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

Tracing the ancestry of modern bread wheats

Caroline Pont1,2, Thibault Leroy2,3,22, Michael Seidel4,22, Alessandro Tondelli5,22, Wandrille Duchemin1,22, David Armisen1,22, Daniel Lang4,22, Daniela Bustos-Korts6,22, Nadia Goué1,2, François Balfourier1, Márta Molnár-Láng1, Jacob Lage6, Benjamin Kilian10,11, Hakan Özkan12, Darren Waite13, Sarah Dyer14, Thomas Letellier15, Michael Alaux15, Wheat and Barley Legacy for Breeding Improvement (WHEALBI) consortium16, Joanne Russell17, Beat Keller1,8, Fred van Eeuwijk6, Manuel Spannagl14, Klaus F. X. Mayer4,5, Robbie Waugh17,20,21, Nils Stein4,11, Luigi Cattivelli5,23, Georg Haberer4,7, Gilles Charmet1,23 and Jérôme Salse1,2,3,*
Even in wheat, many individuals are being sequenced

Tracing the ancestry of modern bread wheats

Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome
Even in wheat, many individuals are being sequenced

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

The pan-genome can be accurately represented as a graph

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

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

<table>
<thead>
<tr>
<th>B73</th>
<th>Mo17</th>
<th>0h43</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGCGT</td>
<td>ACCGT</td>
<td>AGTCT</td>
</tr>
<tr>
<td>ACGAGT- - - -CATGA</td>
<td>ACGNNTAAAAACATGA</td>
<td>ACGAGTAA- -CANNA</td>
</tr>
<tr>
<td>CGTAA</td>
<td>CGCAA</td>
<td>CGCAA</td>
</tr>
</tbody>
</table>

Gene1

Gene2
This pattern differentiates ranges

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

B73
Mo17
Oh43

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

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?

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

Pan-genome

Database

PHG
PHG Imputation tool Pan-genome Database

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

Node = haplotype

Edge = connection between nodes
A population of chromosomes provides the basis for haplotype groups

- Cluster haplotypes at each anchor region
- Reduce memory footprint
- Increase haplotype coverage for better quality
Haplotypes at a single gene in the PHG

Gene 1

- Haplotype group 1
- Haplotype group 2
- Haplotype group 3
- Haplotype group 4
Haplotypes at a single gene in the PHG

Gene 1

Haplotypes:
- T1
- T2
- T3
- T4

Haplotype group 1
Haplotypes at a single gene in the PHG

Gene 1

Consensus Haplotype 1

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

\[ \lambda_{c,2} = \frac{1}{N} \]

\[ \lambda_{c,2} = \frac{1}{2N} \]
Expected *time* for two lineages to coalesce

\[ E(t_{c,2}) = \frac{1}{\lambda_{c,2}} = 2N \]

Probability for \( k \) lineages

\[ \lambda_{c,k} = \binom{k}{2} \lambda_{c,2} = \frac{k(k-1)}{2} \lambda_{c,2} \]

Time for \( k \) lineages

\[ E(t_k) = \frac{2}{k(k-1)} E(t_2) \]
10 Haplotypes contain 90% of common variation

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. Haplotype capture and can be used to model this.
Haplotype annotation

- Frameshift mutations
- Alternative splicing
- Promoter strength
- Expression level
- Deleterious mutations
- Yield estimate
- etc.
Building the PHG database

High-depth WGS
Genome assemblies
Reference genome
GVCF files
Reference intervals
Consensus sequence
Other regions
PHG DB
Building the PHG database
PHG Imputation tool Computational framework 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 40 Kb for hapids = 2 Gb

Whole genome storage
50,000 genomes x 2 Gb = 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

Using a docker image makes it easier to replicate analyses
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

Skim sequence data from new individual:

ATTC  AT  GAA  ATCC
Deduce the best path through anchors

Predicted Genotype:
The PHG runs a Hidden Markov Model

Reference genome

Short-read sequences

Alignment

Haplotype graph

Haplotype counts

HMM to make PATH

Pedigree information

Haplotypes or SNPs

PHG DB

Key element: Finding a Path through the graph
Bi-parental cross: restrict to parent haplotypes
Aligning skim sequence through the graph helps identify recombination events in the progeny.
Use case: Chibas sorghum breeding

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
- Sequence
- Estimate Breeding Value
- Extract DNA
- Impute Genotypes
- Select & Cross

\[ y = X\beta + \varepsilon \]
Challenge: Genomic Selection in relevant time frame

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

\[ y = X\beta + \varepsilon \]
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

1. Make crosses
2. Select parents
3. Evaluate
4. Calculate breeding value
5. Impute
6. Genotype progeny
7. 0.1x
Cost-effective haplotype prediction for genomic selection on large progeny populations

1. Make crosses
2. Skim Sequence progeny
3. Impute
4. WGS 2-3x
5. Select parents
6. Evaluate
7. PHG
8. Calculate breeding value
Use case: positional cloning in wheat

You tell us how this might be used!
PHG

Imputation tool

Computational framework

Pan-genome

Database
PHG Imputation tool

Computational framework

Database

Pan-genome

Combines reference genome and WGS
Imputation tool

PHG

Computational framework
Makes haplotype information easily accessible

Database

Pan-genome
Combines reference genome and WGS
Software is relatively simple to run

Computational framework

Imputation tool

PHG

Pan-genome

Database

Combines reference genome and WGS

Makes haplotype information easily accessible
PHG Imputation tool

**Computational framework**
- Makes haplotype information easily accessible

**Database**

**Pan-genome**
- Combines reference genome and WGS

**Software is relatively simple to run**

**Robust to technology changes**
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

Integrate

Graphtyper

PHG DB
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