

Wheat CAP: Genomic resources / Development of Practical Haplotype Graph

Eduard Akhunov

Wheat Genetic Resources Center
Department of Plant Pathology
Kansas State University

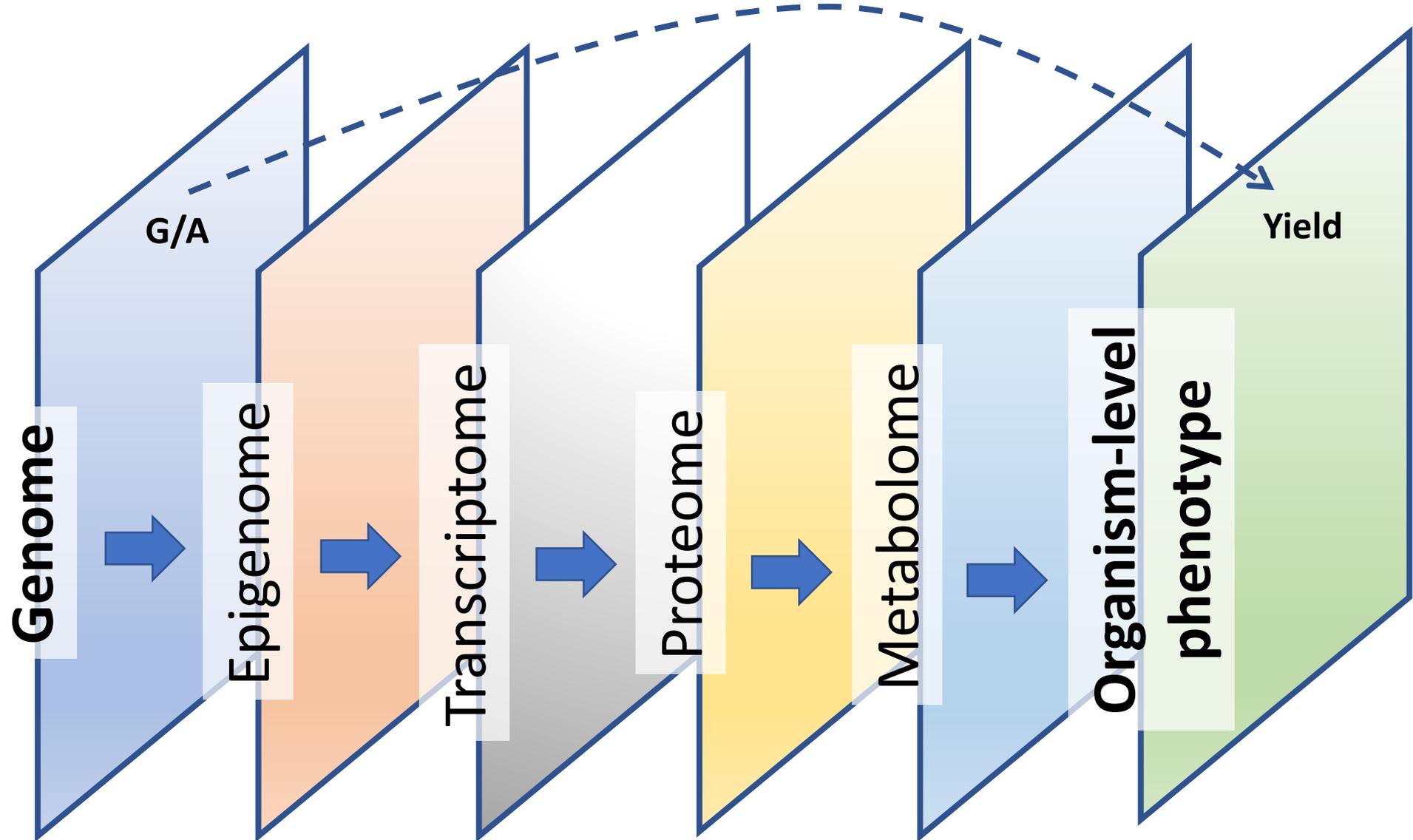


United States Department of Agriculture
National Institute of Food and Agriculture

WheatCAP 22-26, Zoom meeting, Dec. 15 2021



Constructing genome-to-phenome maps is critical for predicting phenotypic outcomes of a particular genotype

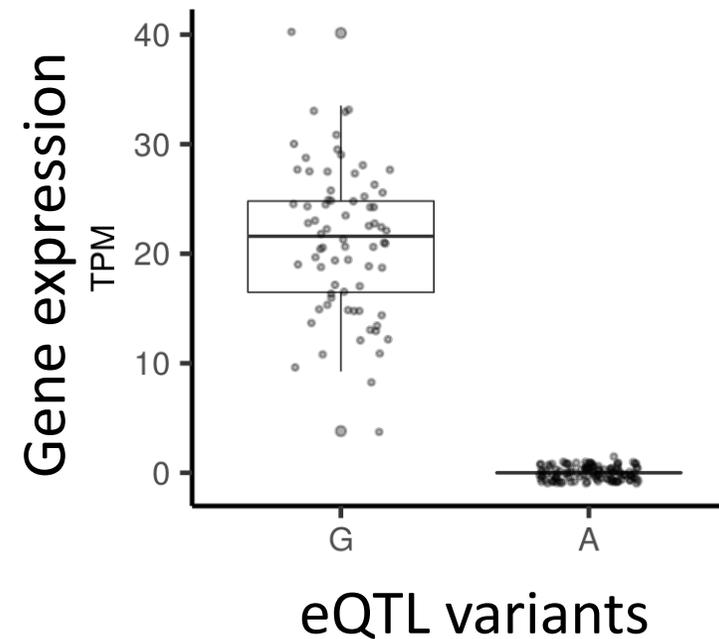
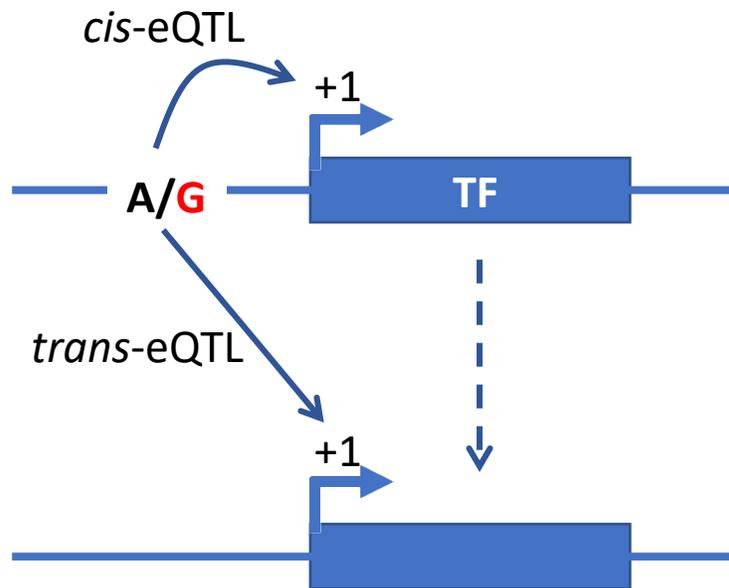


Understanding the genetic control of gene expression variation (eQTL) can improve our ability to detect causal genes and pathways

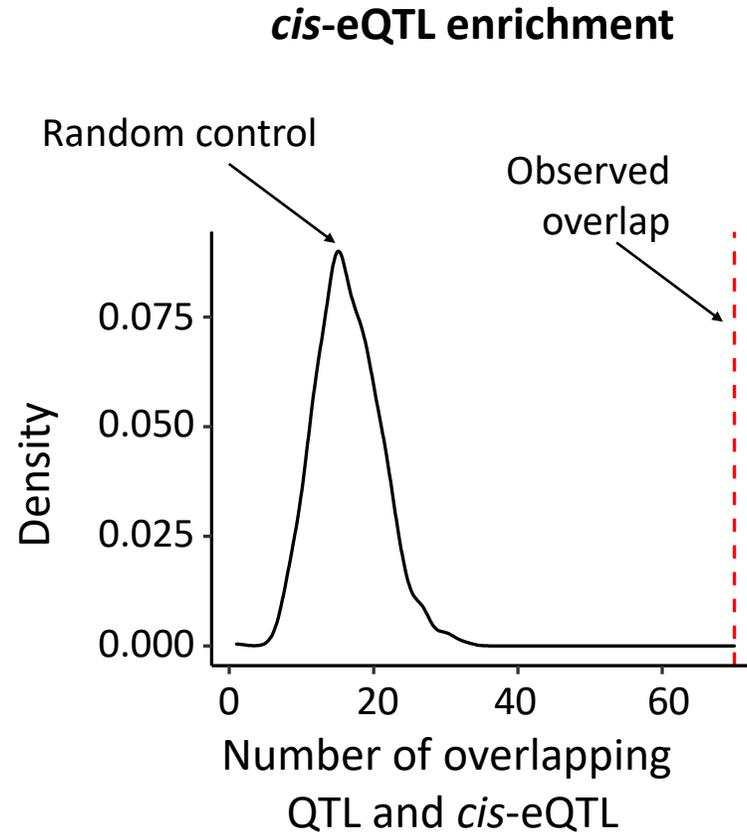
RNA-seq data for 198 diverse wheat accessions

Sequence- and array-based genotyping at 2.4 million SNP sites using Nimblegen wheat regulatory capture (Gardiner et al., 2019)

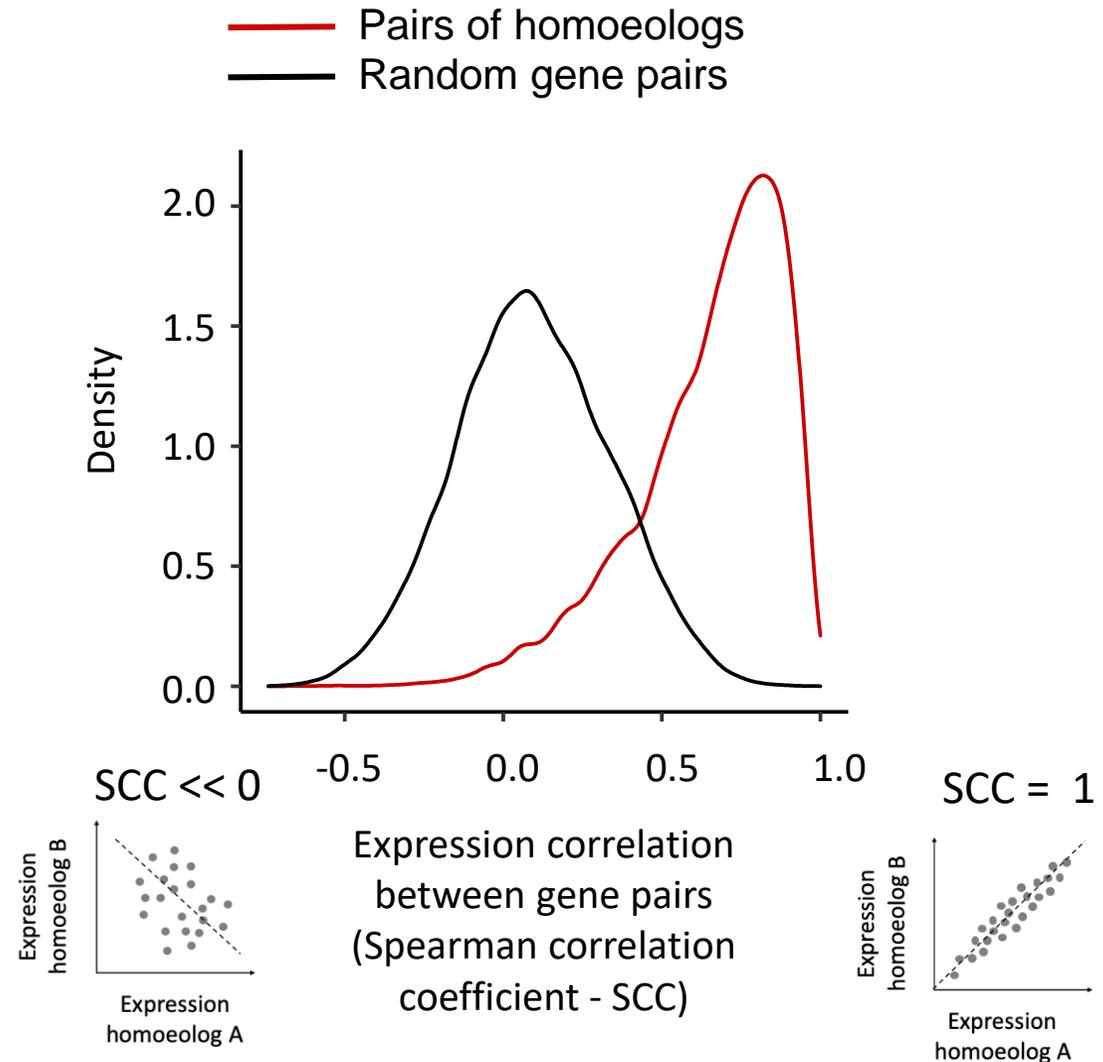
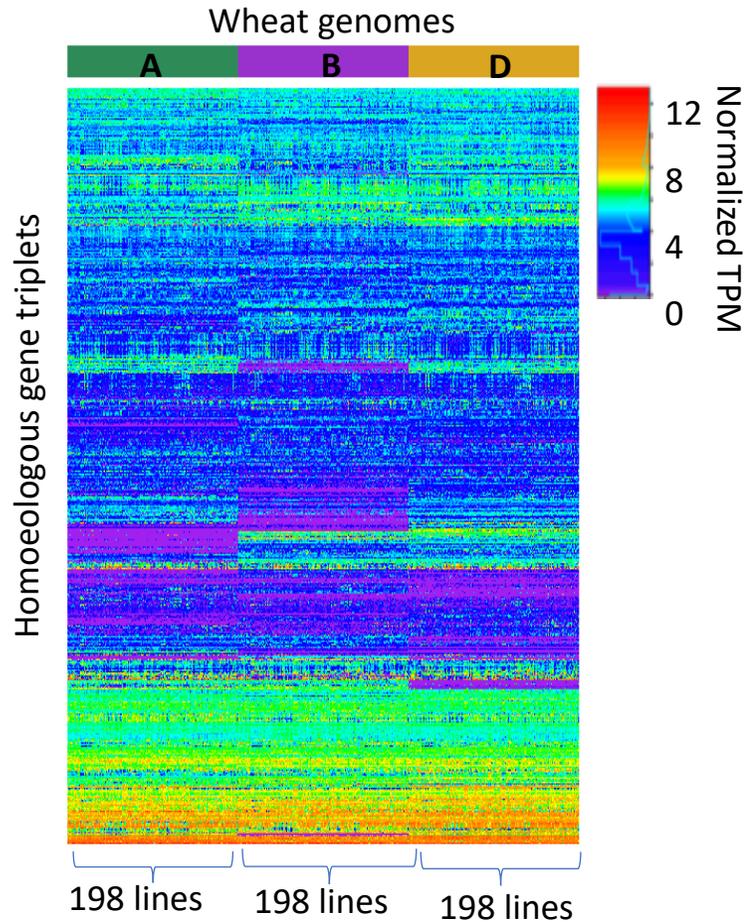
15K cis/trans-eQTL in spikes and seedlings (T3 database)



Interpretation of GWAS: *cis*-eQTL are significantly enriched around (+/- 1 kb) marker-trait associations identified in the WheatCAP and IWYP

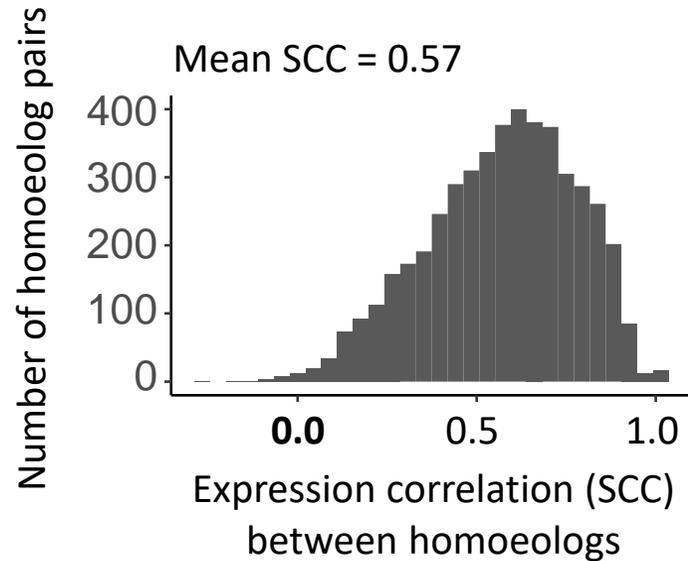
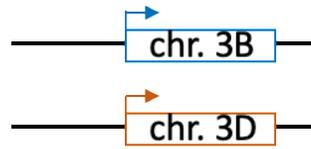


Most wheat accessions carry balanced homoeologs with only small fraction of showing biased (unbalanced) expression

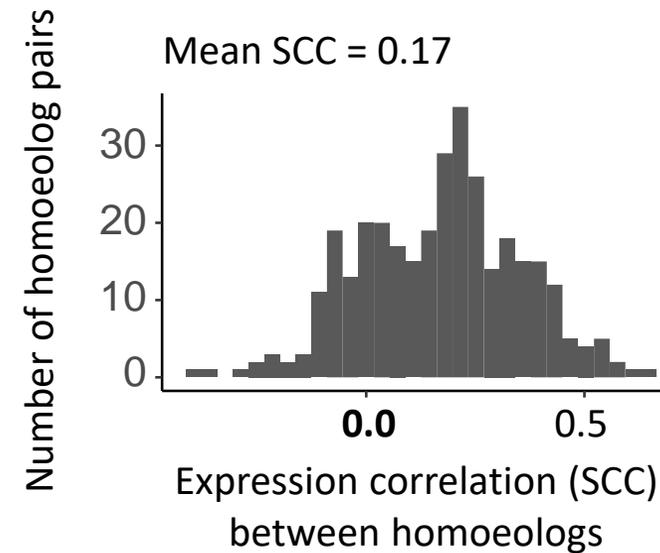
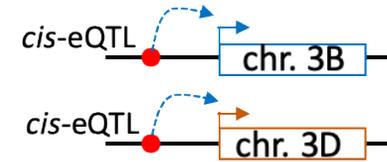


Presence of *cis*-regulatory mutations lead to imbalanced expression of homoeologous genes across wheat accessions

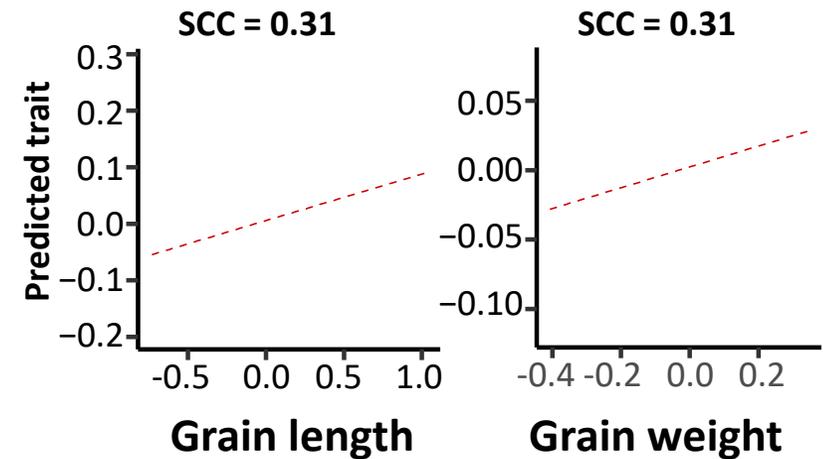
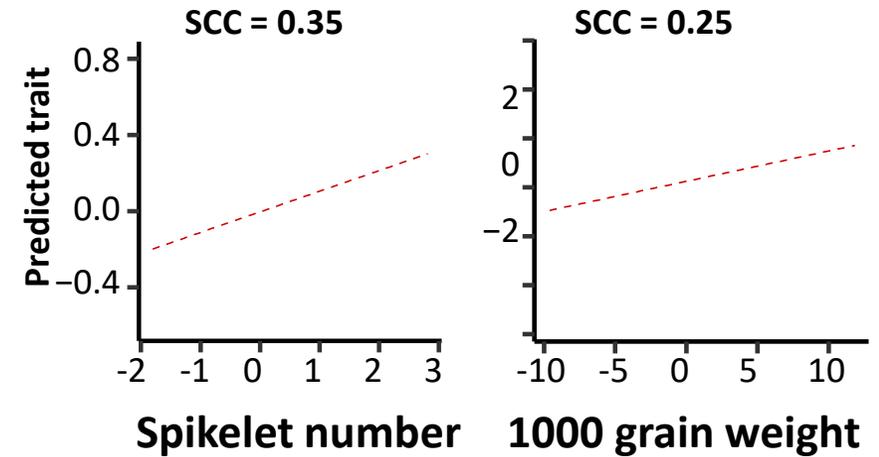
No regulatory mutations



cis-regulatory mutations in both homoeologs lead to biased expression



Trait prediction using ridge regression modeling based on the expression values of biased homoeologs suggest a significant association between the number of dysregulated homoeologs and variation in productivity traits.

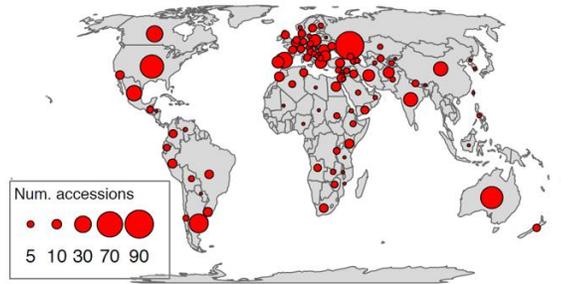


Development of genomic resources for WheatCAP 2022-2026

1. Panel; 200 wheat lines representative of the US wheat breeding germplasm, parents of mapping populations, and wild relatives
2. Characterization of genomic and regulatory diversity by low-pass sequencing
3. Characterization of transcriptomic diversity (Quant-Seq) across 10 tissues at different developmental stages:
 - 1) Seedling root + vegetative at 1 week; 2) Crown with axillary tillers and roots at 4 weeks; 3) Flag leaf sheath + lamina at complete emergence; 4) Root apices at 15 d; 5) Vegetative apex at 3 weeks; 6) Double ridge stage; 7) Terminal spikelet stage; 8) Spikelet at complete gynoecium development; 9) Grain from plants at Zadoks 71 stage (milk development stage); 10) Grain from plants at Zadoks 83 stage (dough development stage).
4. eQTL mapping and identification of variants affecting both gene expression and phenotype

Targeted re-sequencing of wheat germplasm generated a diversity map of the wheat genome

1000 wheat exomes



Exome capture (T3 database)

Natural variation: 1,000 genetically and geographically diverse wheat accessions were re-sequenced using Nimblegen exome capture resulting in discovery > **7 M SNPs**, which are deposited into the T3 database.

337 wheat cultivars representing the genetic diversity of modern US wheat breeding programs were re-sequenced by exome capture (~ **7 M SNPs**).

Induced variation: also used to sequence 1536 4x and 1200 6x EMS mutants and to identify 10,000,000 induced mutations in coding regions. Re-mapped RefSeq v1.0

He at al., 2019, *Nat. Genetics*

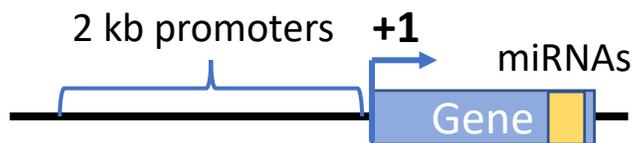
Kuzay et al., 2019, *TAG*

Promoter capture (Gardiner et al, GigaScience, 2019)

Natural variation: Targeted capture of regulatory regions in 250 wheat lines including the Wheat CAP ~**3.2 M SNPs**.

Induced variation: mutations in all promoters of 1,536 tetraploid EMS mutants. Sequenced so far 26% lines: >**4 M mutations**.

Regulatory sequence capture

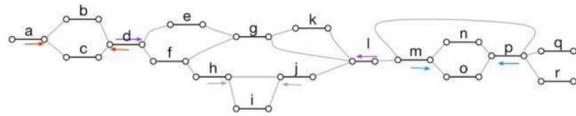


Practical Haplotype Graph (Buckler Lab): a tool for genotyping data storage and genotype imputation



Practical Haplotype Graph

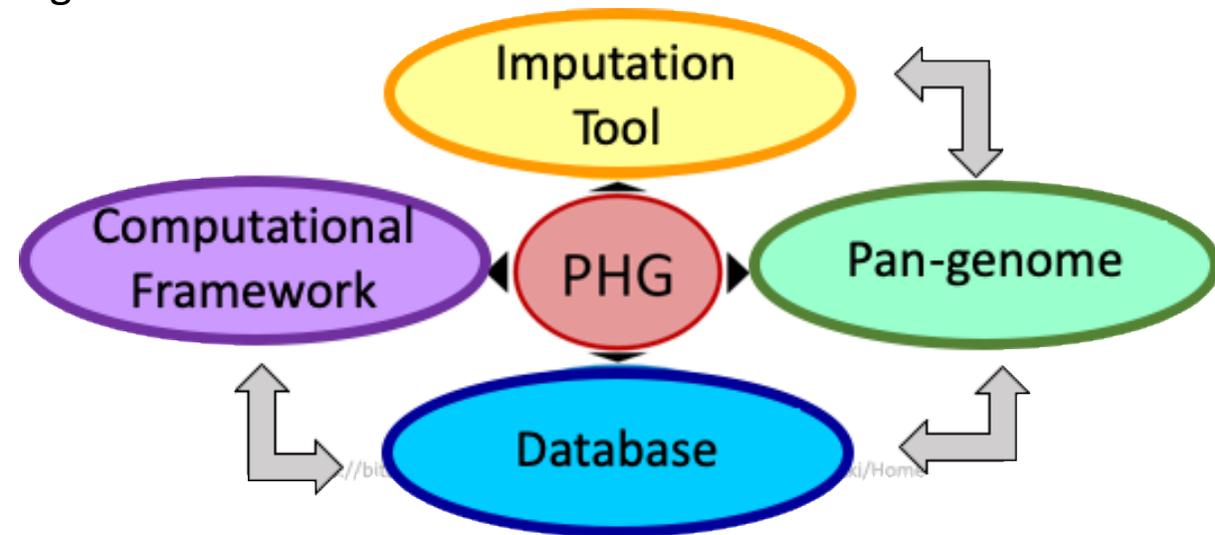
Practical Haplotype Graph



Wheat Practical Haplotype Graph tool: ~400 wheat cultivars representing the genetic diversity of modern US wheat breeding programs were re-sequenced by exome capture. This dataset is used to develop the 1st version of Wheat Practical Haplotype Graph (Jordan et al., 2021, G3). Collaboration between USDA ARS (K. Jordan), USDA ARS Genotyping Labs, T3 database team, and Buckler Lab)

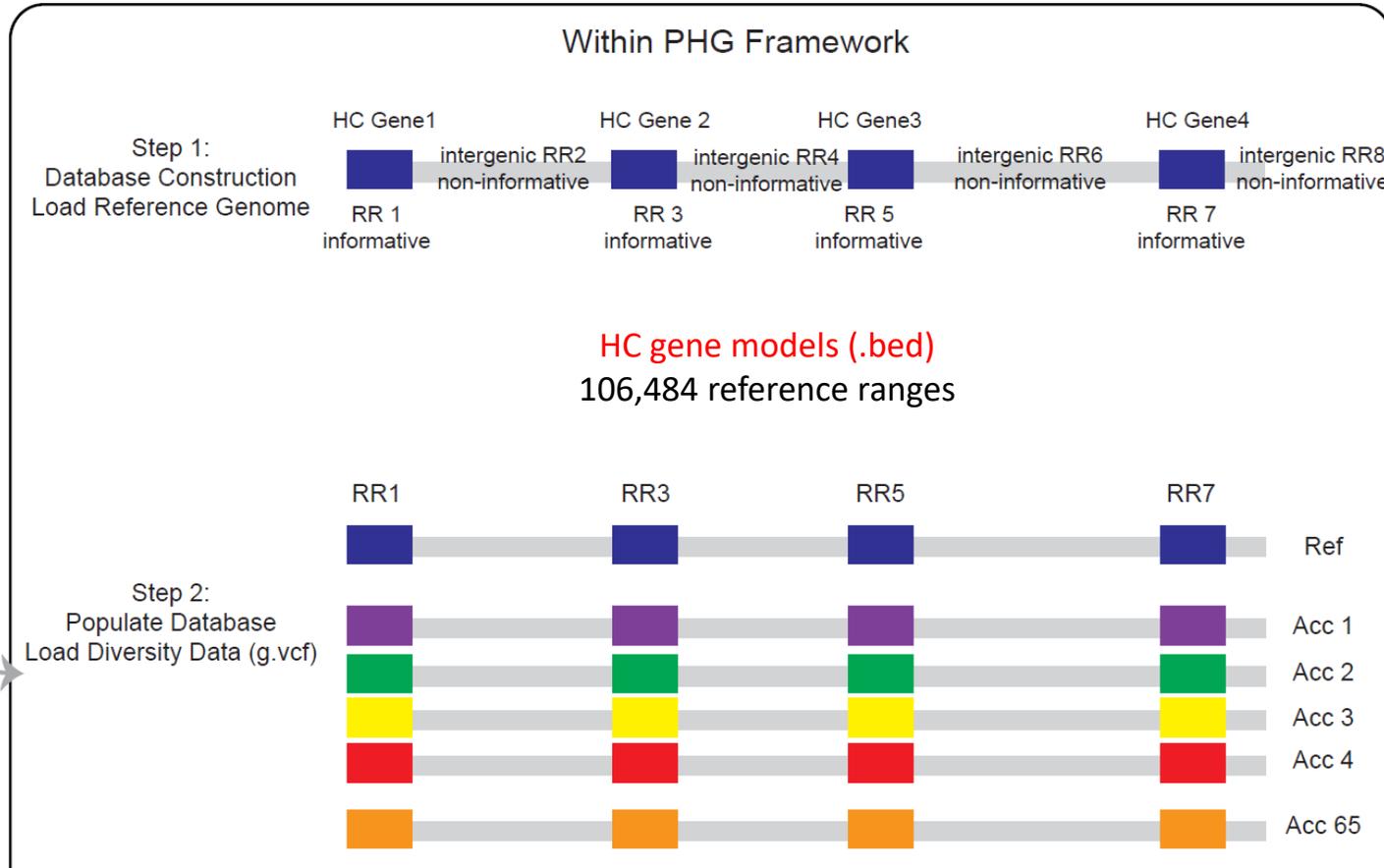
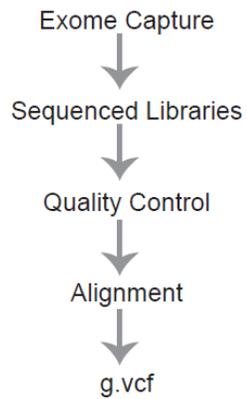
The Practical Haplotype Graph (PHG) Tool

- **Computational Framework (efficient storage and reproducible)**
 - Set of source code configured in Singularity container with all needed bioinformatics software packages
- **Customizable Relational Database**
 - Build customized database with your germplasm
 - Make new database on experiment basis, or add to existing
- **Pan-genome**
 - Reference Genome
 - WGS – representative diversity of input germplasm
 - Capable of storing genome assemblies (SV)
 - More powerful than single reference platform
- **Imputation tool**
 - Generate meaningful data with low sequencing coverage
 - Cost effective with GBS, skim-sequencing, etc...
 - Agnostic platform: Capable of combining different technologies
- **Continuing to improve the capabilities**



PHG: Reference based system (CS RefSeq v1.1)

Outside of PHG :
Generate sequence data
Align to same reference



1: Separate genome
into informative and
noninformative ranges

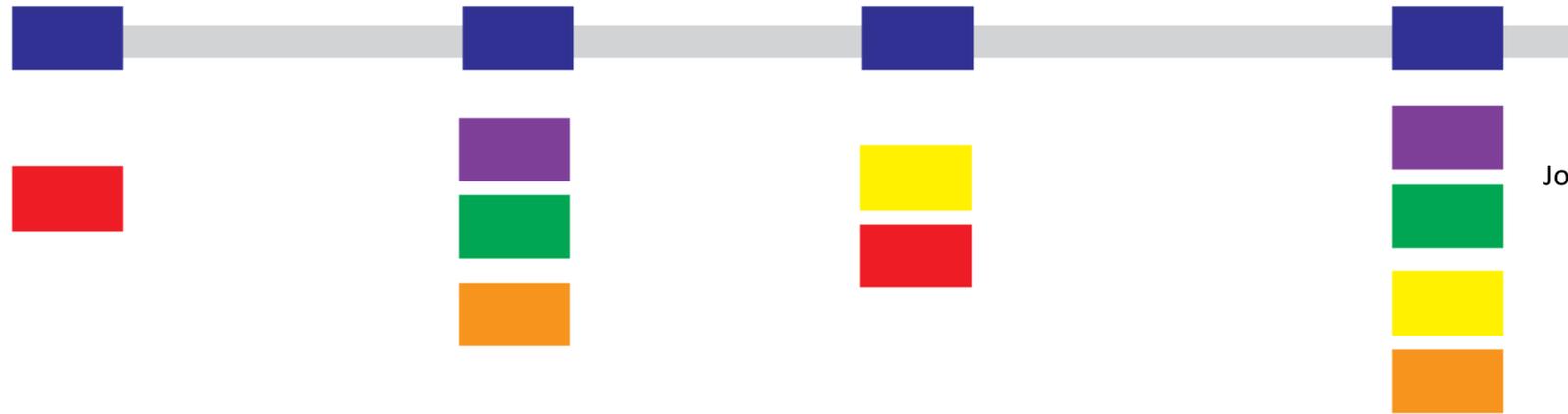
2: Populate the database

Genomes stored
as sequences of
haplotypes
instead of
nucleotides

65 wheat accessions sequenced using exome capture (Jordan et al., 2021, G3)

Create Pan-genome from Diversity Data

Step 3:
Create Consensus
Collapse Haplotypes



Jordan et al, G3, 2021

- Collapse diversity data into consensus haplotypes
- Stores consensus haplotypes sequence by **haplotype ID**
 - Accession information is represented as haplotype IDs in database
 - **Pan-genome represents all diversity in the founding accessions**

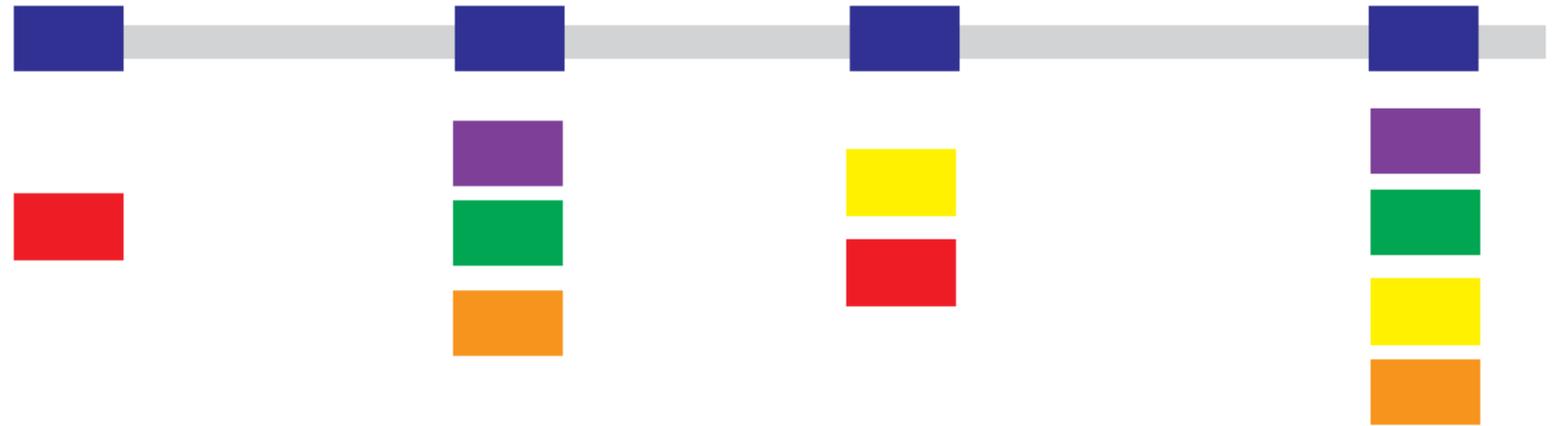
Map low coverage NGS to pan-genome

Exome Capture, GBS,
skimSeq libraries

Quality Control

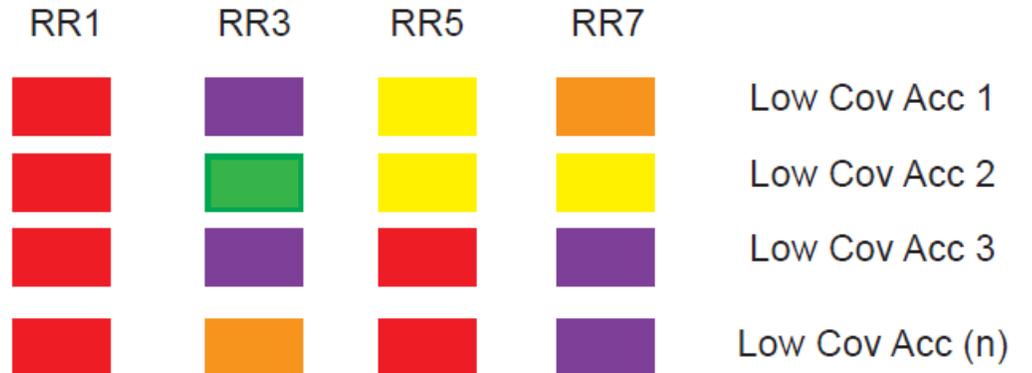
.fq

Step 4:
Impute Skim Data
Find Path through PHG

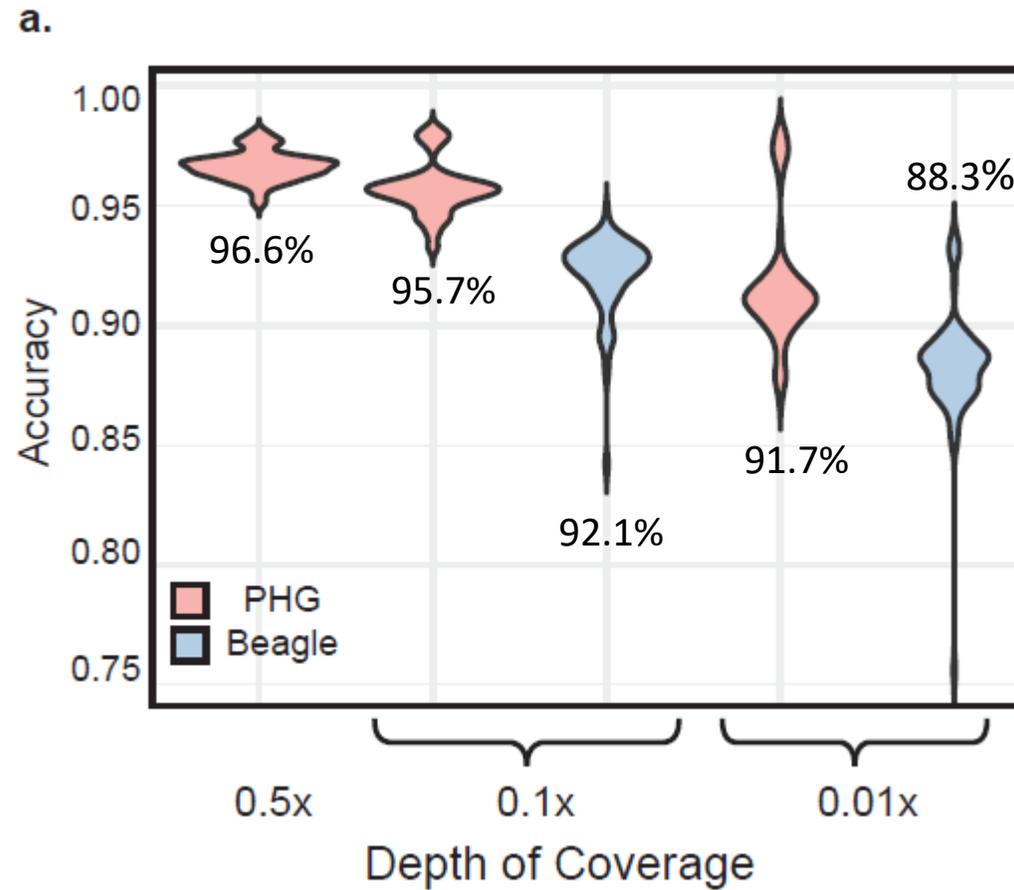


Jordan et al, G3, 2021

- Input GBS, skim seq (fastq)
- Aligns to pangenome haplotypes (minimap2)
- Finds path through the graph (HMM set probability threshold)
- Imputes across missing reference ranges
 - Output: Best path through graph by hap ID



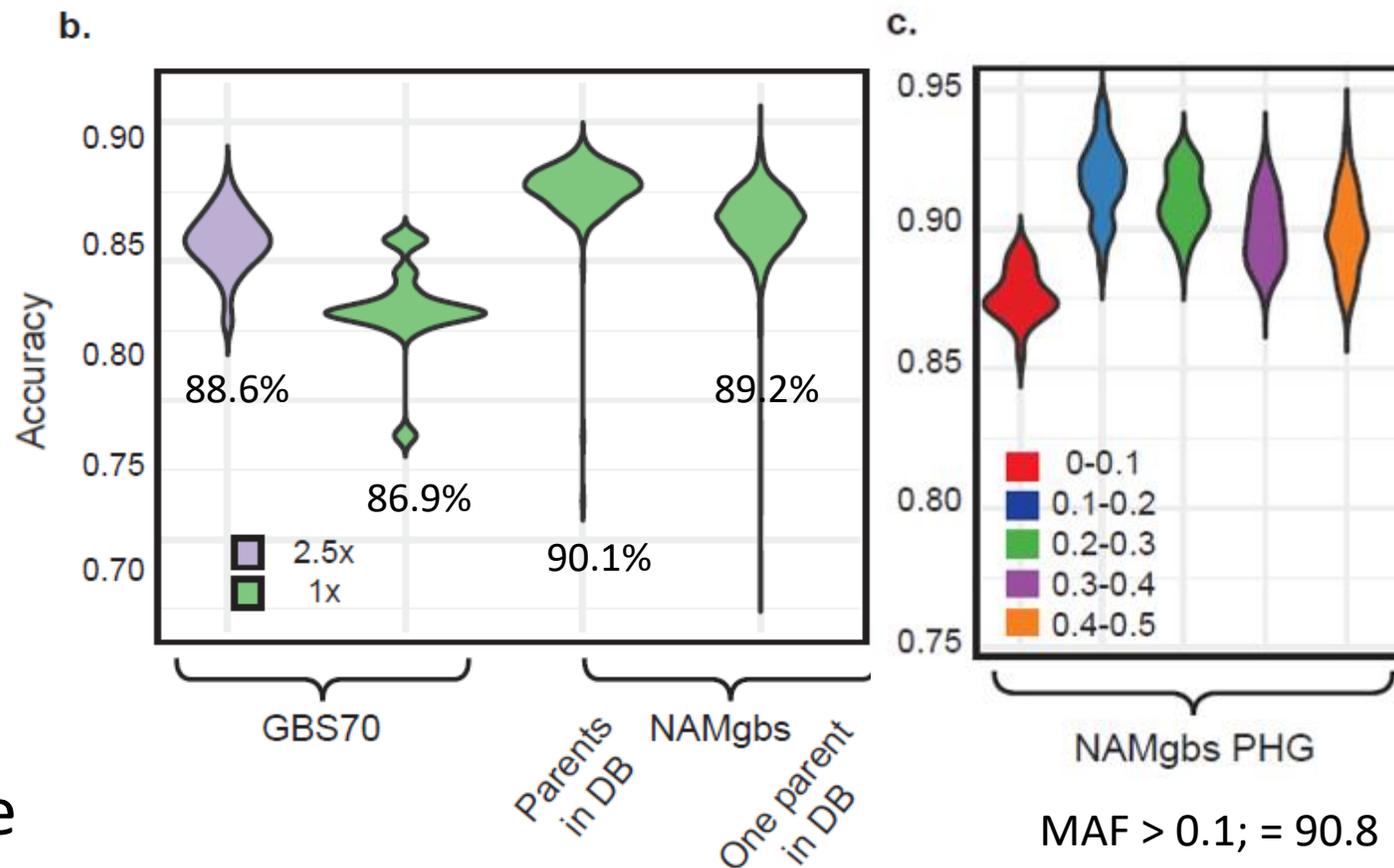
Imputation Accuracy (Wheat PHGv1)



WheatPHGv1 imputed 92% with 0.01x

Imputation Accuracy with GBS $\geq 87\%$

- Used GBS data to test imputation concordance
 - ~5% overlap with exome capture regions
- Concordance improves with representative haplotypes in database
- Concordance improves with more frequent haplotypes in database



Future directions (Wheat PHG v.2)

1. Expand PHG database representation using NGS from 600 wheat accession including lines from all breeding programs
2. Incorporate wheat pan-genome into PHG
3. Incorporate GBS ranges into PHG
4. Incorporate medium density genotyping markers into PHG
5. Integration of PHG into BreedBase (J. Jannink)