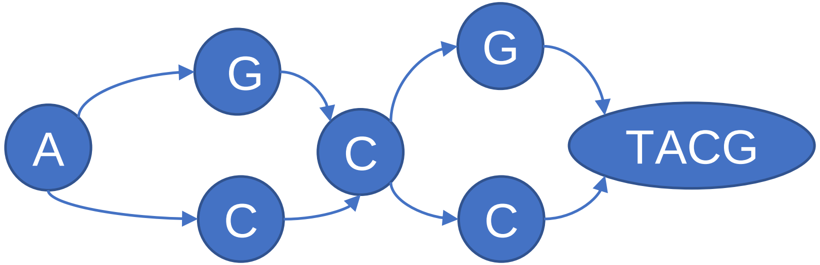


Review: Paten et al. *Genome Res*. 2017, May; **27(5)**: 665-676.

The pan-genome can be accurately represented as a graph

```
B73      AGCGT  ACGAGT----CATGA  CGTAA
Mo17     ACCGT  ACGNNTAAAACATGA  CGCAA
Oh43     AGCCT  ACGAGTAA--CANNA  CGCAA
```

- We have lots of ambiguity
- Intergenic retro-regions can be crazy hard
- Alignment tools are not graph aware

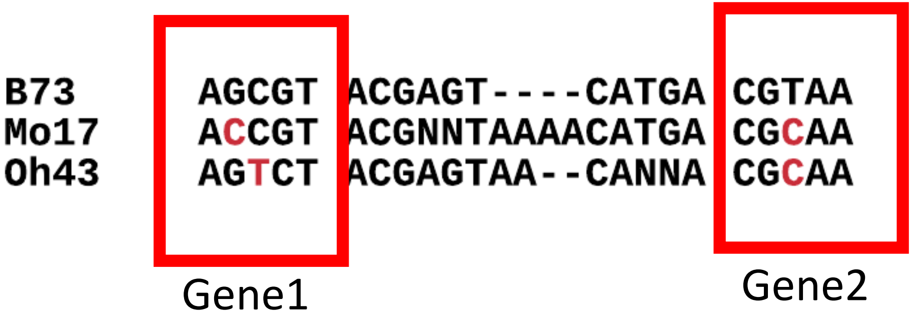


ACK! - NNs

Etc.

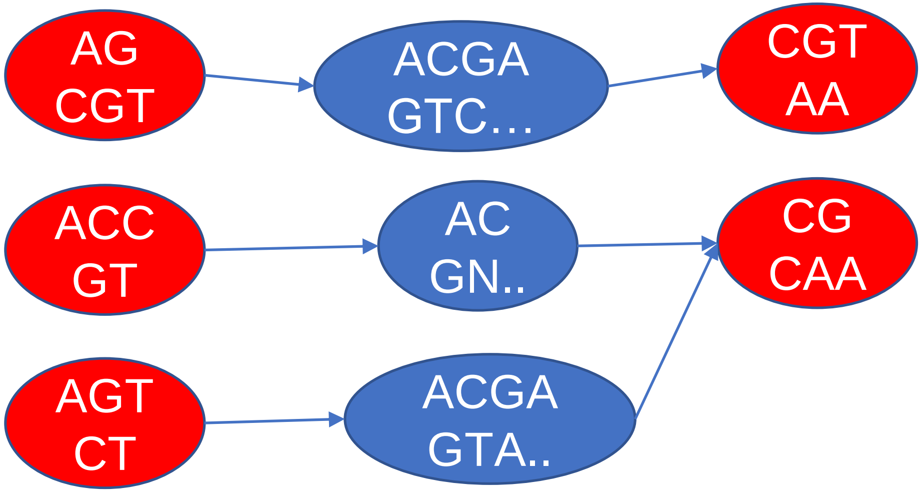
Make this practical

- Biology produces genomes with a consistent pattern
  - Conserved genes (and flanking elements)
  - Non-conserved intergenic regions with tremendous variation



```
B73     AGCGT ACGAGT----CATGA CGTAA
Mo17    ACCGT ACGNNTAAAACATGA CGCAA
Oh43    AGTCT ACGAGTAA--CANNA CGCAA
```

This pattern differentiates ranges



Key elements:

- Path graph
- Anchor and non-anchor ranges
- Haplotypes identified in each range

```
B73     AGCGT  ACGAGT----CATGA  CGTAA
Mo17    ACCGT  ACGNNTAAAACATGA  CGCAA
Oh43    AGTCT  ACGAGTAA--CANNA  CGCAA
```

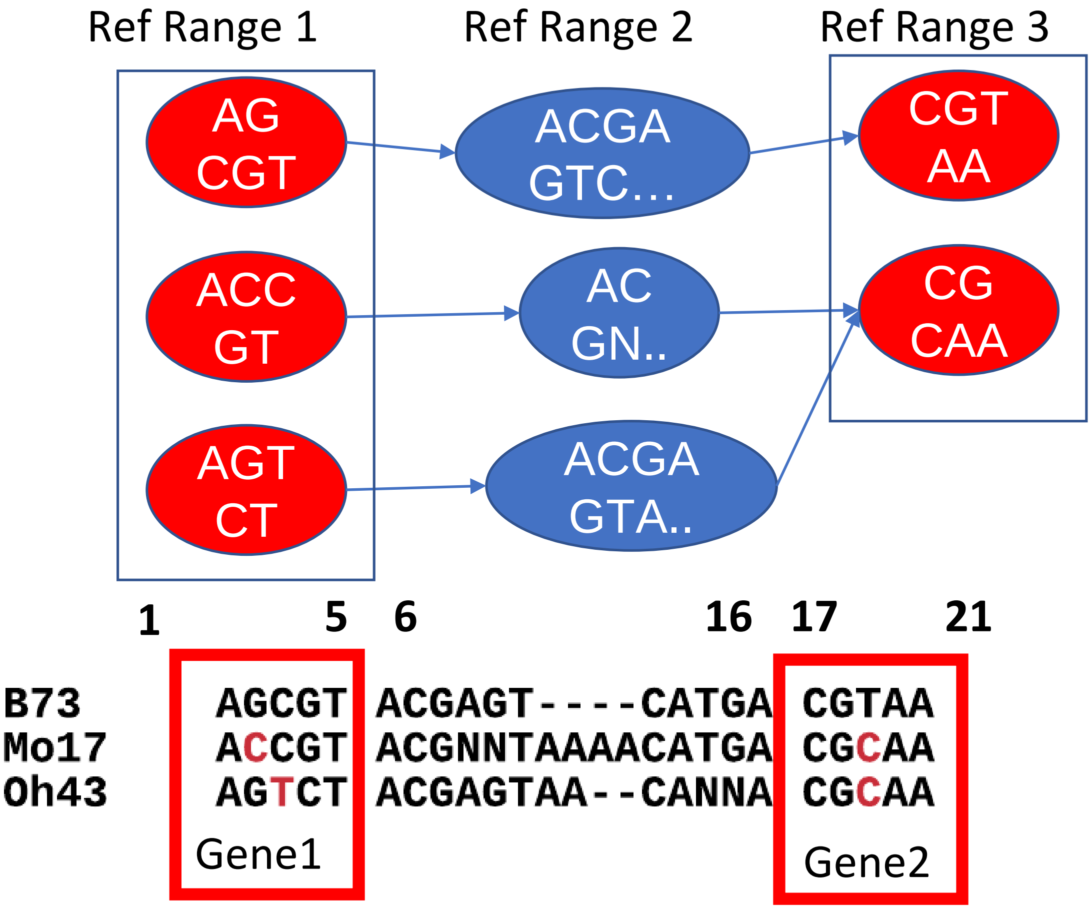
Gene1

Gene2

Anchor vs. non-anchor reference ranges

- Often, this will equate to ***genic*** vs. ***intergenic*** ranges, as annotated in the reference genome sequence
- What's relevant:
    a. ***essential*** (almost always present) vs. ***unessential*** (might be missing in some individuals)
    b. ***easily aligned*** (no repeat motifs) vs. ***not easily aligned*** (repeats, indels)
- Non-anchor regions may often contain genes
- The software doesn't care: figure out what works

Tie ranges to reference sequence



Ref Range 1

Ref Range 2

Ref Range 3

AG CGT

ACC GT

AGT CT

ACGA GTC…

AC GN..

ACGA GTA..

CGT AA

CG CAA

```
        1         5   6          16   17     21
B73      AGCGT  ACGAGT----CATGA   CGTAA
Mo17     ACCGT  ACGNNTAAAACATGA   CGCAA
Oh43     AGTCT  ACGAGTAA--CANNA   CGCAA
         Gene1                    Gene2
```

Key elements:

- Path graph
- Anchor and non-anchor ranges
- Haplotypes identified in each range
- Range coordinates tied to the reference genome

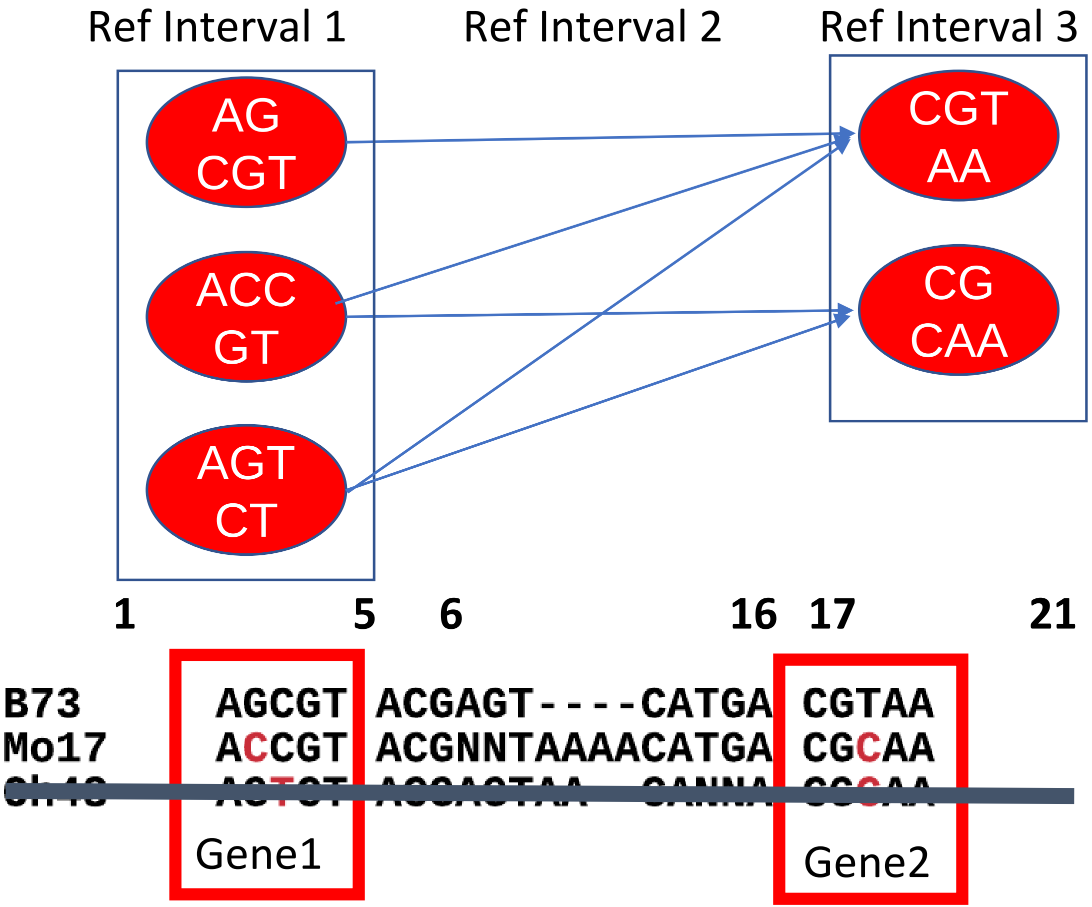
Nomenclature has varied over time

Reference Range = Reference Interval = Genome Interval

Anchor = Genic Interval

You might find these & more in documentation

What about a practical graph?

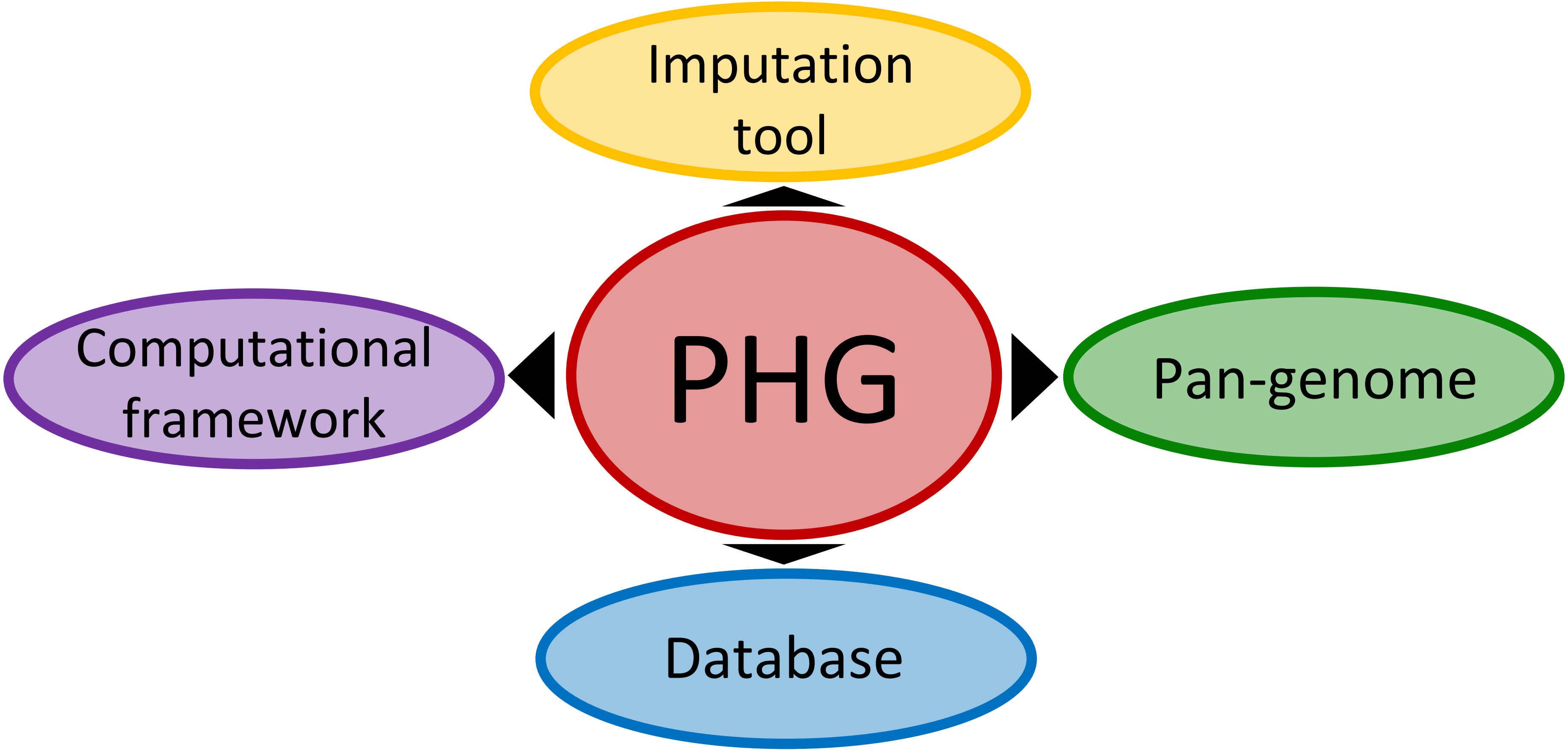

Ref Interval 1
Ref Interval 2
Ref Interval 3
AG CGT
ACC GT
AGT CT
CGT AA
CG CAA
1
5
6
16
17
21
B73 AGCGT ACGAGT----CATGA CGTAA
Mo17 ACCGT ACGNNTAAAACATGA CGCAA
~~Ch46 ACTCT ACCAGTAA--CANNA CGCAA~~
Gene1
Gene2

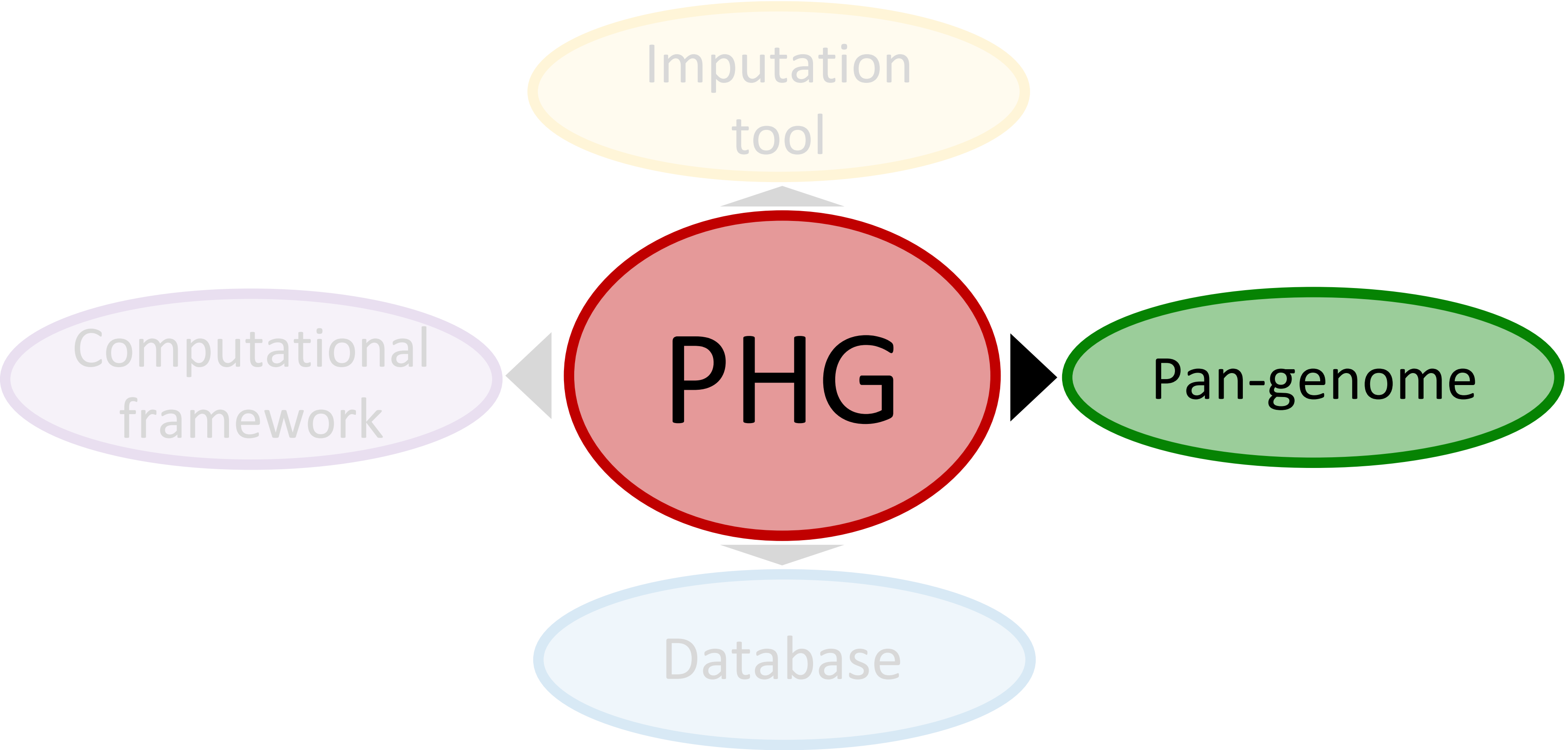
Key elements:

- Path graph
- Anchor and non-anchor ranges
- Haplotypes identified in each range
- Range coordinates tied to the reference genome
- Transition probabilities calculated between anchor haplotypes
- Probabilities specified to the population analyzed

Why is this practical?

- By definition, the community agrees on the reference genome as a coordinate system
- Works around the difficult regions of the genome
- Haplotype identification leads to compressed data
- Cheap short reads align well to the anchors
- Uses off-the-self bioinformatics (e.g., GATK)
- Can be used by both breeding and genomics communities


Imputation tool
Computational framework
PHG
Pan-genome
Database



Imputation tool

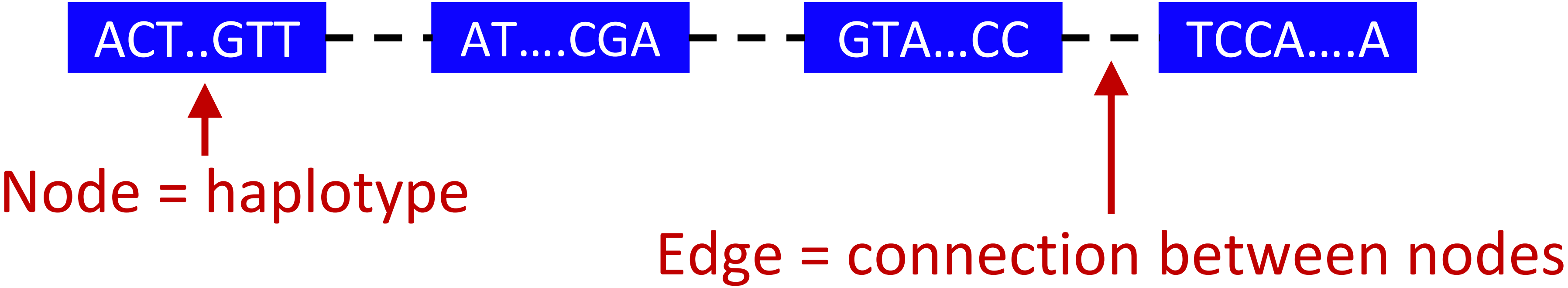
Computational framework

PHG

Pan-genome

Database

A chromosome is a sequence of haplotypes with conserved and non-conserved elements



ACT..GTT

AT….CGA

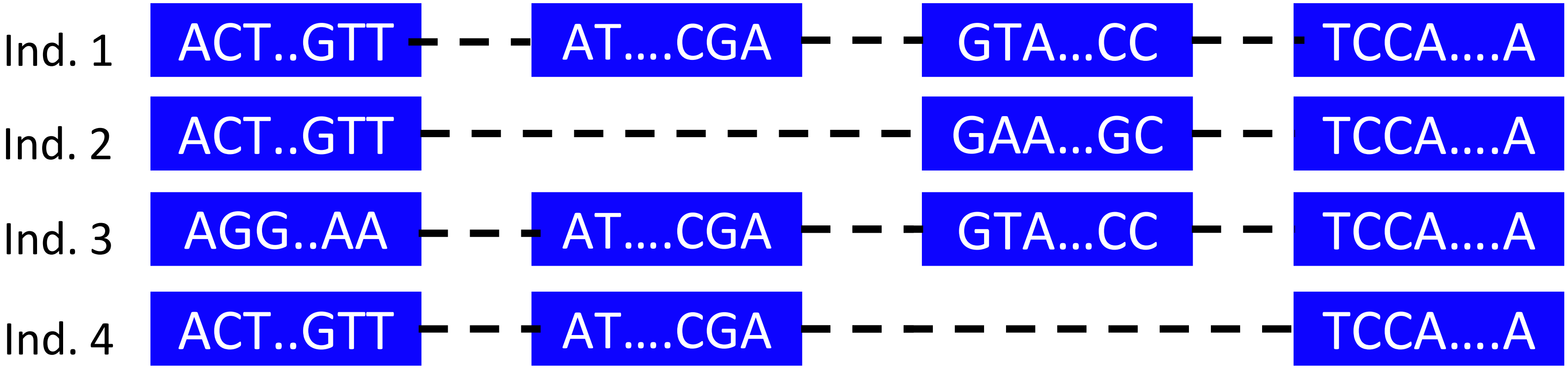
GTA…CC

TCCA….A

Node = haplotype

Edge = connection between nodes

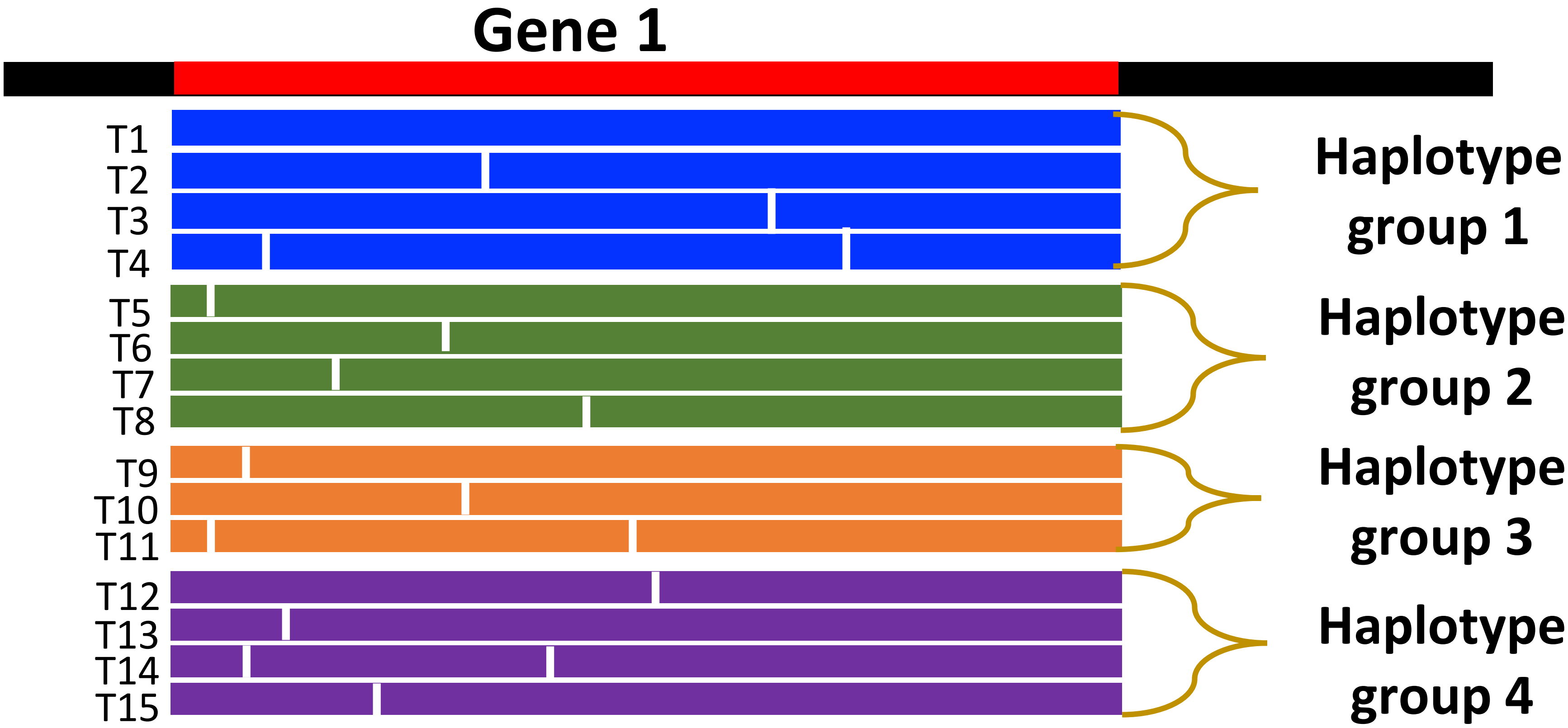
A population of chromosomes provides the basis for haplotype groups



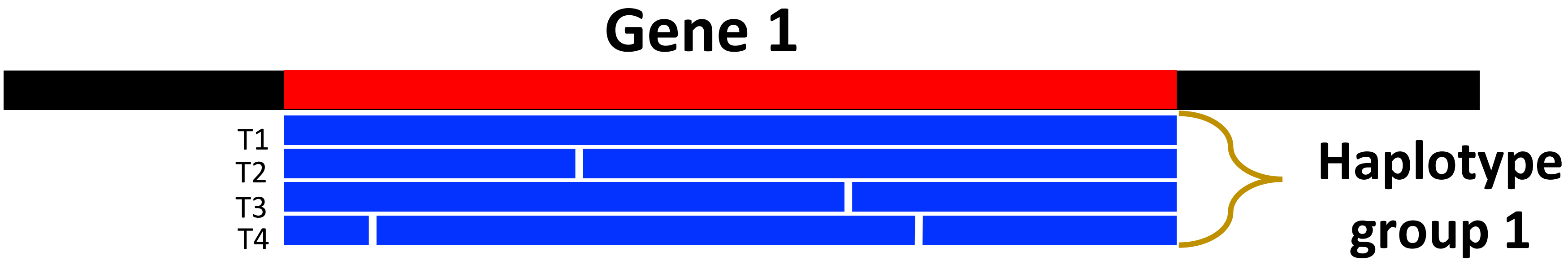
Ind. 1	ACT..GTT	AT....CGA	GTA...CC	TCCA....A
Ind. 2	ACT..GTT		GAA...GC	TCCA....A
Ind. 3	AGG..AA	AT....CGA	GTA...CC	TCCA....A
Ind. 4	ACT..GTT	AT....CGA		TCCA....A

- Cluster haplotypes at each anchor region
- Reduce memory footprint
- Increase haplotype coverage for better quality

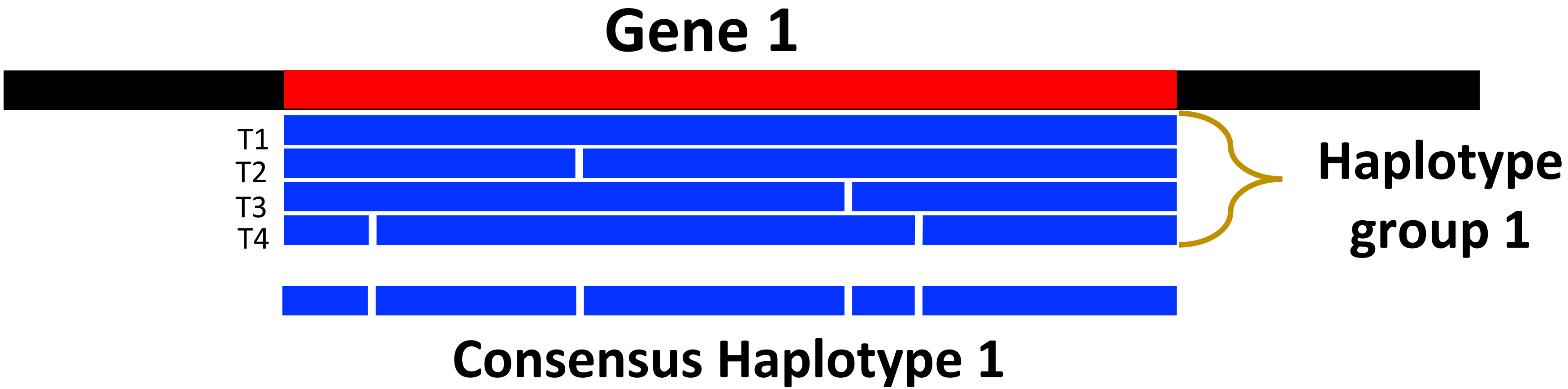
Haplotypes at a single gene in the PHG



Haplotypes at a single gene in the PHG

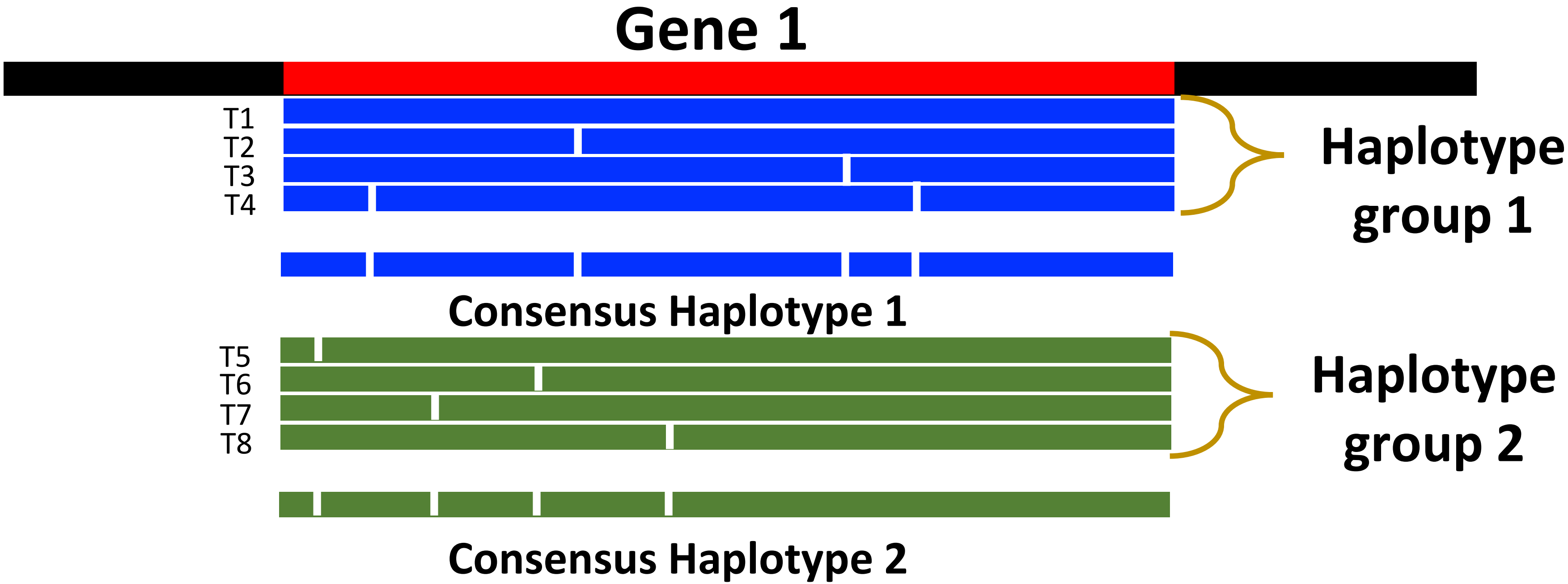


Haplotypes at a single gene in the PHG



All variant sites are maintained within the consensus sequence

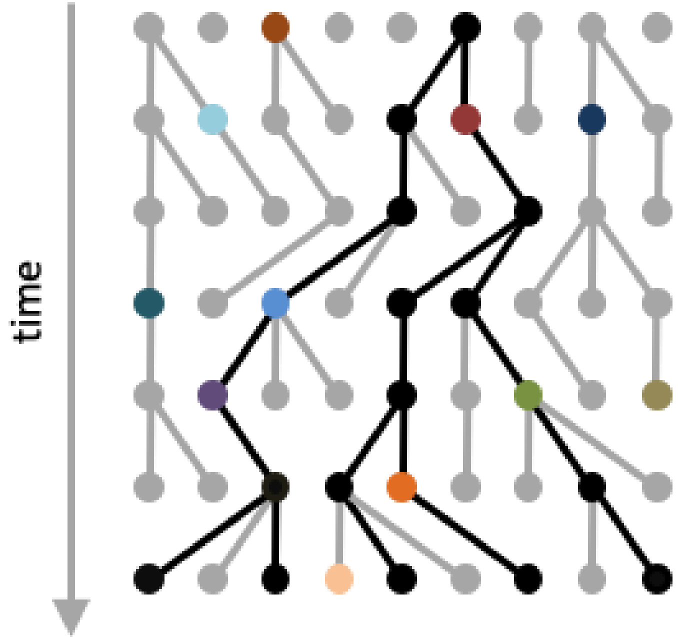
Haplotypes at a single gene in the PHG



All variant sites are maintained within the consensus sequence

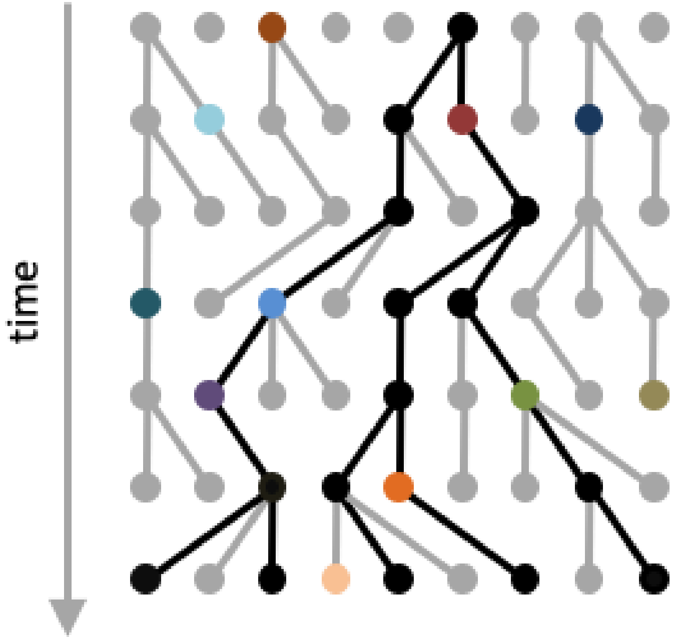
Population genetic detour: haplotype clustering is a good idea

Genealogy of a sample

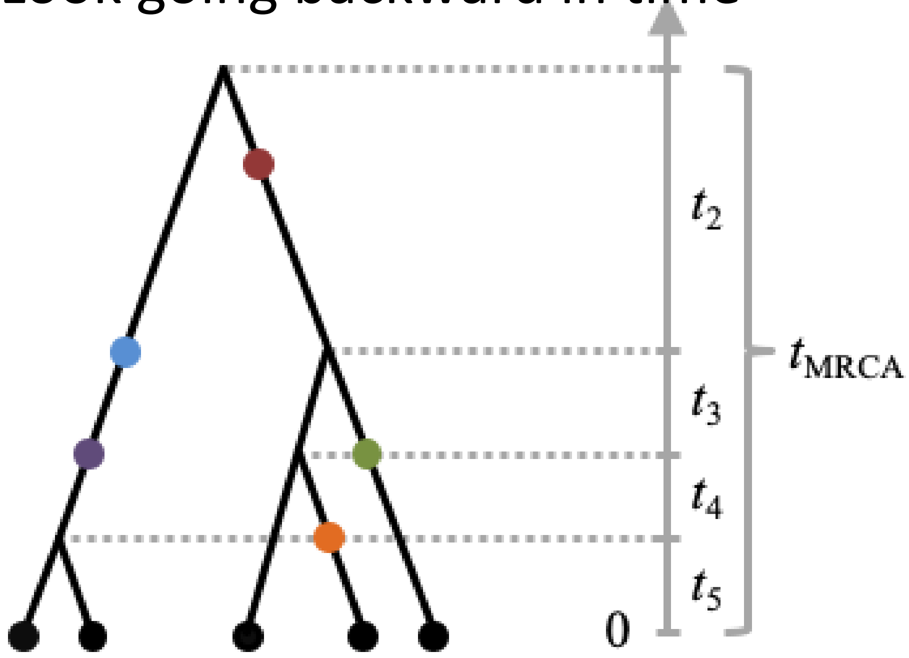


Population genetic detour: haplotype clustering is a good idea

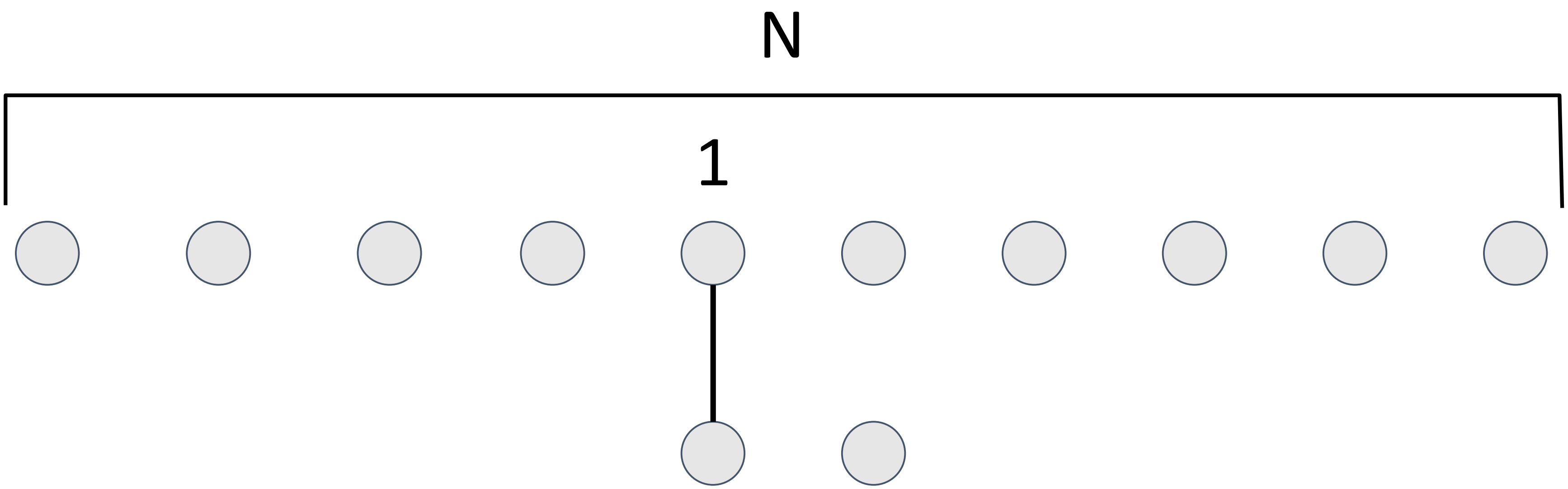
Genealogy of a sample



Look going backward in time



Probability that two lineages will coalesce



$\lambda_{c,2} = 1 / N$

$\lambda_{c,2} = 1 / 2N$

Expected *time* for two lineages to coalesce

$$\mathrm{E}(t_{\mathrm{c},2}) = 1 / \lambda_{\mathrm{c},2} = 2\mathrm{N}$$

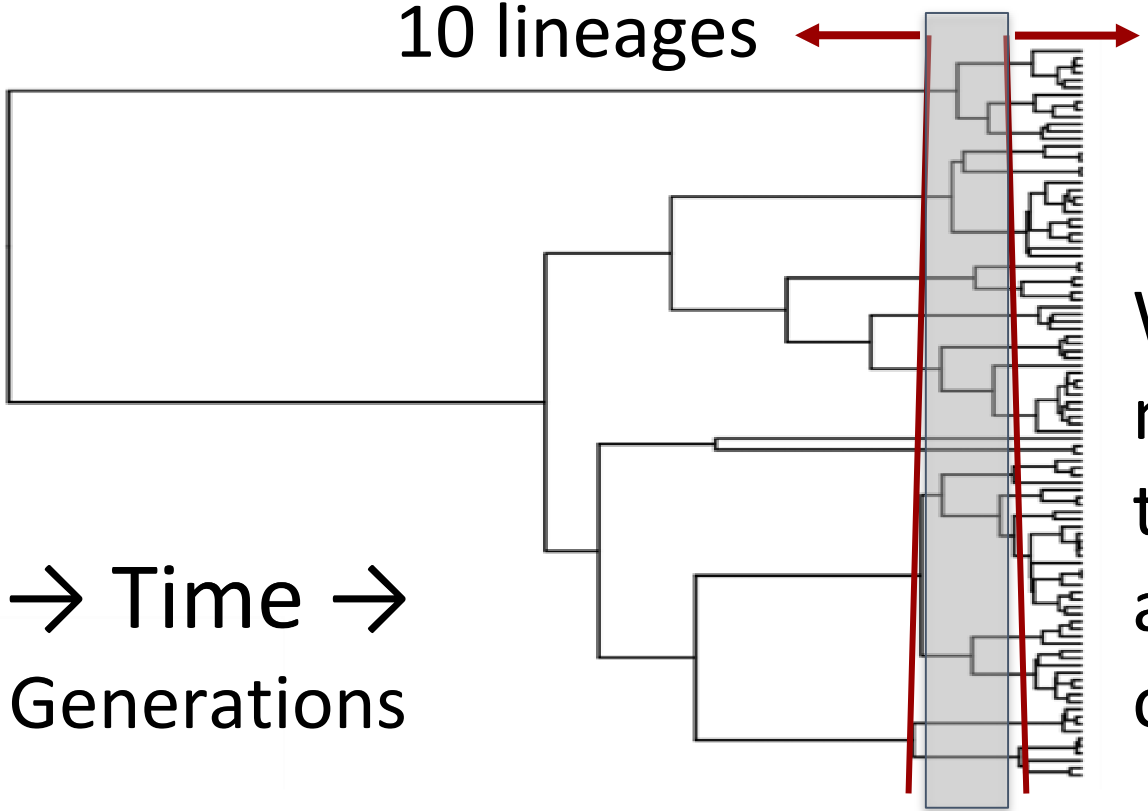
Probability for *k* lineages

$$\lambda_{c,k} = \left(\begin{array}{c} k \\ 2 \end{array} \right) \lambda_{c,2} = \frac{k(k-1)}{2}\lambda_{c,2}$$

Time for *k* lineages

$$E(t_k) = \frac{2}{k(k-1)}E(t_2)$$

10 Haplotypes contain 90% of common variation



10 lineages

> 20 lineages

→ Time →

Generations

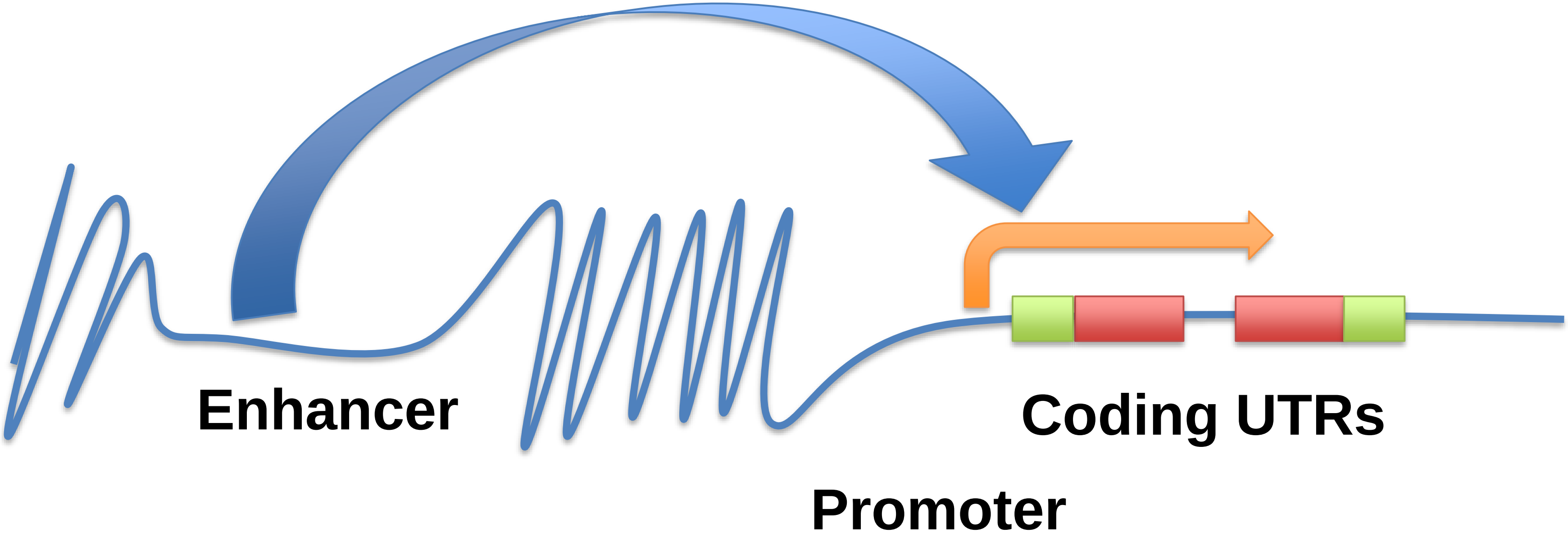
When there are many lineages they coalesce in a short amount of time

Genomic uses

- Once populated with 10-20 quality assemblies
  - E.g. 18% of 2 genes intervals were shared between B73 and W22
  - Custom genomes can be easily produced
  - Dramatically reduces problems with alignment
  - Map based cloning of genes

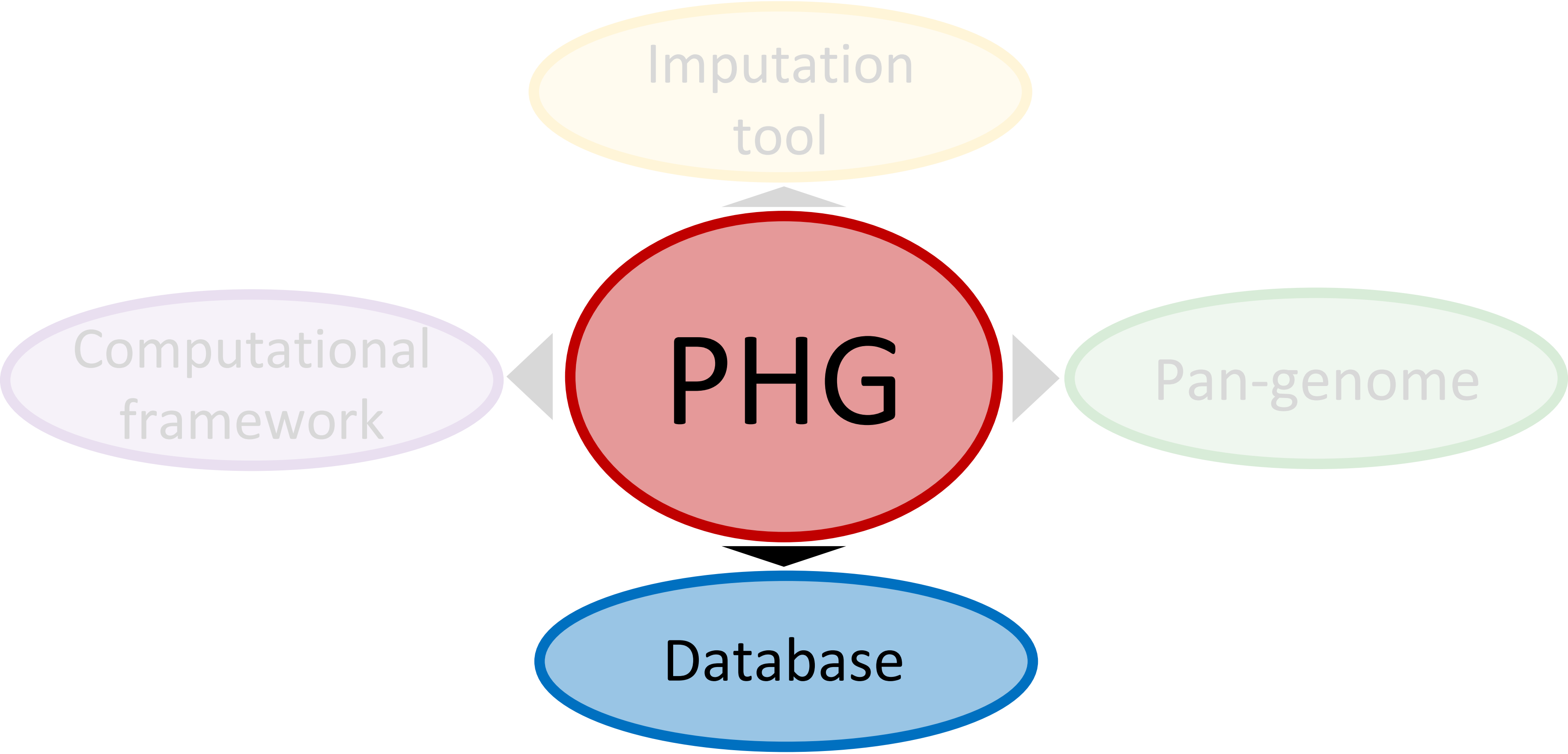
Haplotype epistasis

Likely epistasis between enhancer, promoter, UTRs, splicing, and coding changes. Haplotypes capture and can be used to model this.

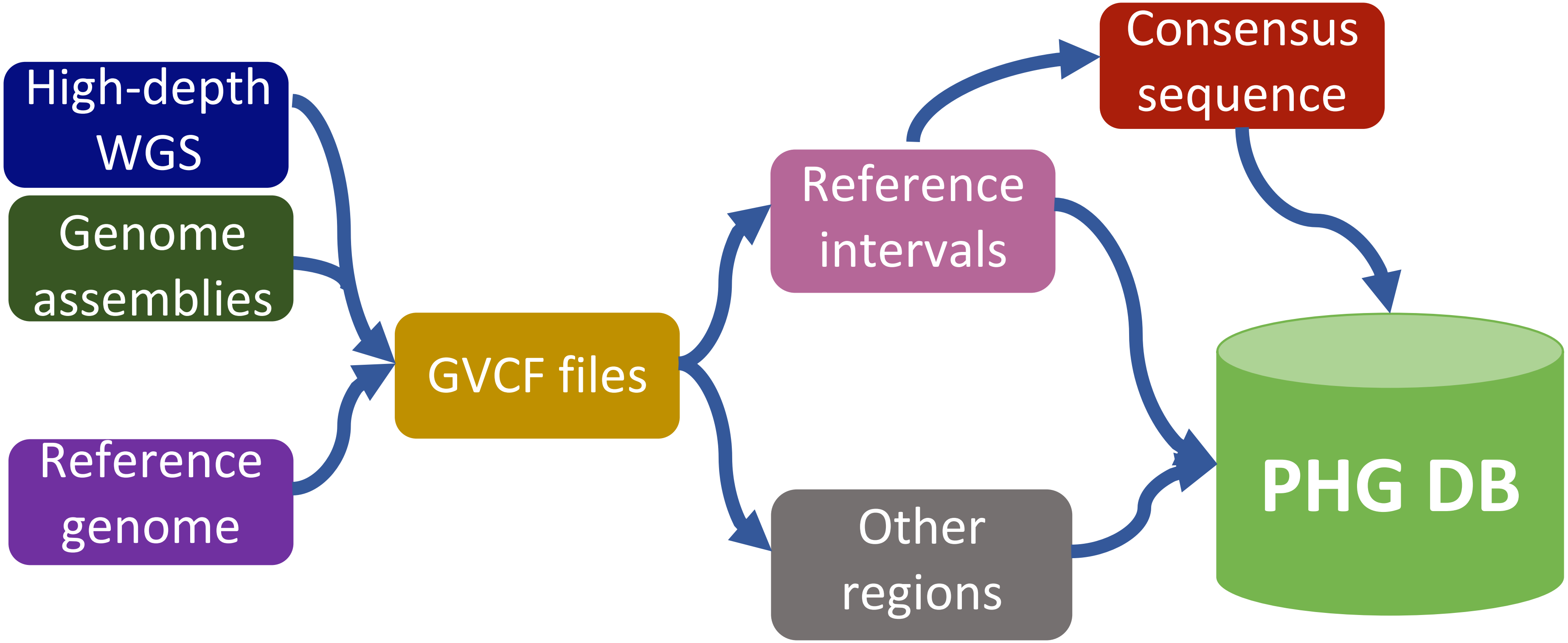


Haplotype annotation

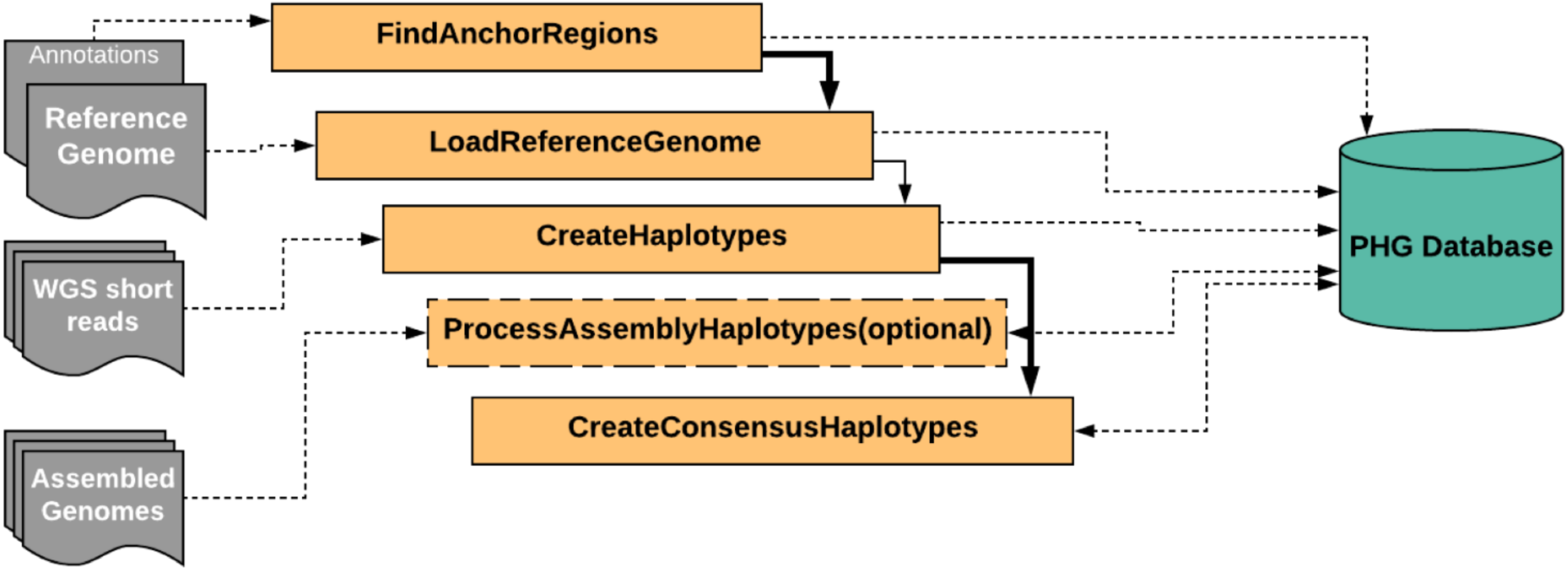
- Frameshift mutations
- Alternative splicing
- Promoter strength
- Expression level
- Deleterious mutations
- Yield estimate
- etc.

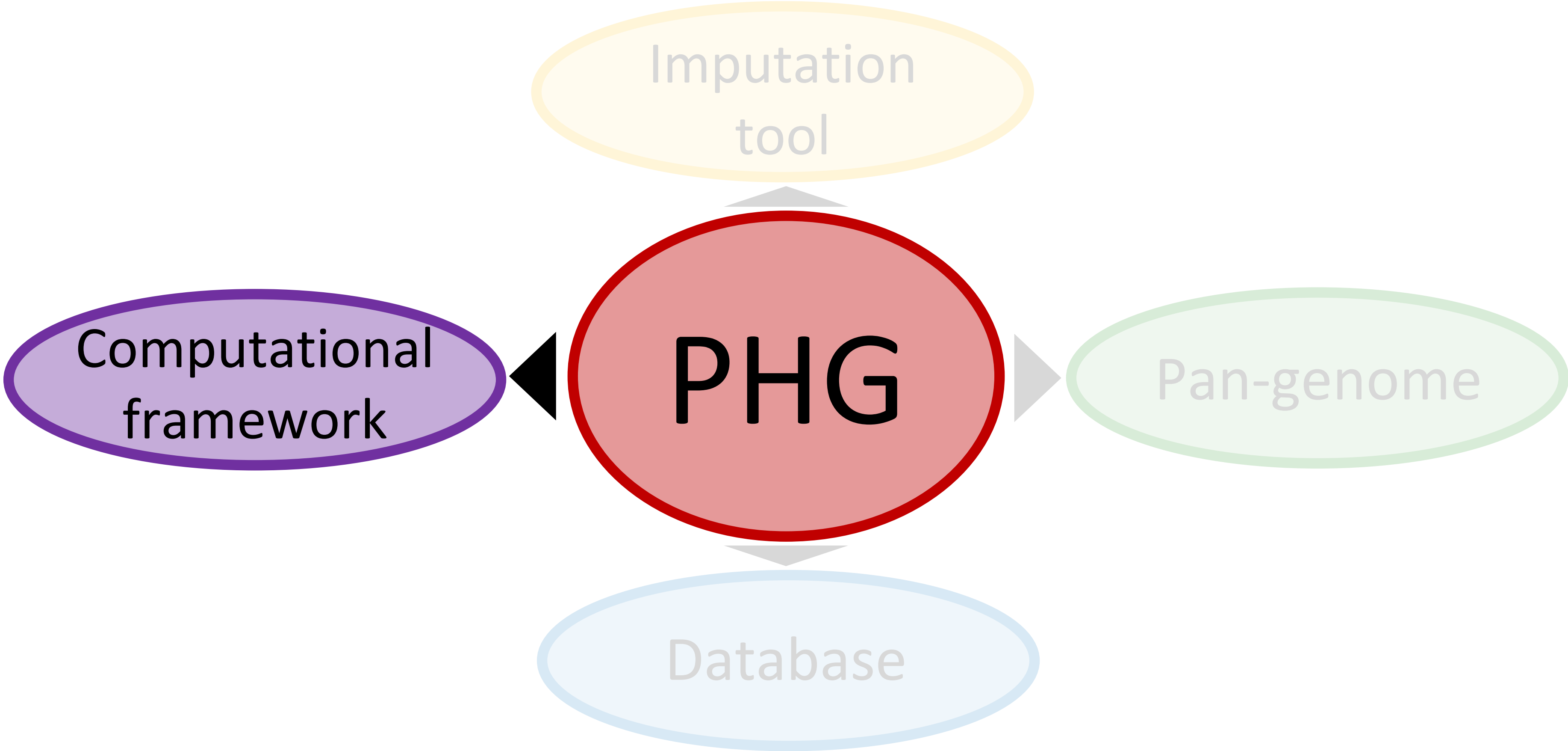

Imputation tool
Computational framework
PHG
Pan-genome
Database

Building the PHG database



Building the PHG database




Imputation tool
Computational framework
PHG
Pan-genome
Database

In-memory storage of a species

~10 consensus haplotypes might capture

>90% of common variation in a species

>99% of variation in breeding populations

Haplotype storage

10 haplotypes x 2 Gb = 20Gb

50,000 genomes x 40Kb for hapids = 2Gb

Whole genome storage

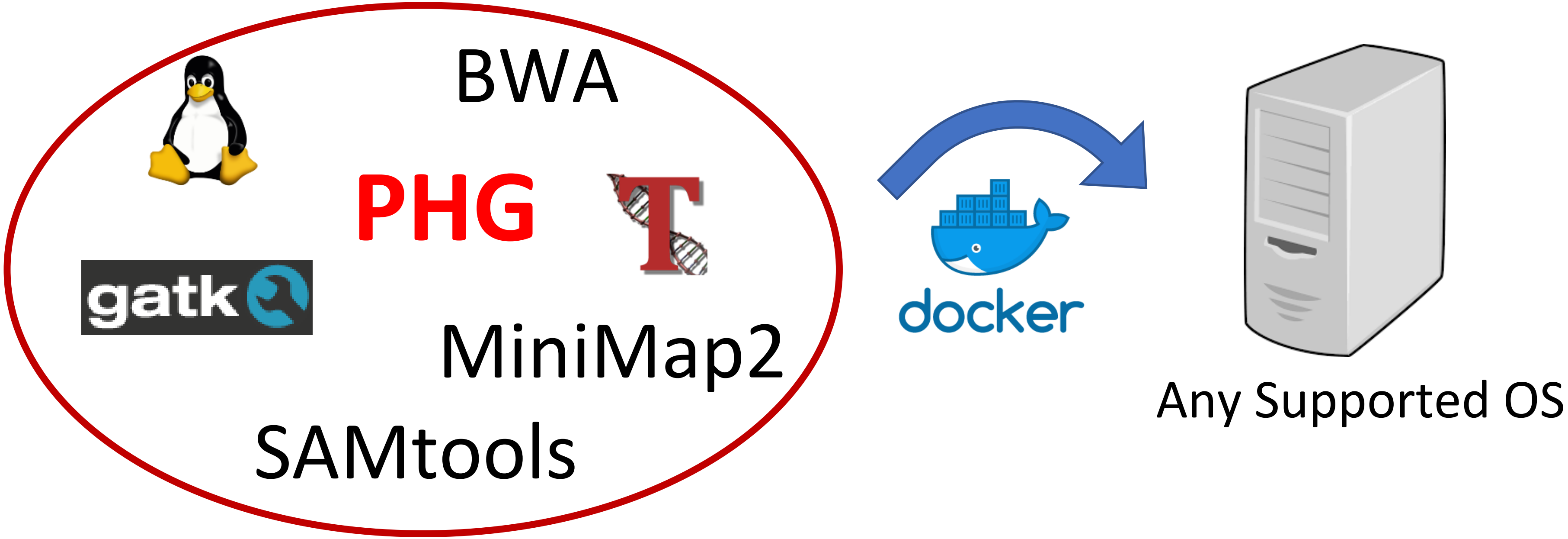
50,000 genomes x 2Gb = 100,000 GB or 100Tb

The PHG computational framework

- Create sqlite or postgres database with all haplotypes
- Software:
    - Populates database
    - Generates graph in-memory from the database
    - Uses the in-memory graph to predict new haplotypes
- Pipeline uses software from several sources
- Distributed as a Docker image

Designed to be relatively straightforward to run

A docker image captures the computing software environment



BWA

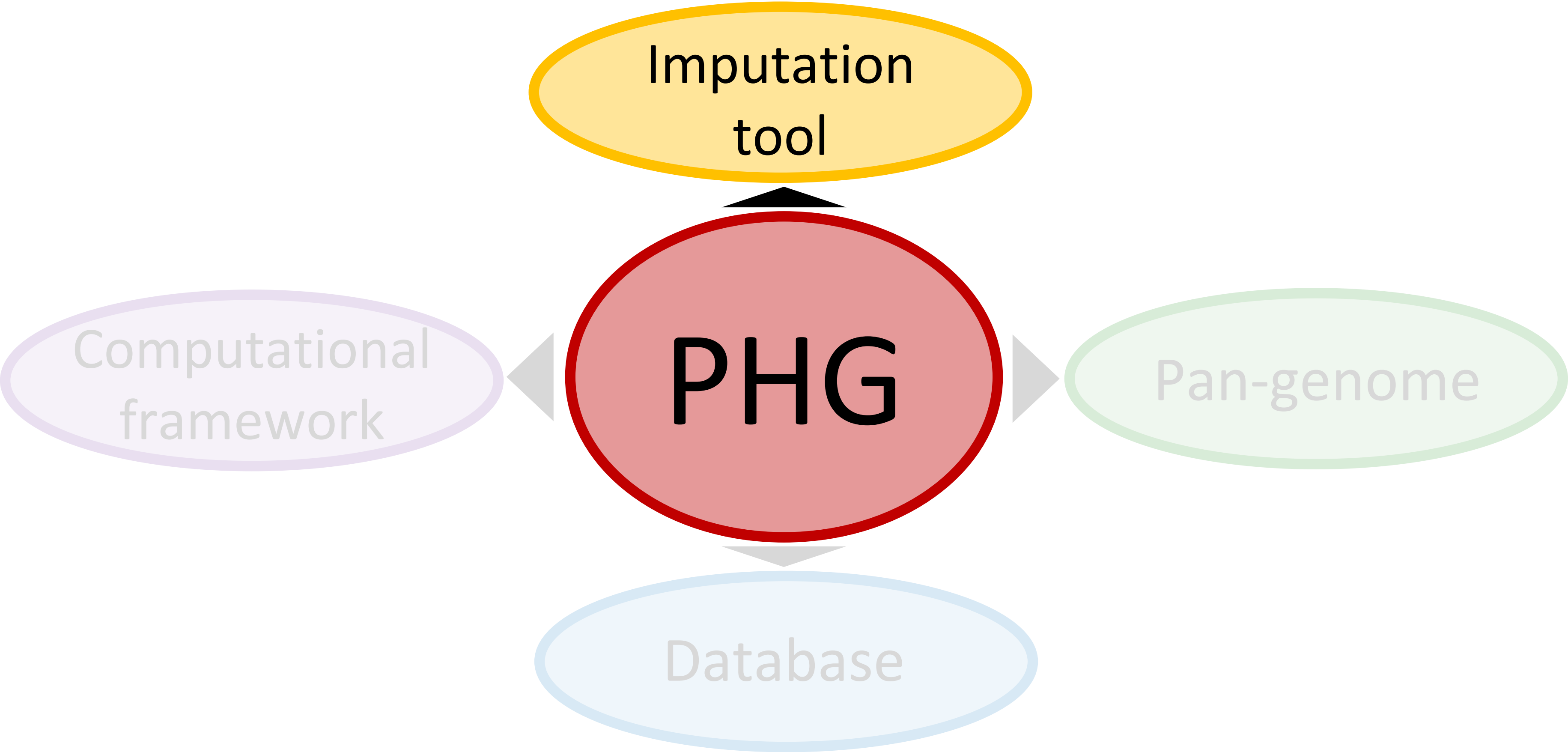
PHG

MiniMap2

SAMtools

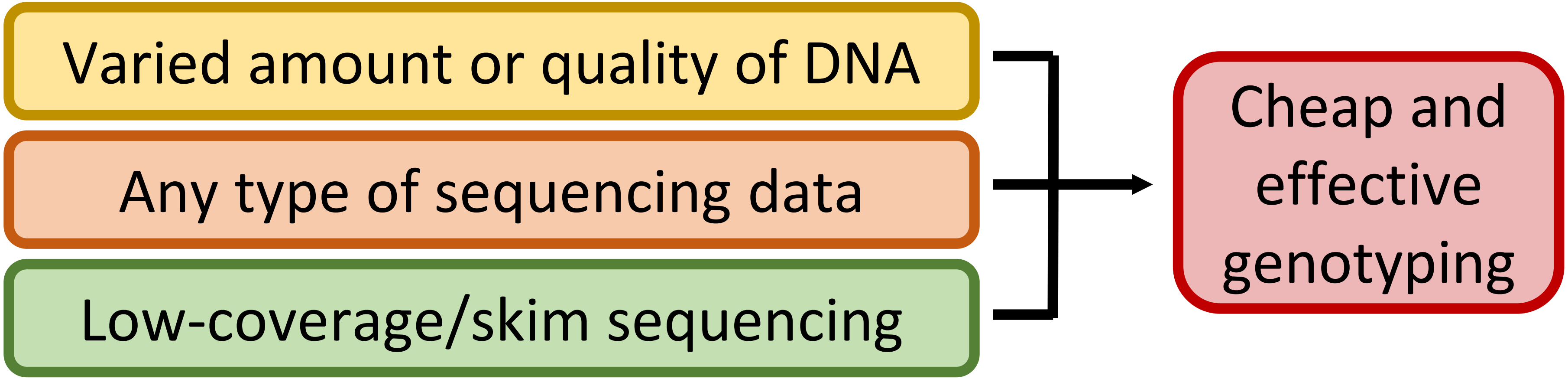
Any Supported OS

Using a docker image makes it easier to replicate analyses


Imputation tool
Computational framework
PHG
Pan-genome
Database

The PHG imputes using sequence from any source

Interchangeable vendors give:



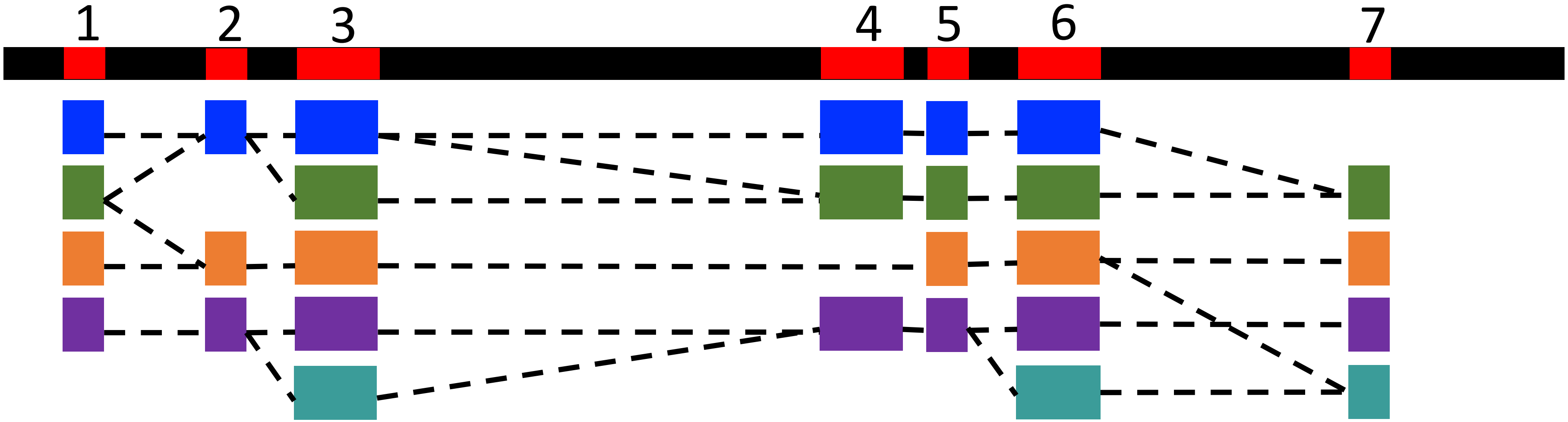
Varied amount or quality of DNA

Any type of sequencing data

Low-coverage/skim sequencing

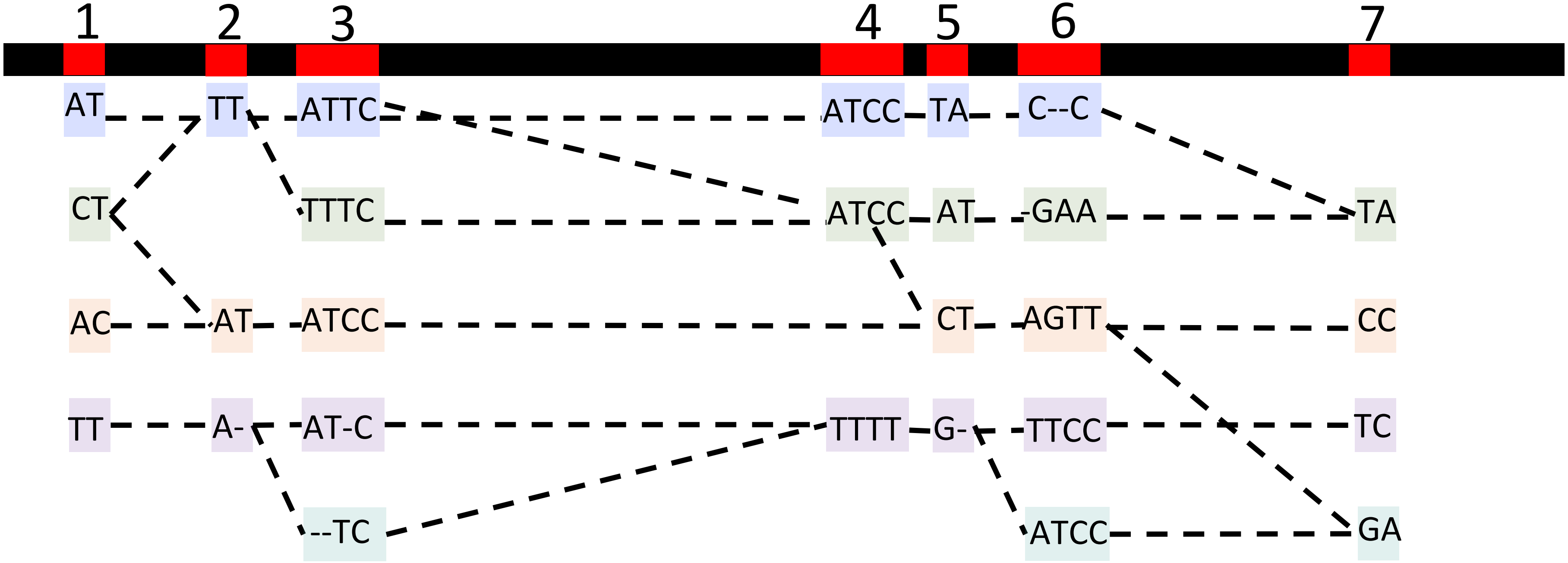
Cheap and effective genotyping

Within-anchor variant calling



Haplotypes for new individuals are imputed based on similarity to haplotypes in the graph

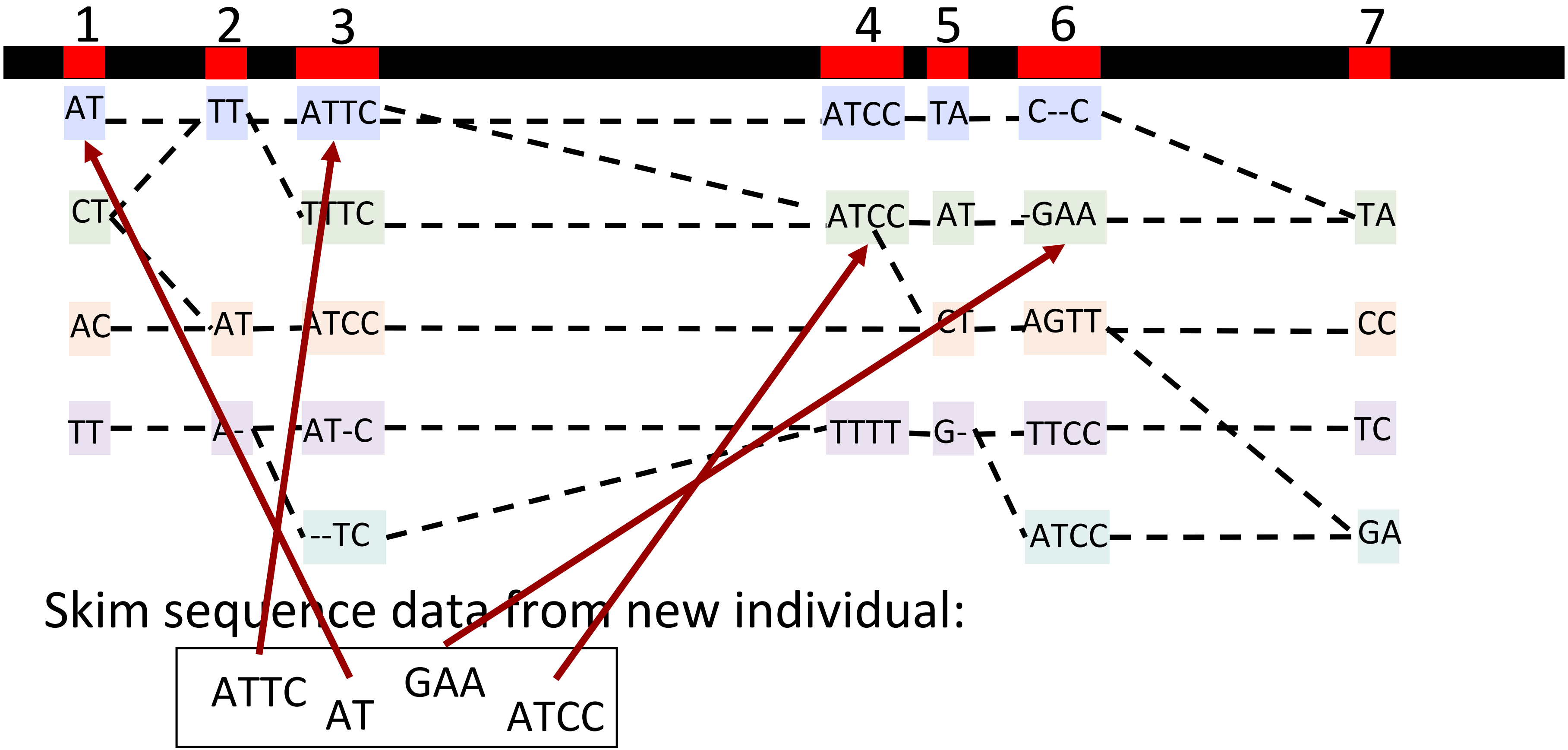
Across-anchor variant calling



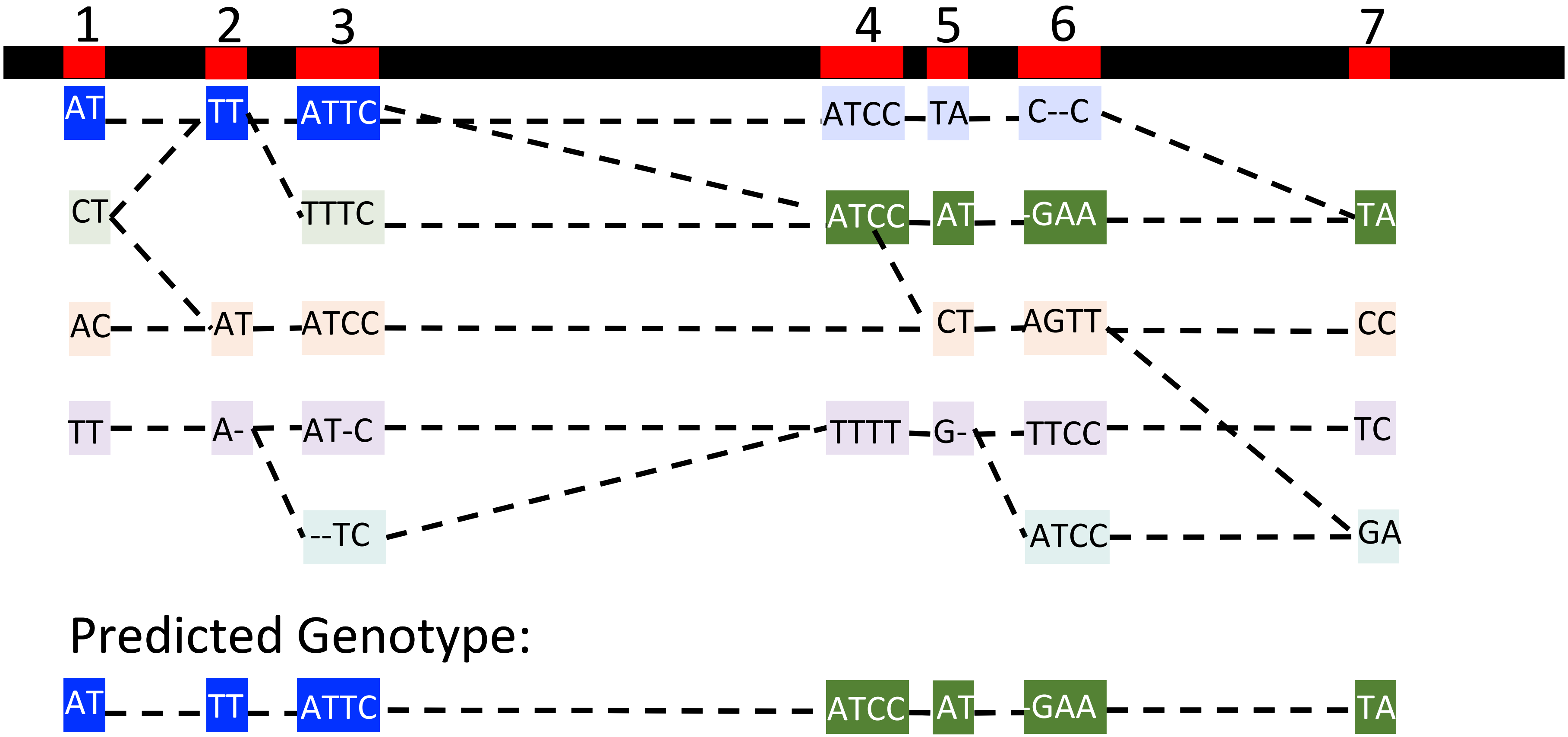
Skim sequence data from new individual:

ATTC AT GAA ATCC

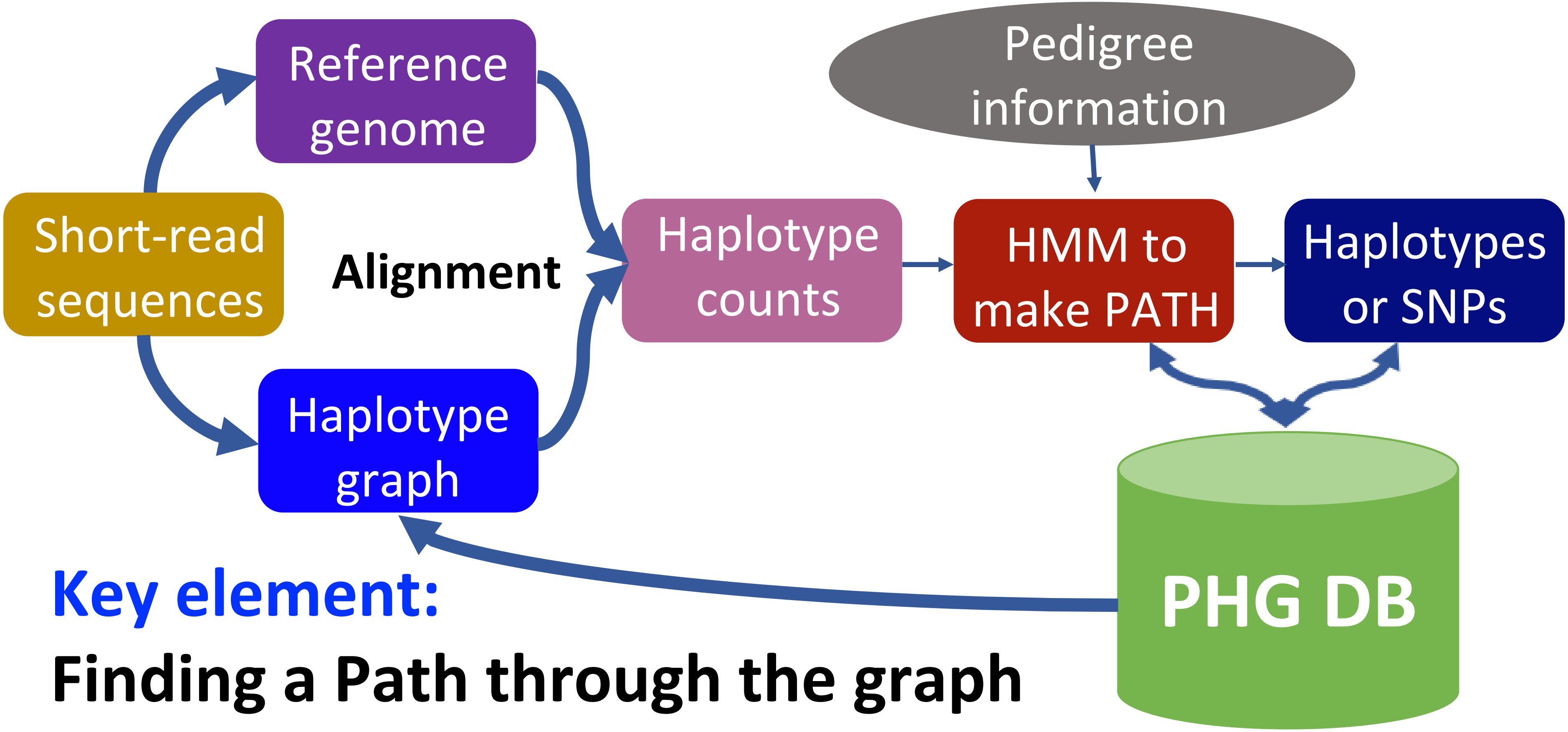
Align skim sequence to haplotypes



Deduce the best path through anchors



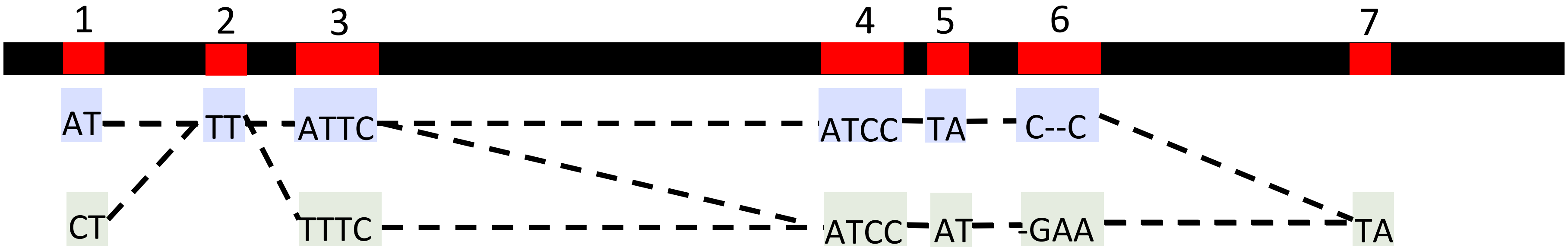
The PHG runs a Hidden Markov Model



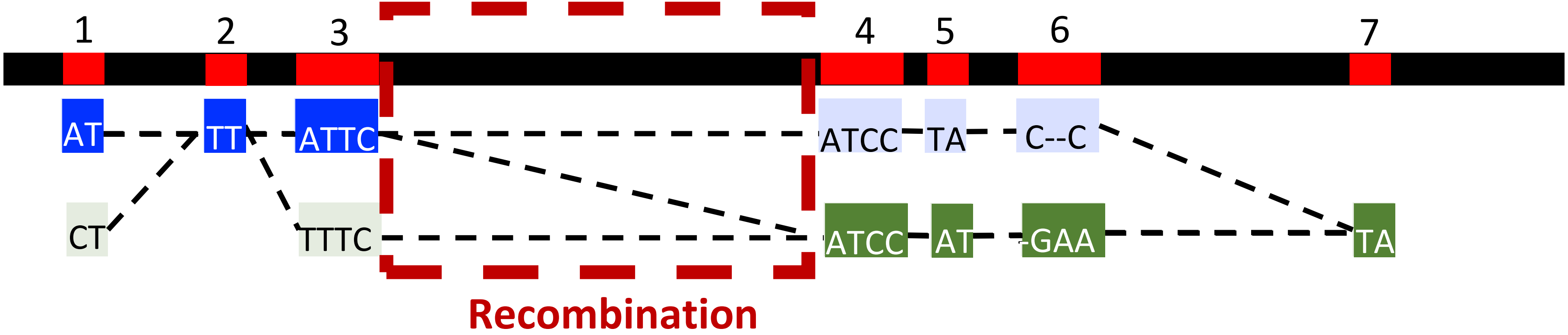
Key element:

Finding a Path through the graph

Bi-parental cross: restrict to parent haplotypes



Identify intervals with recombination



Aligning skim sequence through the graph helps identify recombination events in the progeny

Use case: Chibas sorghum breeding

CHIBAS

Key Traits

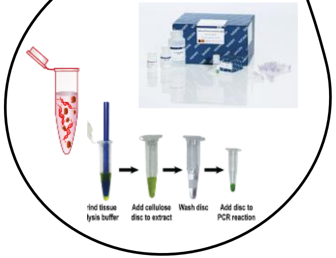
- Grain yield
- Stalk sugar content
- Biomass

2015 sugarcane aphid outbreak: most popular varieties no longer viable

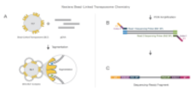
Challenge: Genomic Selection in relevant time frame

Collect Samples

Extract DNA



Sequence



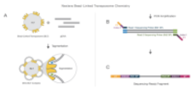
Impute Genotypes

Estimate Breeding Value

$y = X\beta + \varepsilon$

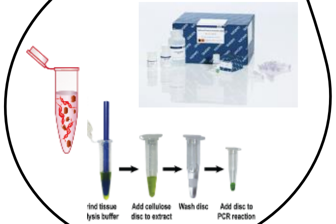
Select & Cross

Challenge: Genomic Selection in relevant time frame

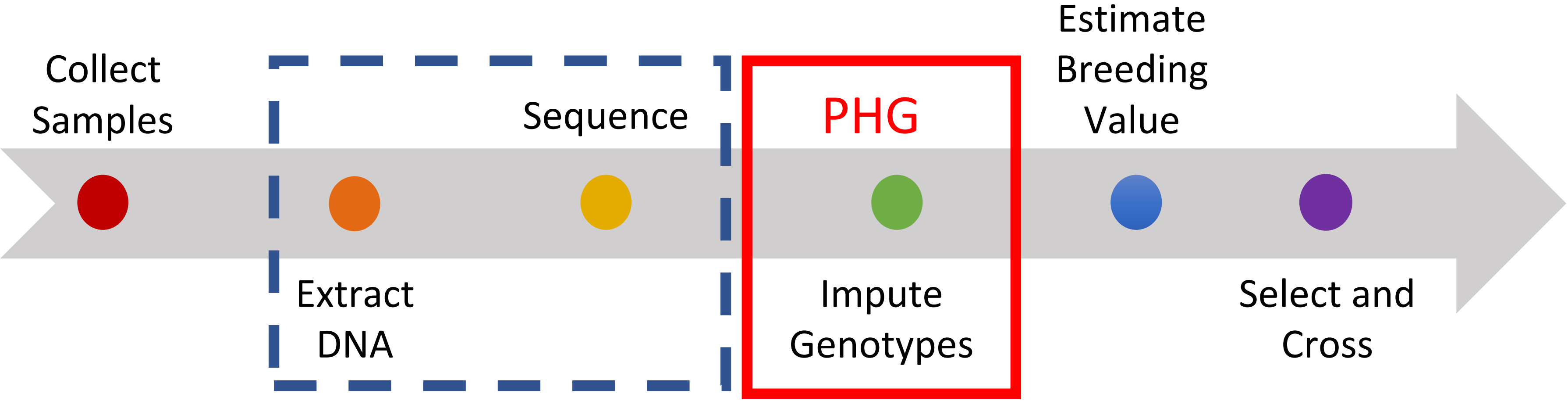


$y = X\beta + \varepsilon$

- Parents must be selected in time to make crosses
- Genomic selection requires cheap, scalable genotyping technologies

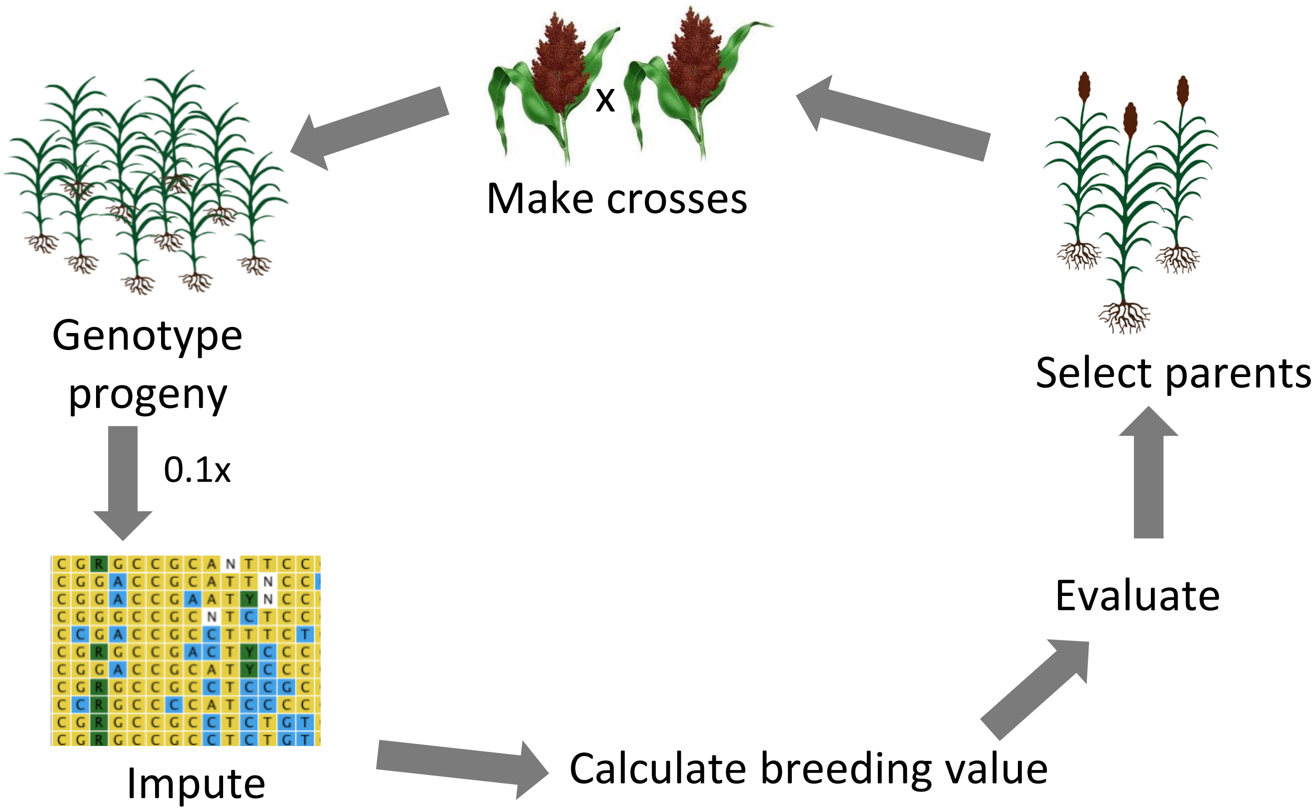


How do we make Genomic Selection cheap and scalable?



We need a system that is robust to technology changes

Cost-effective haplotype prediction for genomic selection on large progeny populations



x

Make crosses

Genotype progeny

0.1x

```
C G R G C C G C A N T T C C
C G G A C C G C A T T N C C
C G G A C C G A A T Y N C C
C G G G C C G C N T C T C C
C C G A C C G C C T T T C T
C G R G C C G A C T Y C C C
C G G A C C G C A T Y C C C
C G R G C C G C C T C C G C
C C R G C C C C A T C C C C
C G R G C C G C C T C T G T
C G R G C C G C C T C T G T
```

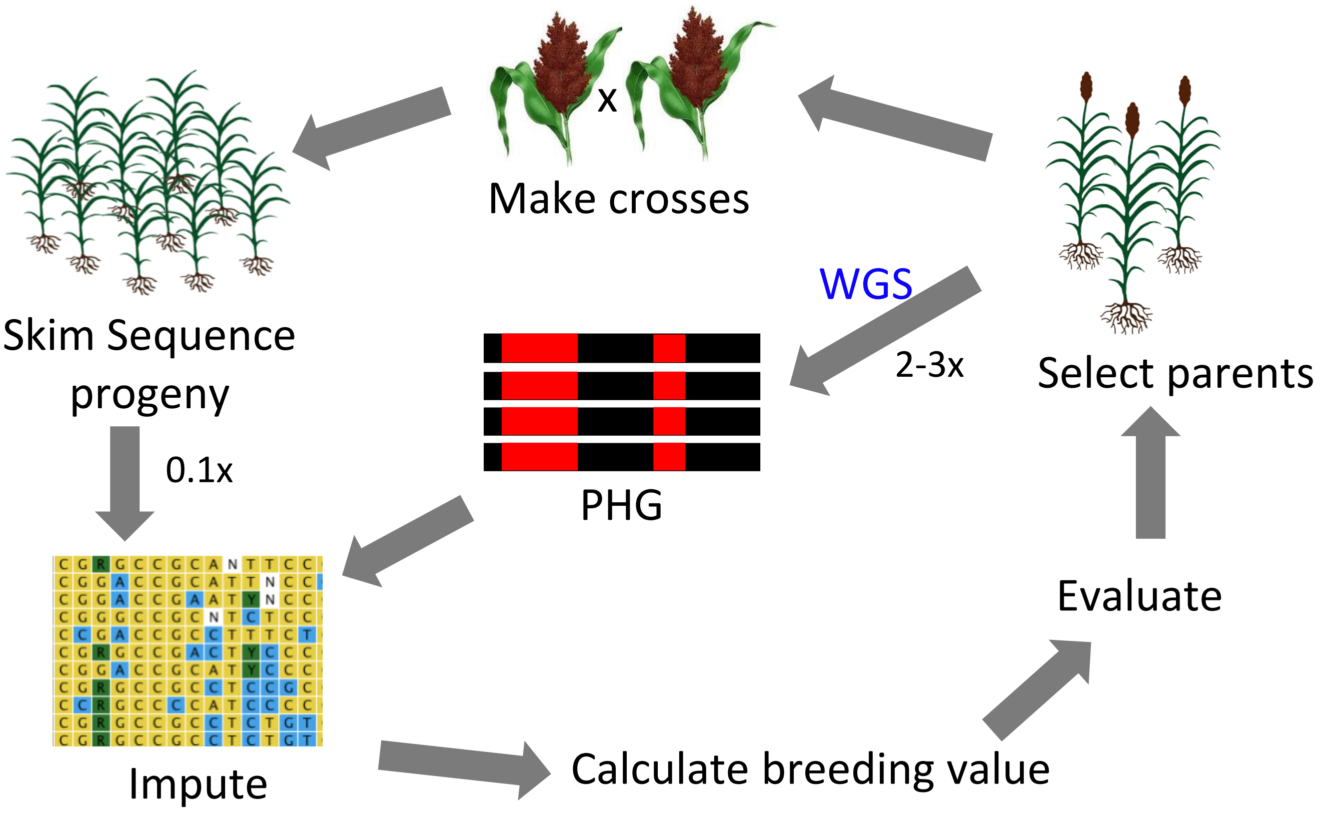
Impute

Calculate breeding value

Evaluate

Select parents

Cost-effective haplotype prediction for genomic selection on large progeny populations



x

Make crosses

Skim Sequence progeny

0.1x

C	G	R	G	C	C	G	C	A	N	T	T	C	C
C	G	G	A	C	C	G	C	A	T	T	N	C	C
C	G	G	A	C	C	G	A	A	T	Y	N	C	C
C	G	G	G	C	C	G	C	N	T	C	T	C	C
C	C	G	A	C	C	G	C	C	T	T	T	C	T
C	G	R	G	C	C	G	A	C	T	Y	C	C	C
C	G	G	A	C	C	G	C	A	T	Y	C	C	C
C	G	R	G	C	C	G	C	C	T	C	C	G	C
C	C	R	G	C	C	C	C	A	T	C	C	C	C
C	G	R	G	C	C	G	C	C	T	C	T	G	T
C	G	R	G	C	C	G	C	C	T	C	T	G	T

Impute

Calculate breeding value

Evaluate

Select parents

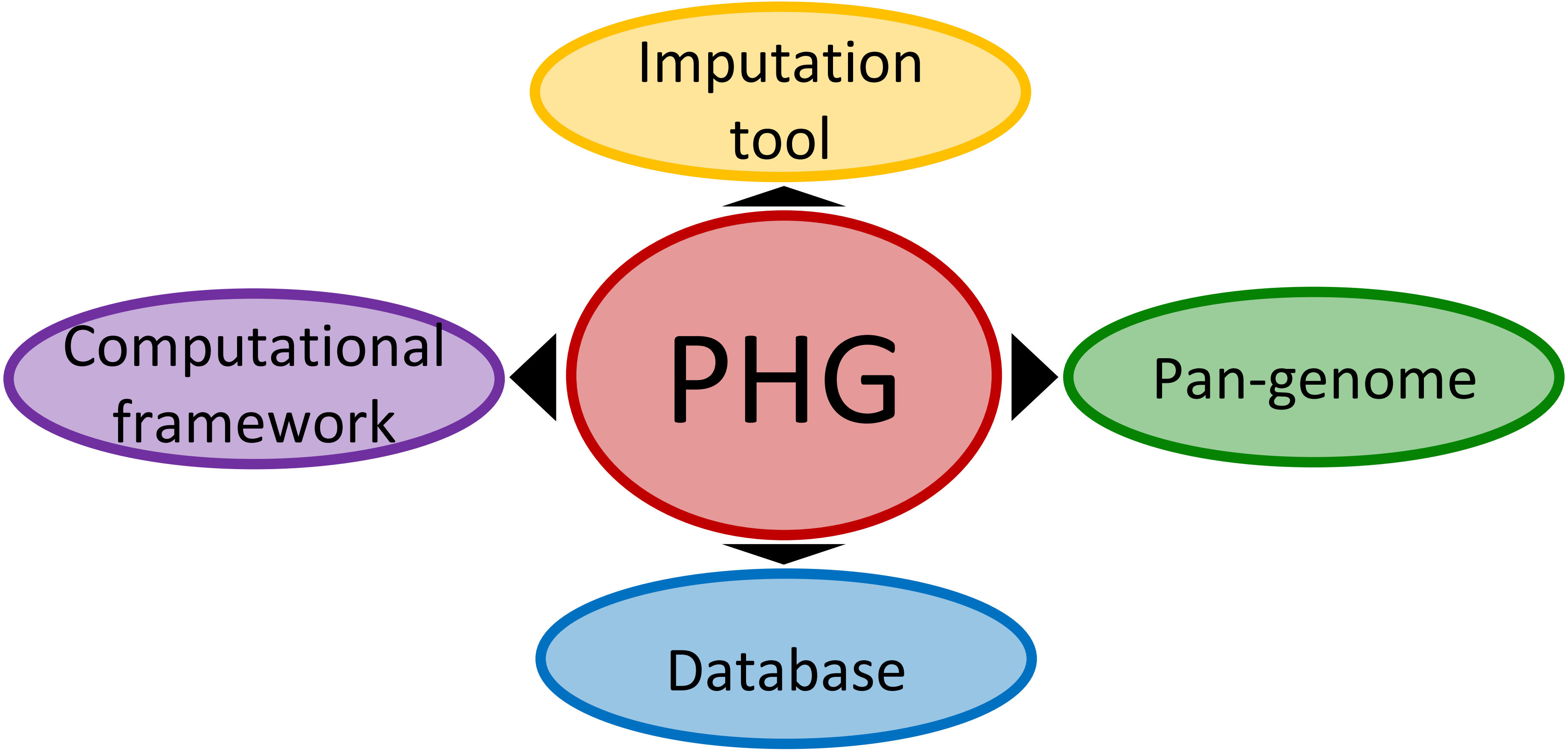
WGS

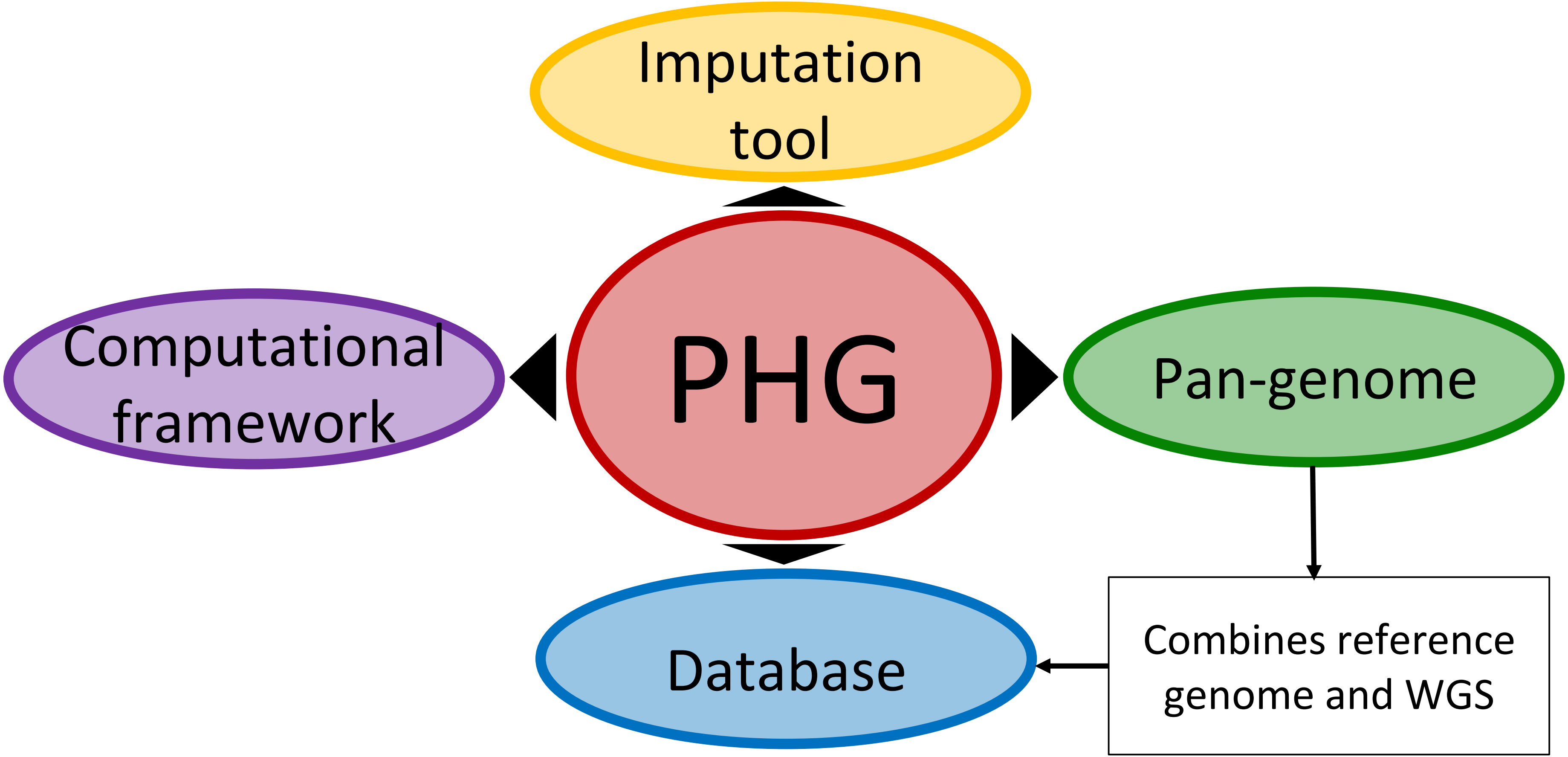
2-3x

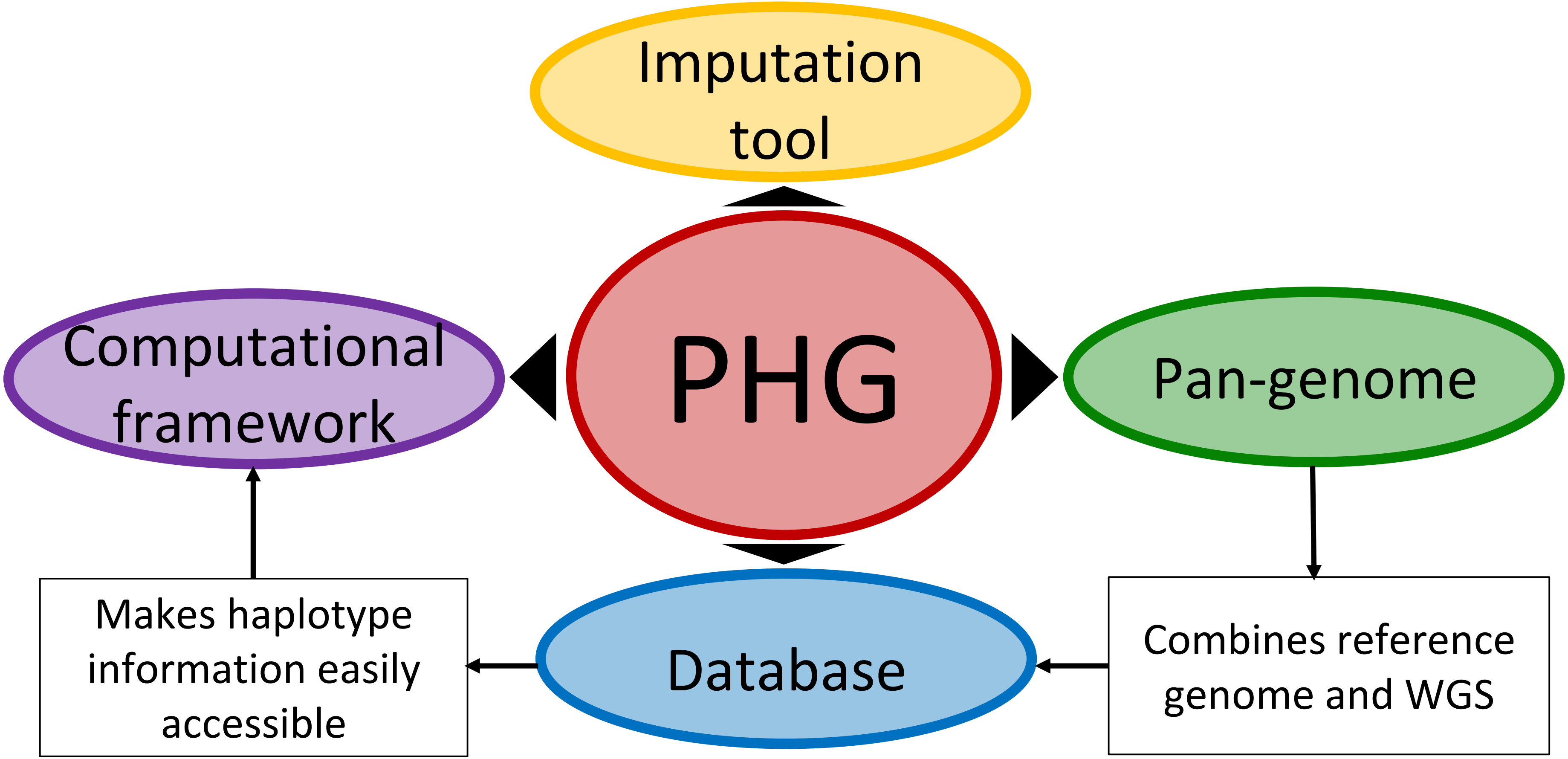
PHG

Use case: positional cloning in wheat

You tell us how this might be used!


Imputation tool
Computational framework
PHG
Pan-genome
Database


Imputation tool
Computational framework
PHG
Pan-genome
Database
Combines reference genome and WGS



Imputation tool

Computational framework

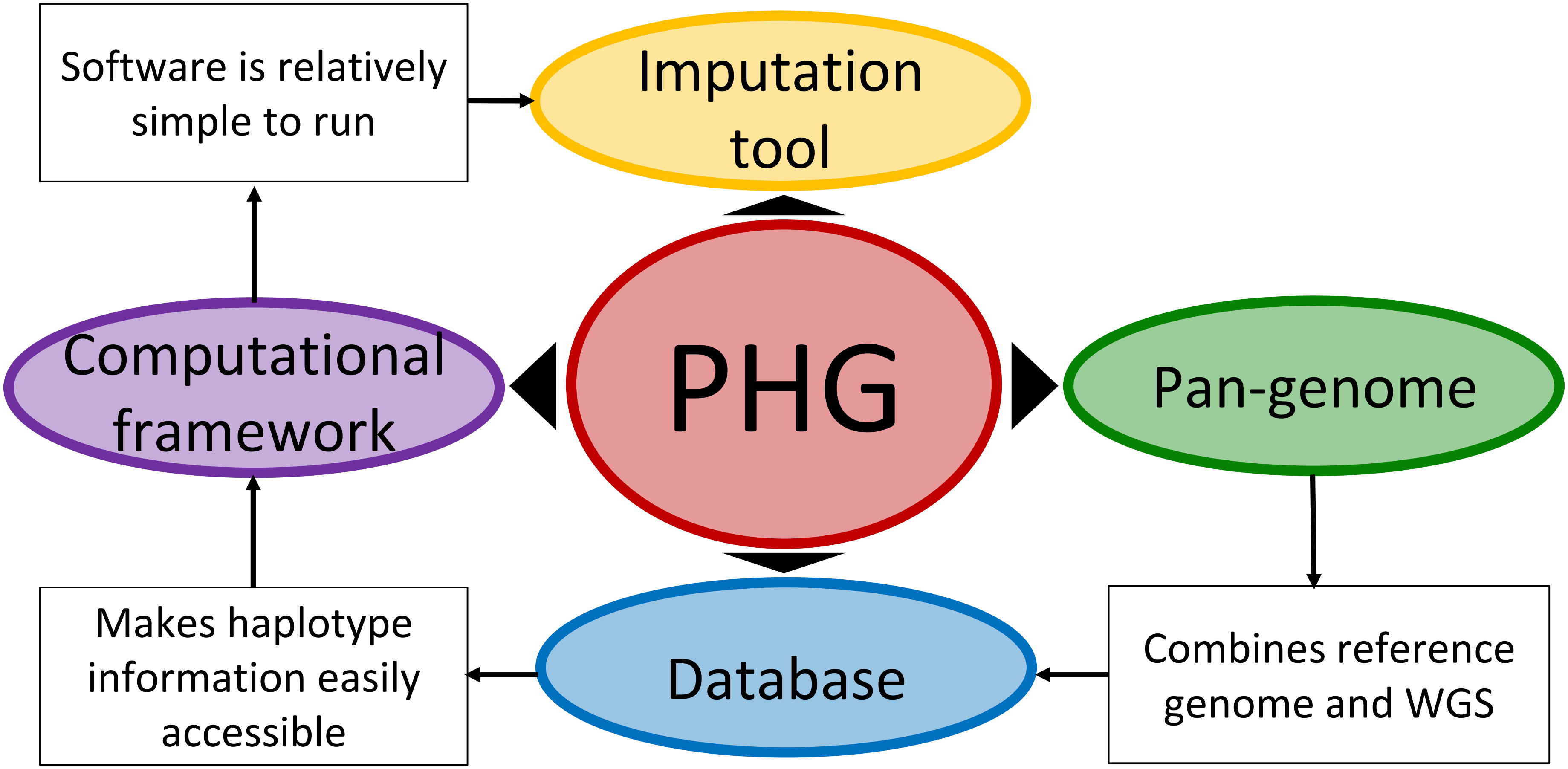
PHG

Pan-genome

Database

Makes haplotype information easily accessible

Combines reference genome and WGS


Software is relatively simple to run

Imputation tool

Computational framework

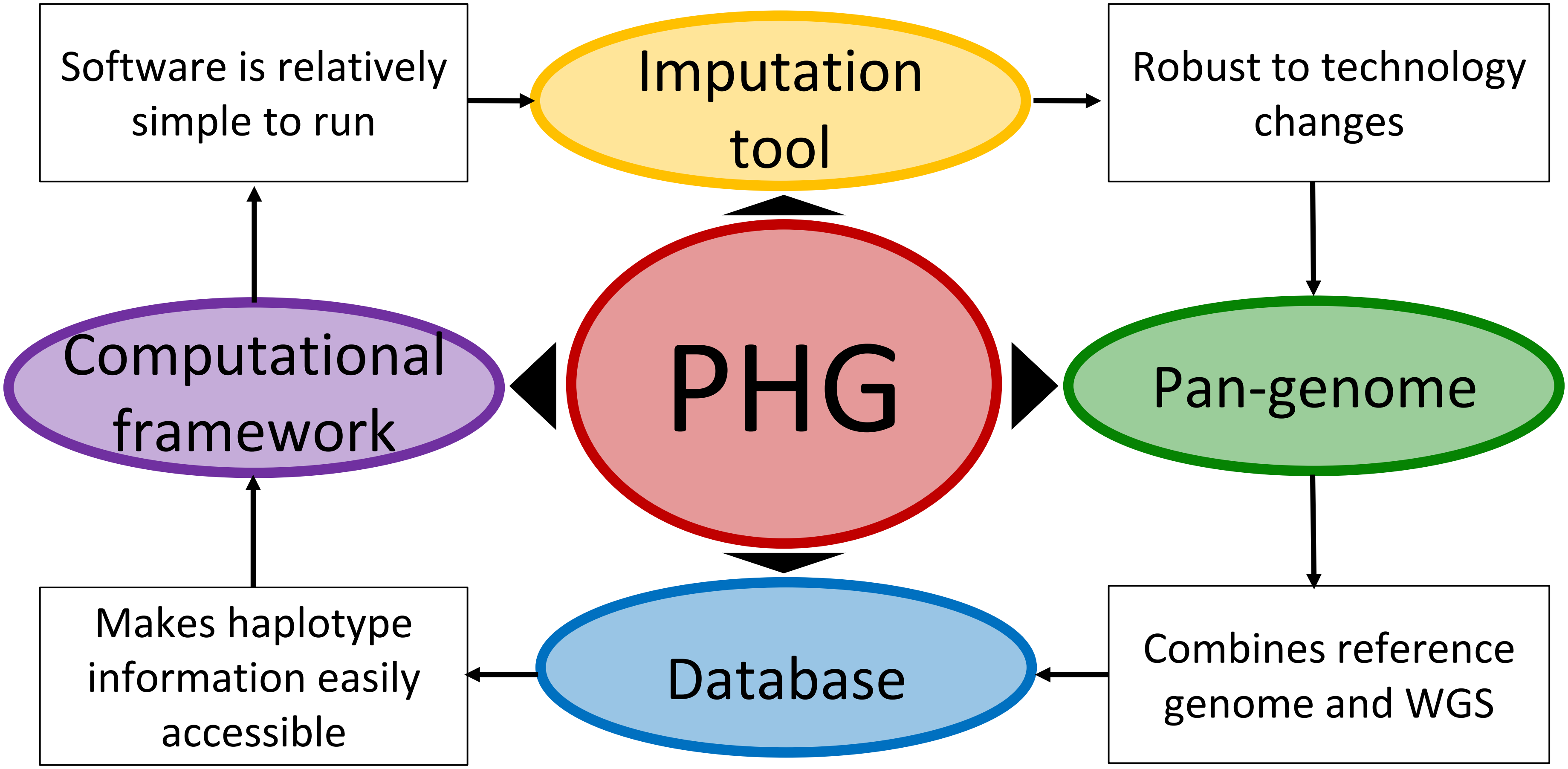
PHG

Pan-genome

Makes haplotype information easily accessible

Database

Combines reference genome and WGS



Software is relatively simple to run

Imputation tool

Robust to technology changes

Computational framework

PHG

Pan-genome

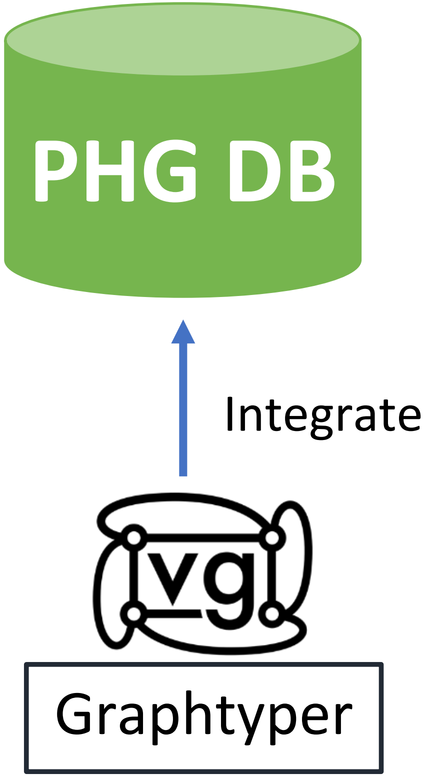
Makes haplotype information easily accessible

Database

Combines reference genome and WGS

What does the PHG change?

- Easy to produce custom genomes for a breeding program
- Replaces GBS, rAmpSeq, and low coverage informatic pipelines
- Facilitates use of low coverage random sequence data



Limitations of the PHG

- 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

Where are we going?

- You tell us!!!
- Improve haplotype identification with low coverage
- Storage of rare allele amendments to consensi
- Improve GS performance
- GUI drivers in TASSEL, R?, Jupyter?
- Robust annotation of haplotypes

PHG