Genotyping Tools and Resources: PHG

Katherine Jordan, USDA-ARS, HWWGRU, Manhattan, Kansas

Annual WheatCap Meeting

PAG, San Diego

January 15, 2023

The Practical Haplotype Graph (PHG) Tool

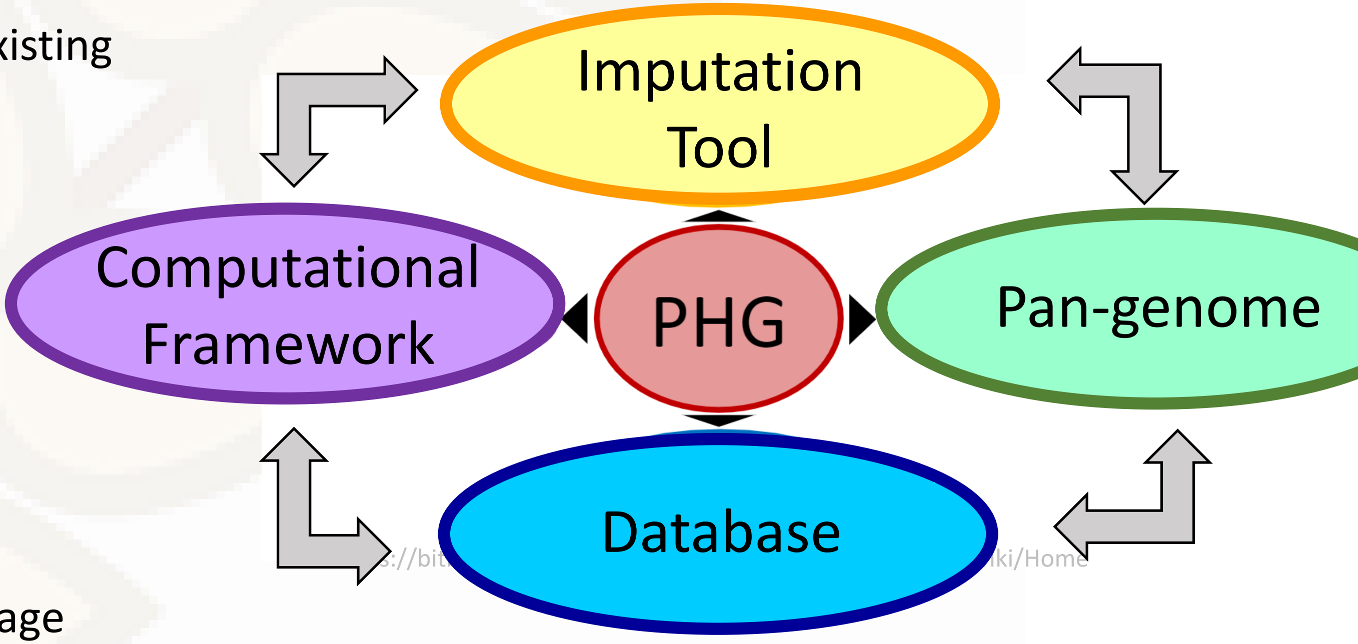
- Computational Framework (efficient storage and reproducible)
  - Source code configured in Singularity container with all needed bioinformatics software packages

- 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
  - Can store 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: Combines different technologies

- Continuing to improve the capabilities



https://bitbucket.org/bucklerlab/practicalhaplotypegraph/wiki/Home

Lessons from WheatPHGv1

Jordan et al, G3, 2021

- Reference Ranges CSv1.1 genes; 65 founding accessions

- Imputation accuracy is best with matched data, 92% with 0.01x
  - Best with matched data (genic ranges/EC data), but >87% with GBS

- Concordance improves with representative haplotypes in database
  - With representative haplotype PHG accurately imputes across alien segment
  - Imputation is 89% accurate with one parent in database with GBS data

- Concordance improves with more frequent haplotypes in database
  - > 90% accurate with MAF > 0.1 (MAF based on database founders)

WheatCap PHG version2; Newer DB version 0.35

- Reference ranges – Coordinates based on genes from RefSeq v2.1

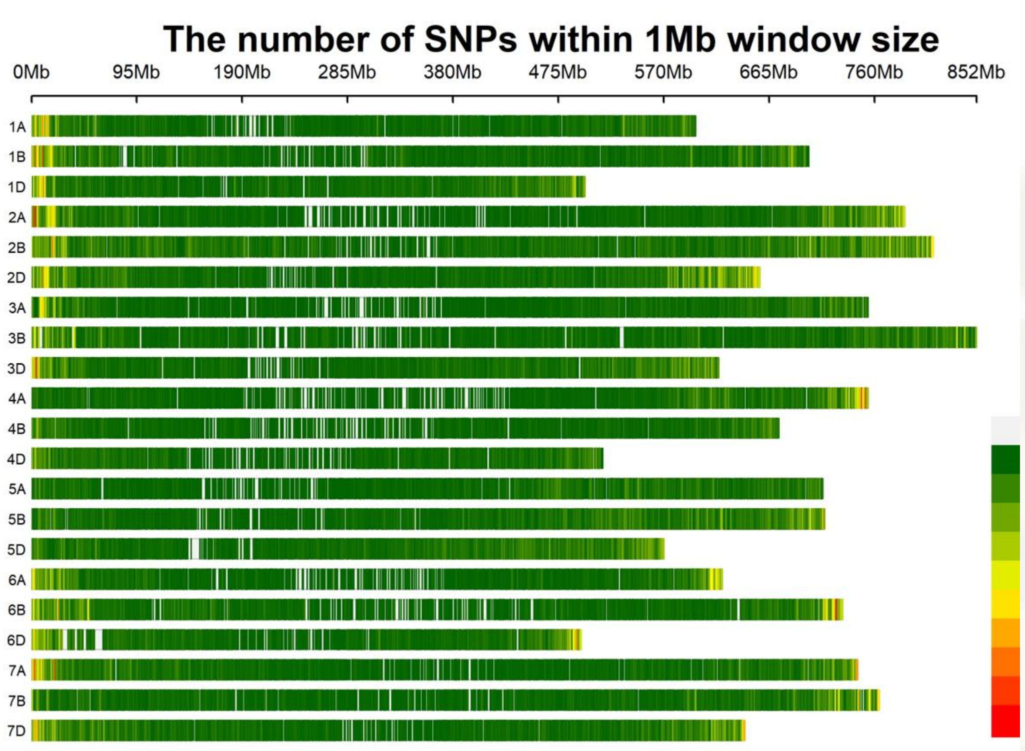
- 472 taxa sequenced using Exome Capture
  - 90 Southern Great Plains
  - 94 Northern Great Plains
  - 95 Southern and Eastern US
  - 193 Pacific Northwest region

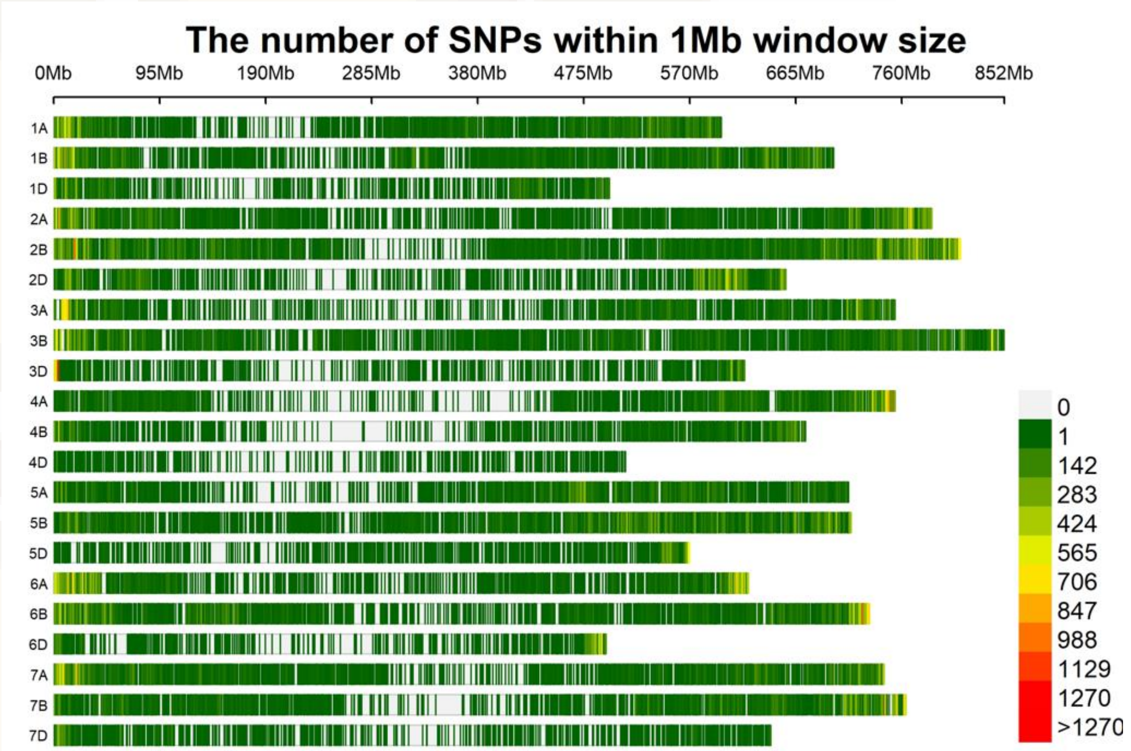
- Database footprint 146Gb

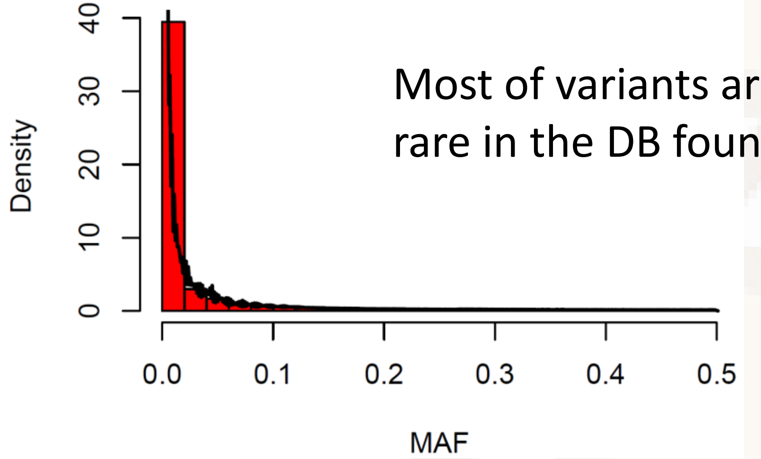
- T3 has access to this database to use for imputation

Market Class	PHG v2
Spring	48
HardRedWinter	59
HardRedSpring	13
SoftRedWinter	39
SoftWinter	42
Winter	35
SoftWhiteWinter	14
HardWhiteWinter	14

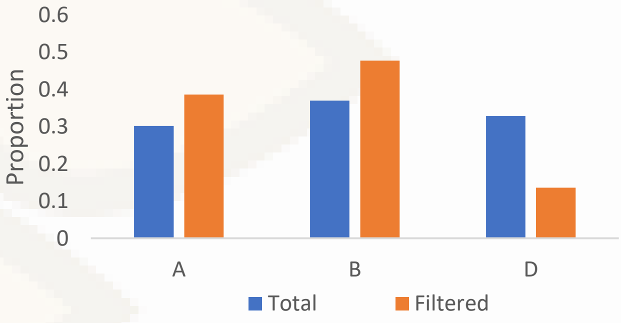
PHGv2 Founders > 5 million segregating variants

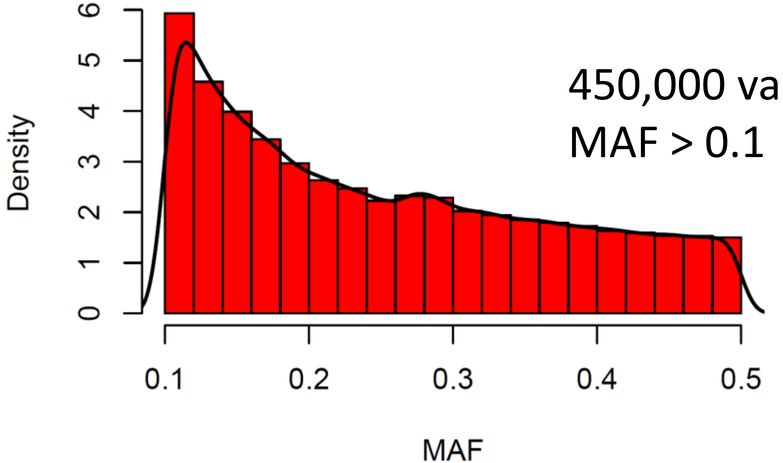






Most of variants are rare in the DB founders



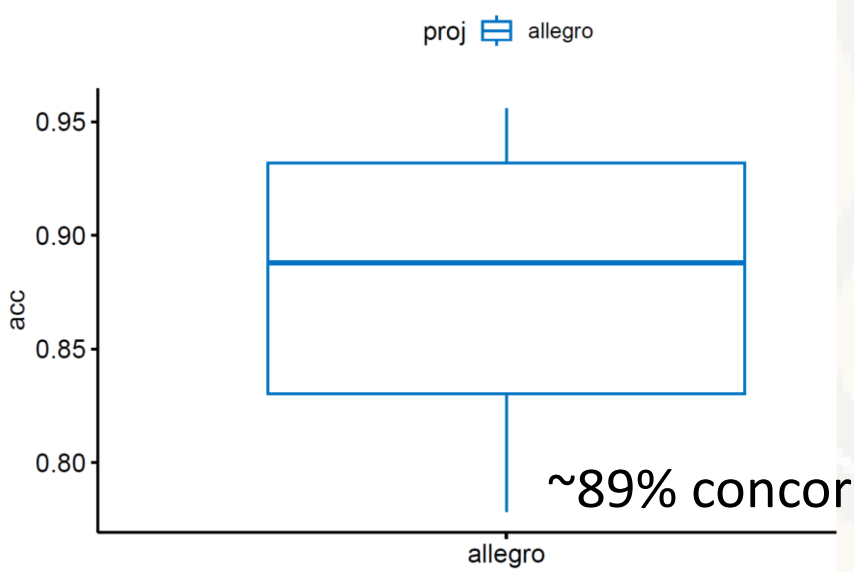


450,000 variants MAF > 0.1

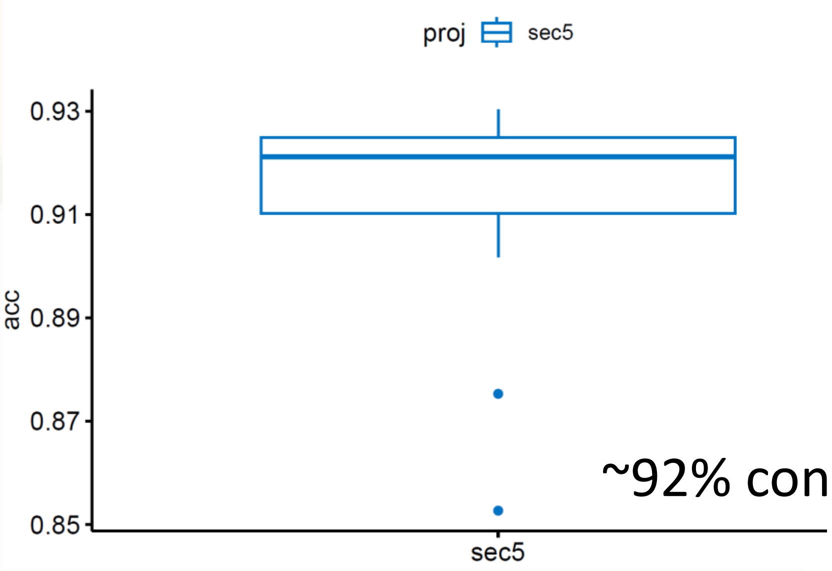
Imputation Test Cases (fastq files)

- Allegro data; 95 SWW lines
  - Wheat Cap database:~400 lines
  - PHGv2 Reference: CSv1.1
  - 106M SE 100bp/taxa = ~0.3x RR cov
  - Compared to Allegro calls (Brian Ward)

- Skim Exome Capture; 12 HWW lines
  - Winter Wheat database: 83 lines
  - PHGv2 Reference: CSv2.1
  - 491,526 PE reads/taxa = ~0.4x RR cov
  - Compared to GATK pipeline ~20x data



~89% concordant



~92% concordant

Imputation Test Cases (vcf files) Clay Birkett

- T3 crew testing new database for imputation from vcf files

Genotype Protocol	PHG founder accession	Not PHG founder
Infinium 90K	94%	79%
Infinium 9K	93%	71%
GMS	89%	
Jason 3K chip (*.fq)	97%	

Protocol	Down sample	Markers	Accuracy
Skim Exome Capture 13 accessions **not in PHG	10	76,147	94%
	30	25,608	94%
	100	7,618	94%
	300	2,865	93%

- More data points = Better imputation accuracies
- Different genotyping methods give different concordance (RR coverage?)

Summary

- PHGv2 with CSv2.1 genome is available for imputation via T3 staff
  - .fastq or .vcf imputation
  - Includes all market classes (more inclusive than PHGv1 - 65 accessions)

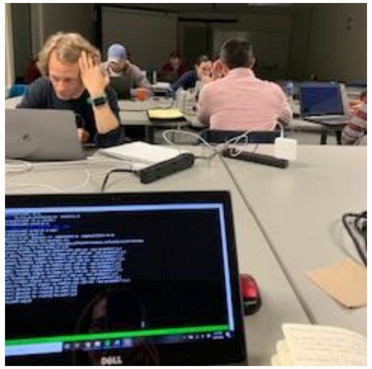
- Imputation accuracies – compared to previously constructed HQ variants
  - PHG founders accuracies better than non-founders
  - Confounded by germplasm 'discrepancies' ?
    - Still not as concordant as PHG founders (consistent with PHGv1 conclusions)

- Room for improvement
  - work in progress – testing imputation parameters

PHG Wheat Group

- Cornell-USDA Buckler group
- Peter Bradbury
- Lynn Johnson
- Terry Casstevens
- Jean Luc Jannink
- Clay Birkett
- David Waring
- Jason Fiedler
- Brian Ward
- Bikash Poudel
- Eduard Akhunov
- Alina Akhunova
- All PHG Hackathon participants

USDA Agricultural Research Service
U.S. DEPARTMENT OF AGRICULTURE



WHEAT CAP
Coordinated Agricultural Project

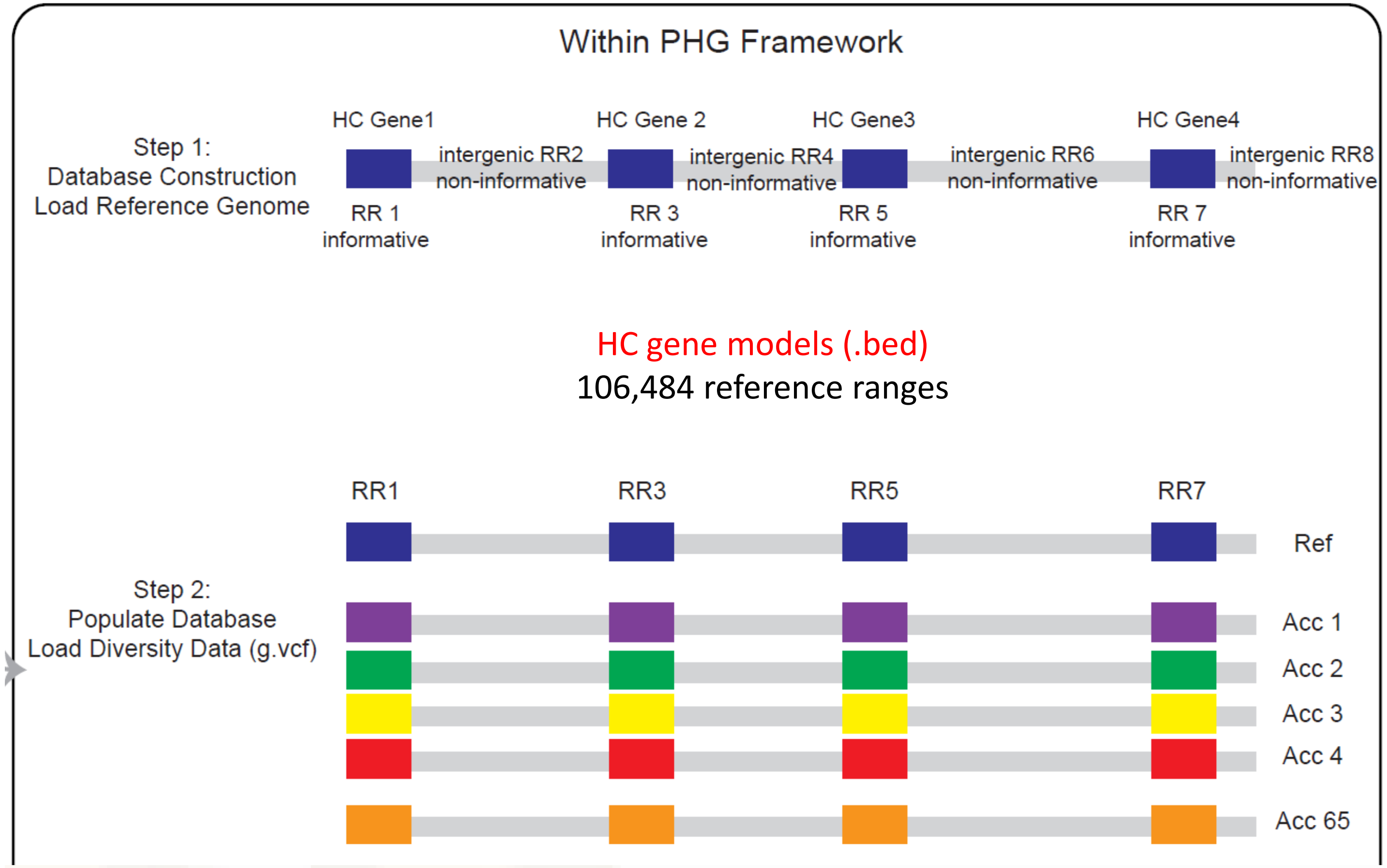
Discussion on Wheat PHG

- Filter the imputed datasets to maximize uses?
  - Test genomic selection models with imputed data
  - ~450,000 markers from WheatCapv2 likely more accurate than rarer variants
    - WheatCap: various mid-density inputs discussed today with T3 imputation
    - Mid-density genotyping platforms appear to impute differently -> (Reference ranges, coverage)

- Continue to test parameters to get better accuracies across reference ranges
  - mxDiv; number of consensus haps, minimap2 sensitivity, etc…
  - How will it handle hets, currently we are ignoring hets
  - Broaden founders? Currently 472 taxa
- Input on parameters, reference ranges, assemblies, one for all or tailored PHGs?

- New PHG version is out: 1.x (September 2022)
  - Output: imputed g.vcf files (likely to combine multiple projects)
  - More computationally efficient but does not currently support wheat chromosome lengths

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



HC gene models (.bed)

106,484 reference ranges

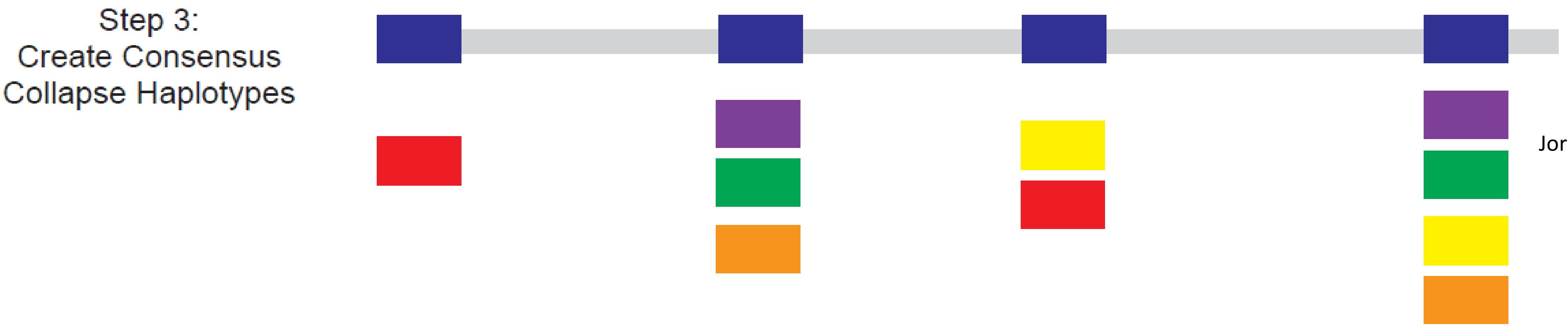
1: Separate genome into informative and noninformative ranges

2: Populate the database

Genomes stored as sequences of haplotypes instead of nucleotides

65 wheat accessions sequenced using Exome capture (Krasileva, et al, PNAS, 2017)

Create Pan-genome from Diversity Data

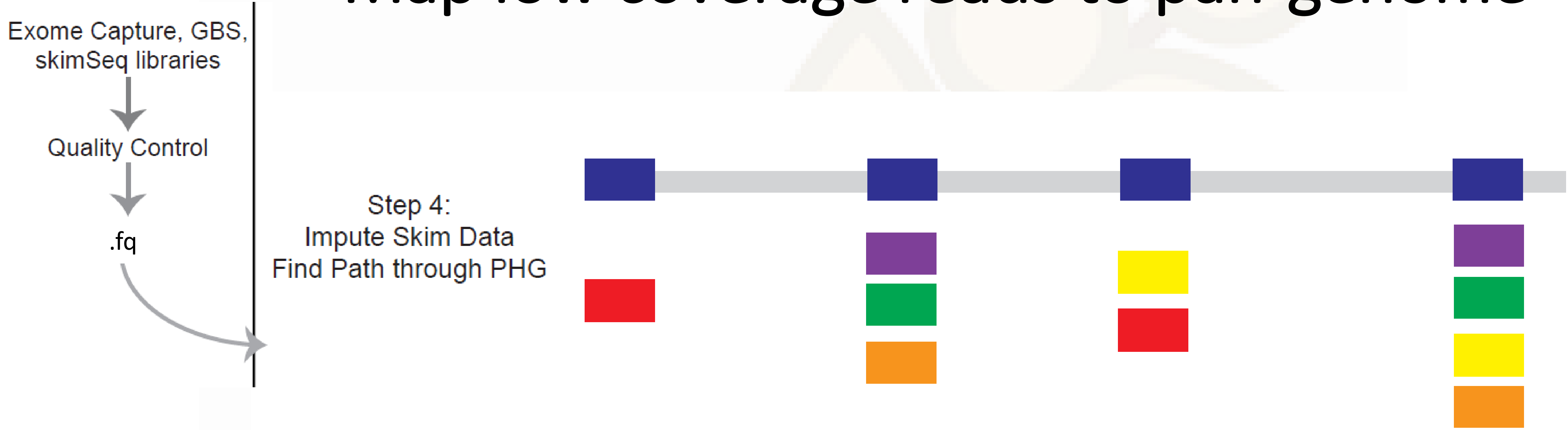


Jordan et al, G3, 2021

- Collapse diversity data into consensus haplotypes
- Parameters in config file that help with haplotype collapsing
  - Diversity (max diversity) & Number Taxa etc… (keep low frequency haplotypes)

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

Map low coverage reads 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

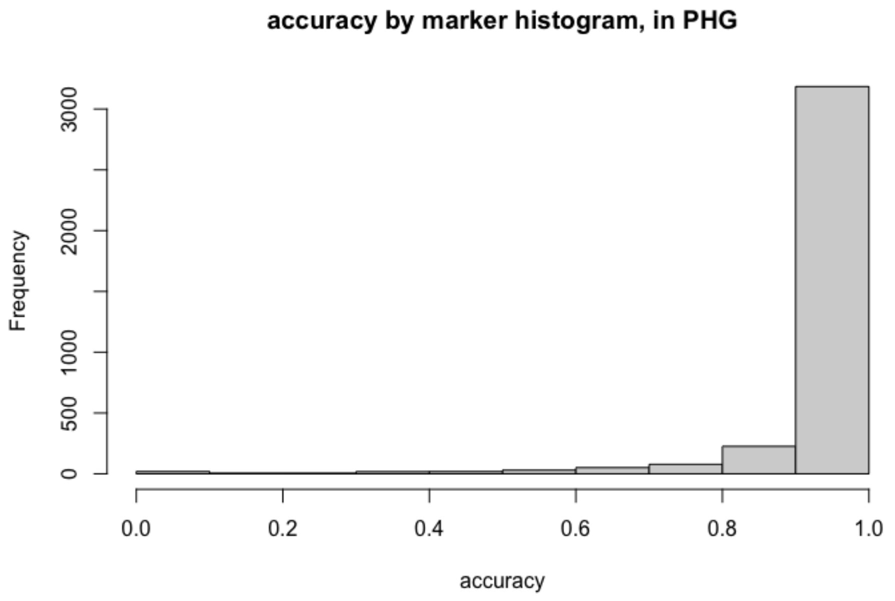
- Input GBS, skim seq (fastq)
- Aligns to pangenome haplotypes (minimap2)
- Finds path through the graph (HMM set probability threshold)
- Imputes across missing reference ranges
  - Output: Best path through graph by hap ID

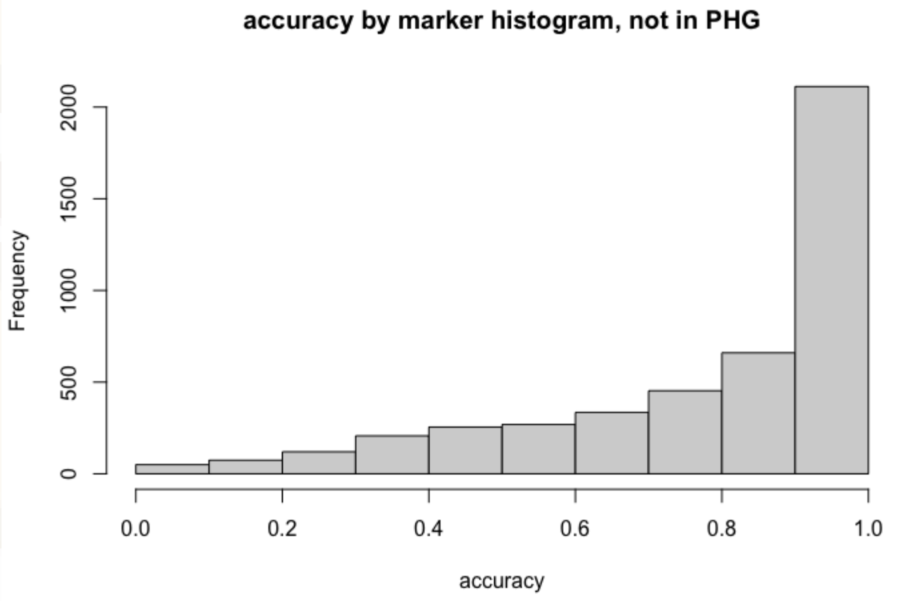
RR1	RR3	RR5	RR7	
🟥	🟪	🟨	🟧	Low Cov Acc 1
🟥	🟩	🟨	🟨	Low Cov Acc 2
🟥	🟪	🟥	🟪	Low Cov Acc 3
🟥	🟧	🟥	🟪	Low Cov Acc (n)

Accuracy of down-sampled data

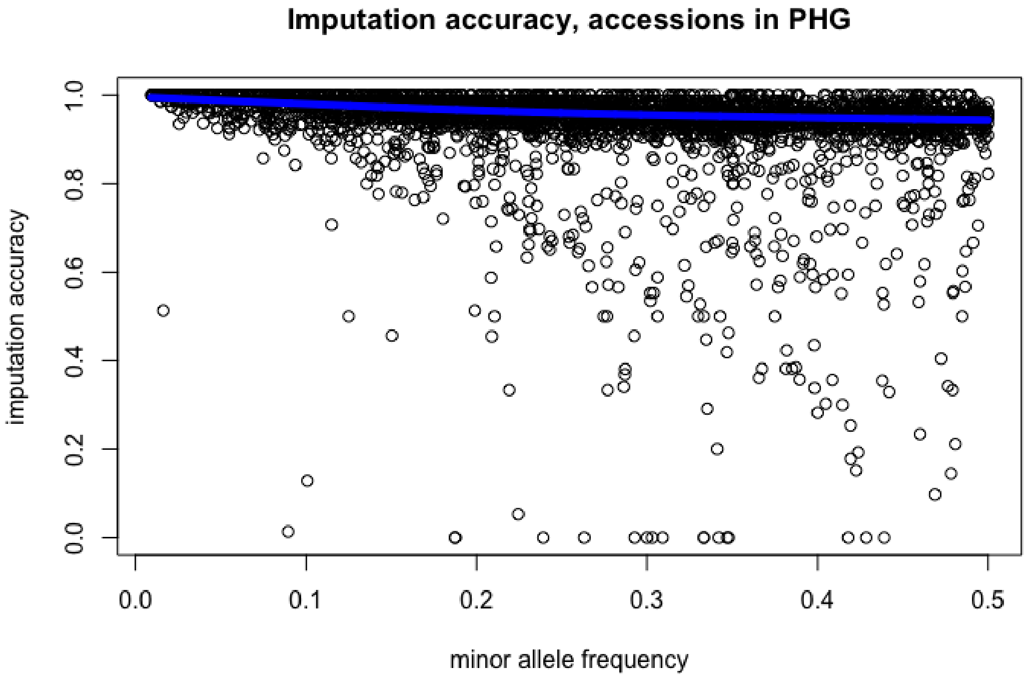
Protocol	Down sample	Markers	Accuracy
Exome Capture 13 accessions	10	76,147	94%
	30	25,608	94%
	100	7,618	94%
	300	2,865	93%
Illumina 90K 79 accessions	1	21,814	93%
	10	2,486	93%
	30	1054	93%
	100	553	87%

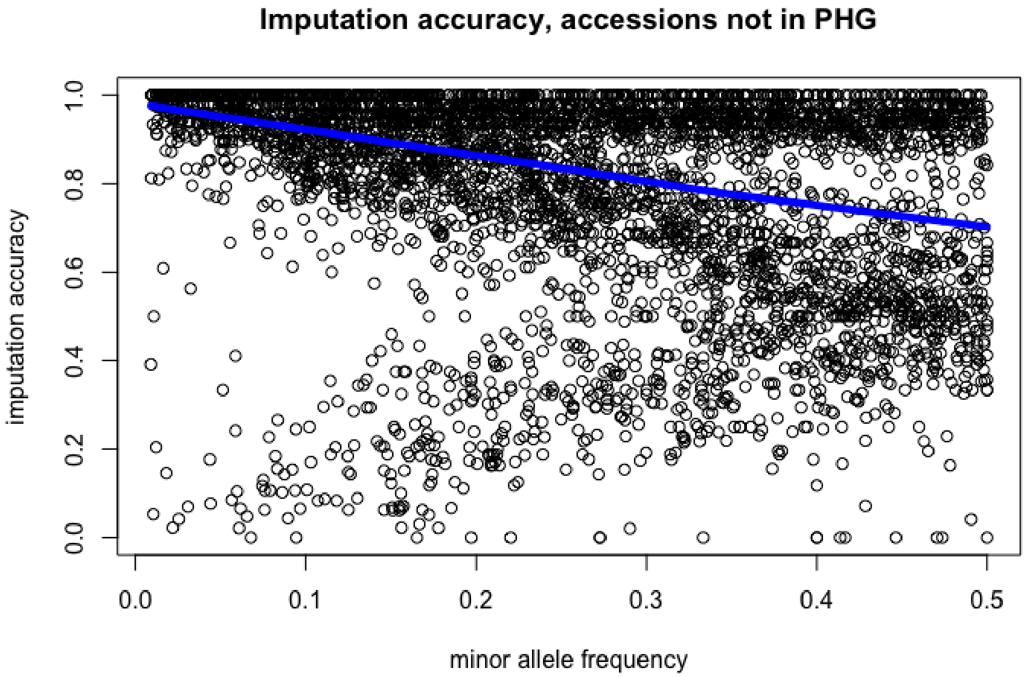
Accuracy by marker, Illumina 90K





Accuracy by minor allele frequency
Illumina 90K


Imputation accuracy, accessions in PHG
imputation accuracy
minor allele frequency


Imputation accuracy, accessions not in PHG
imputation accuracy
minor allele frequency