






















ORIGINAL RESEARCH

A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction

Sarah E. Jensen¹  | Jean Rigaud Charles²  | Kebede Muleta³  |
 Peter J. Bradbury^{4,6}  | Terry Casstevens⁴  | Santosh P. Deshpande⁵  |
 Michael A. Gore¹  | Rajeev Gupta⁵  | Daniel C. Ilut¹  | Lynn Johnson⁴  |
 Roberto Lozano¹  | Zachary Miller⁴  | Punna Ramu⁴  | Abhishek Rathore⁵  |
 M. Cinta Romay⁴  | Hari D. Upadhyaya⁵  | Rajeev K. Varshney⁵  |
 Geoffrey P. Morris³  | Gael Pressoir²  | Edward S. Buckler^{1,4,6}  |
 Guillaume P. Ramstein⁴ 

¹Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

²Chibas and Department of Agriculture and Environmental Sciences, Quisqueya University, Port-au-Prince, Haiti

³Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA

⁴Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA

⁵International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Telangana 502324, India

⁶United States Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA

Correspondence

Sarah Jensen, Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA. Email: sej65@cornell.edu

Funding information

Advanced Research Projects Agency - Energy, Grant/Award Number: DE-AR0000598; Agricultural Research Service; United States Agency for International Development, Grant/Award Number: AID-OAA-LA-16-00003; Bill and Melinda Gates Foundation

Abstract

Successful management and utilization of increasingly large genomic datasets is essential for breeding programs to accelerate cultivar development. To help with this, we developed a *Sorghum bicolor* Practical Haplotype Graph (PHG) pangenome database that stores haplotypes and variant information. We developed two PHGs in sorghum that were used to identify genome-wide variants for 24 founders of the Chibas sorghum breeding program from 0.01x sequence coverage. The PHG called single nucleotide polymorphisms (SNPs) with 5.9% error at 0.01x coverage—only 3% higher than PHG error when calling SNPs from 8x coverage sequence. Additionally, 207 progenies from the Chibas genomic selection (GS) training population were sequenced and processed through the PHG. Missing genotypes were imputed from PHG parental haplotypes and used for genomic prediction. Mean prediction accuracies with PHG SNP calls range from .57–.73 and are similar to prediction accuracies obtained with genotyping-by-sequencing or targeted amplicon sequencing (rhAmpSeq) markers. This study demonstrates the use of a sorghum PHG to

Abbreviations: BLAST, Basic Local Alignment Search Tool; CDS, coding sequence; GBS, genotyping-by-sequencing; GS, genomic selection; gVCF, genomic variant call format; HMM, Hidden Markov Model; PHG, Practical Haplotype Graph; PIC, polymorphic information content; SNP, single nucleotide polymorphism; VCF, variant call format; WGS, whole genome sequence.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *The Plant Genome* published by Wiley Periodicals, Inc. on behalf of Crop Science Society of America

impute SNPs from low-coverage sequence data and shows that the PHG can unify genotype calls across multiple sequencing platforms. By reducing input sequence requirements, the PHG can decrease the cost of genotyping, make GS more feasible, and facilitate larger breeding populations. Our results demonstrate that the PHG is a useful research and breeding tool that maintains variant information from a diverse group of taxa, stores sequence data in a condensed but readily accessible format, unifies genotypes across genotyping platforms, and provides a cost-effective option for genomic selection.

1 | INTRODUCTION

The goal of plant breeding is to develop improved cultivars with high yield potential and better-quality traits while reducing input requirements and environmental impact. Over time, plant breeding techniques have shifted from unconscious selection for domestication traits to deliberate selection and breeding program design. As the field of genetics developed, plant breeders began to use marker-assisted selection to associate genetic markers with desirable traits and inform breeding decisions (reviewed in Ramstein, Jensen, & Buckler, 2018). First proposed by Meuwissen, Hayes, and Goddard (2001), genomic selection (GS) is an extension of marker-assisted selection that uses genome-wide markers to make predictions about individual performance. Many studies have shown that GS can accelerate the breeding process and rate of genetic gain without significantly increasing program costs (e.g., Bernardo & Yu, 2007; Heffner, Lorenz, Jannink, & Sorrells, 2010; Heslot, Jannink, & Sorrells, 2015; Meuwissen et al., 2001; Muleta, Pressoir, & Morris, 2019; Poland et al., 2012). Increasingly dense marker or haplotype maps in major crops like maize (*Zea mays*; Bukowski et al., 2018) and sorghum (*Sorghum bicolor* (L.) Moench; Lozano et al., 2019) can now be leveraged to inform breeding decisions.

Despite its apparent utility, wide-scale adoption of GS has been slow, particularly in developing countries where labor is less expensive and labor-intensive phenotypic selection is therefore cheaper than GS for many traits (Ribaut, de Vicente, & Delannay, 2010). Genomic selection costs depend on multiple components, including sample collection, processing, and sequencing costs, as well as bioinformatics requirements. Wider adoption of GS will require fast genotyping methods that are cheaper than phenotypic selection for all or most traits of interest. One way to make genotyping cheaper is to capitalize on pre-existing genomic data. Pan-genomic analyses capture a wider range of genetic variation and are becoming more common as genotyping costs decrease and the quantity of sequence data increases (e.g., Alonso-Blanco et al., 2016; Morris et al., 2013; Shakoor et al., 2014). Tools that can extrapolate from existing whole-genome sequences

(WGS) or other marker technologies and impute genotypes for individuals in a breeding population can utilize pre-existing WGS to minimize the amount of new data that breeders need to generate.

Here we introduce a method for genotyping breeding populations from low-coverage sequence data called the Practical Haplotype Graph (PHG). The PHG uses a graph of haplotypes to represent the variability in a breeding program and can merge genotypes from WGS and marker technologies. It takes a pan-genome approach to marker identification and relies on having a limited number of recombination events in each generation (Bouchet et al., 2017; Mace et al., 2009). In a breeding population, all major haplotypes in the progeny population are contained within a set of founder haplotypes. Moderate recombination shuffles founder haplotypes in subsequent generations, but progeny genotypes can be reconstructed if the approximate recombination positions can be identified. The PHG is designed to reconstruct haplotypes for a breeding population based on low-coverage random sequence, which reduces the quantity of new sequence data that needs to be collected for each cycle of GS.

In building a PHG database, moderate- to high-coverage WGS reads are aligned to the reference genome (Figure 1a) and collapsed into common, consensus haplotypes based on sequence divergence (Figure 1b, 1c). These consensus haplotypes are important for the PHG because they reduce the complexity of the graph, make it possible to associate traits with shared haplotypes, and fill in (impute) missing sequence information, while maintaining unique haplotypes. Once created, consensus haplotypes can be used to predict genotypes for new individuals. Skim sequence data are aligned to consensus haplotypes to find the best path through the graph, and single nucleotide polymorphism (SNP) variants from the predicted haplotype path can be written to a variant call format (VCF) file. The result is a set of genome-wide SNP variant calls for each taxon, imputed from skim sequence (Figure 1d).

We tested the efficiency and accuracy of the PHG in a sorghum breeding program at Chibas in Port au Prince, Haiti. The climate in Haiti offers year-round growing conditions with three growing seasons annually. Currently, the primary

sorghum breeding program at Chibas uses phenotypic recurrent selection and takes about a year per selection cycle. The main targets of selection include high yield, stalk sugar content, and sugarcane aphid resistance. Genomic selection has the potential to speed up the breeding cycle so that selections can be made up to three times per year, but is only useful if the potential increased genetic gain per cycle outweighs additional genotyping costs (Muleta et al., 2019). We tested a PHG built with 24 founders of the Chibas breeding program (referred to here as the “founder PHG”), and another built with the 24 Chibas founders plus an additional 374 taxa that are representative of overall sorghum diversity (referred to as the “diversity PHG”). With the Chibas breeding program as our test case, we evaluated whether the PHG can impute from low-coverage sequence, how PHG-imputed markers compare to other genotyping methods, and how much background genomic information is needed for a breeding program. Our results show that the PHG is an effective tool to impute genotypes for GS from random skim sequence.

2 | MATERIALS AND METHODS

2.1 | Phenotypic data

Phenotypes for height, brix, juice weight, leaf weight, earliness, stem weight, and grain yield were collected at the Regional Biofuels Technical and Knowledge Center field sites in Port au Prince, Haiti. The Chibas training population (250 taxa) was assayed in five experimental conditions (experiments) in Port au Prince, Haiti. Experiments differed by sowing date: Experiment 1, 7 Sept. 2017; Experiment 2, 7–10 Nov. 2017; Experiment 3, 8–10 Nov. 2017 (irrigated), Experiment 4, 3 Feb. 2018; and Experiment 5, 16 Feb. 2018. All experiments were rainfed except for Experiment 3. Taxa were evaluated under randomized complete block designs with four replicates in each experiment, except for Experiments 4 and 5, which had two replicates each. For each trait, genotype (taxon) means were estimated as fixed effects under a linear mixed model with independent and identically distributed effects for experiments, interactions between genotypes and experiments, and replicates within experiments. Linear mixed models were fit with the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015).

2.2 | Sample collection and DNA extraction

Strips of mature leaf tissue measuring 4 by 0.5 cm were collected from 1-mo-old sorghum plants growing at the Chibas field site in Port au Prince, Haiti. Fresh tissue samples were placed in 96-well costar tube plates and lyophilized for 48 h before being shipped to Intertek for DNA extraction. Samples

Core Ideas

- A pan-genome Practical Haplotype Graph captures key haplotypes in a breeding program.
- The PHG accurately imputes whole-genome variants from only 0.01x sequence coverage.
- The PHG can unify genotypes from many types of input sequence data.

were packed on silica beads and shipped at room temperature to the Intertek facility in Alnarp, Sweden, and DNA was extracted using their metal bead DNA extraction protocol.

2.3 | Library preparation and sequencing

The DNA samples from 24 population founders were used to make TruSeq Nextera sequencing libraries at the Genomics facility at Cornell University. Samples from all 24 founders were pooled and sequenced in a single lane of 2 by 150 bp reads on an Illumina NextSeq500 instrument resulting in an average of 8x coverage per individual. The Chibas training population consists of 238 individuals. Samples in the training set were pooled in a single lane with 2,736 other individuals and sequenced at 2 by 150 bp reads on an Illumina NextSeq500 instrument, resulting in approximately 0.1x coverage for each individual. Genotyping-by-sequencing (GBS) data for comparison with PHG genotypes were from Muleta et al. (unpublished data, 2019).

2.4 | Building the sorghum PHG

A sorghum practical haplotype graph was built using scripts in the `p_sorghumphg` bitbucket repository and PHG version 0.0.9. Instructions for building a new PHG can be found on the PHG Wiki, available on Bitbucket at <https://bitbucket.org/bucklerlab/practicalhaplotypegraph/wiki/Home> (Figure 2).

2.4.1 | Creating and loading reference ranges

Reference ranges for the PHG were chosen based on conserved gene annotations. Conserved coding sequences (CDS) were selected as likely functional genomic regions where reads are easier to map unambiguously. Coding sequences from the sorghum version 3.1 genome annotations and the version 3.0 reference genome were downloaded from the Joint Genome Institute and compared to a Basic Local Alignment

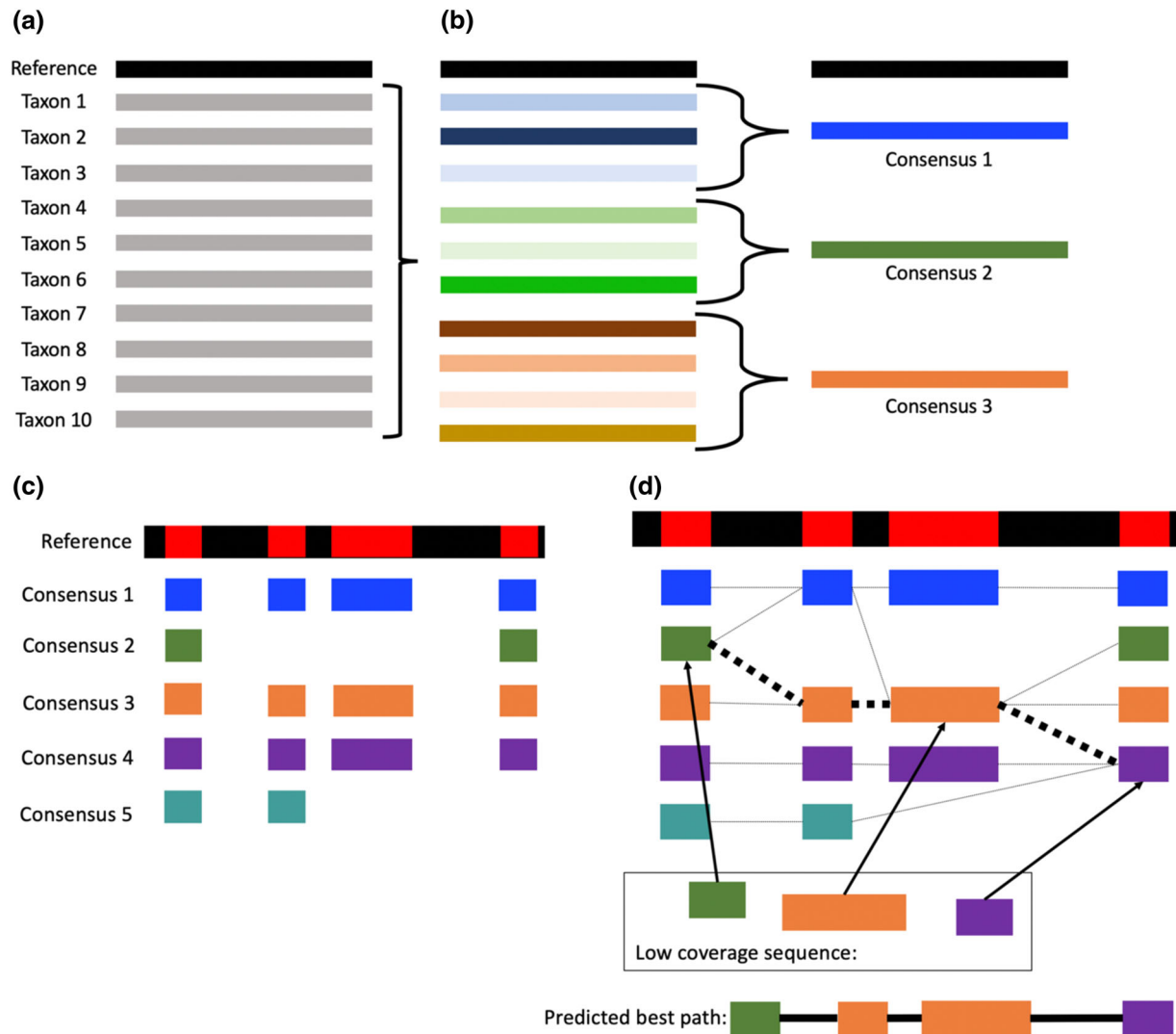


FIGURE 1 Practical haplotype group (PHG) database and haplotype creation. (a) Whole genome sequence data are aligned to a reference genome and loaded into the PHG database. (b) A set of designated reference ranges is chosen, and input data are condensed to produce consensus haplotypes at each reference range. Colors are used to indicate sequence similarity across taxa, and only a single reference range is shown. In this range, 10 genotypes are condensed into three major consensus haplotypes, with minor differences within a consensus haplotype maintained as variant sites. (c) Unique consensus haplotypes are built for each reference range across the genome. Reference ranges are indicated as red regions of the black reference genome bar. (d) Low-coverage sequence data are aligned to the consensus haplotypes and a hidden Markov model (HMM) links reference ranges across the genome to predict genome-wide haplotypes (black edges reflect HMM edge weights between haplotypes, with the most likely haplotype based on the low coverage sequence data highlighted as a bold, dashed line)

Search Tool (BLAST) database containing CDS for *Zea mays*, *Setaria italica*, *Brachypodium distachyon*, and *Oryza sativa* (Bennetzen et al., 2012; Ouyang et al., 2007; Schnable et al., 2009; Vogel et al., 2010) that was created using BLAST+ command line tools (Altschul et al., 1997). The sorghum version 3.1 CDS annotations and version 3.0 reference genome (McCormick et al., 2017) were compared to the four-species database with blastn default parameters. These species were used because they have high-quality genome assemblies and annotations and cover a diverse set of grasses. Sorghum gene intervals were kept if there was at least one hit to the four-species database, and gene start and end coordinates were

used to create initial reference intervals. Initial gene intervals were expanded by 1,000 bp on either side of the gene coordinates, and intervals within 500 bp of each other were merged to form a single reference range. The resulting dataset contains 19,539 intervals spaced across the genome, which we designated “genic reference ranges,” while the intervals between genic reference ranges were added to the database as 19,548 “intergenic reference ranges.” The LoadGenomeIntervals pipeline was used to add reference genome sequence to the database for both genic and intergenic ranges, whereas sequence data from additional taxa were added only to the genic reference ranges.

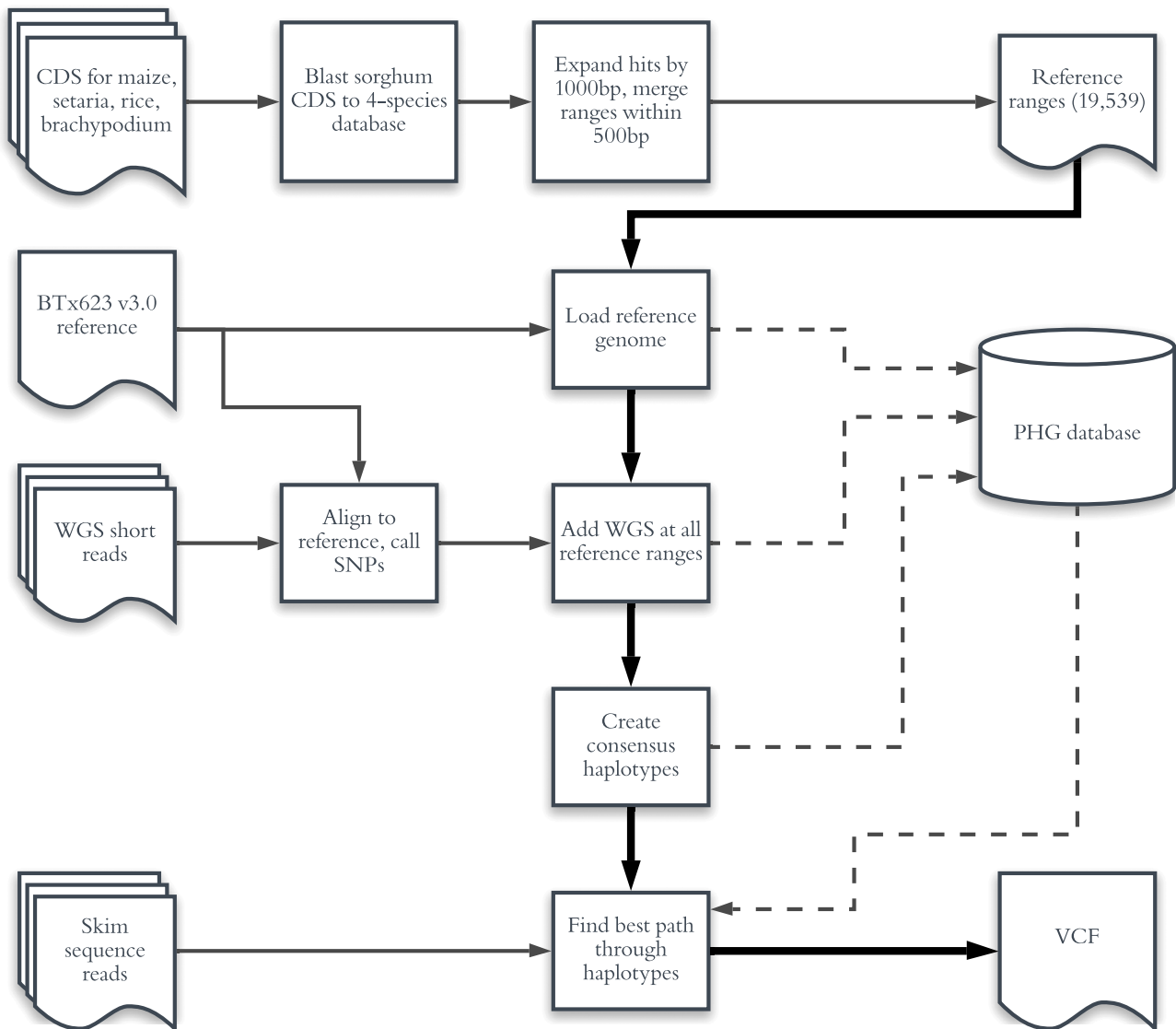


FIGURE 2 Steps in building and using the sorghum practical haplotype group (PHG). The main PHG processing steps are indicated with bold arrows. Steps that store or read information from the PHG database are indicated with dotted lines. CDS, coding sequence; WGS, whole genome sampling; SNP, single nucleotide polymorphism; VCF, variant call format

2.4.2 | Adding haplotypes from diverse taxa and creating consensus haplotypes

Sequence data were aligned to the version 3.0 sorghum BTx623 reference genome with BWA MEM (Li & Durbin, 2009; McCormick et al., 2017). Taxa in the PHG are as follows: 24 founder individuals from the Chibas sorghum breeding program, 274 previously-published taxa (42 from Mace et al., 2013; 232 from Valluru et al., 2019), and 100 taxa from the ICRISAT mini-core collection, for a total of 398 taxa. No de novo genome assemblies are included. Variants were called with Sentieon's HaplotypeCaller pipeline (Sentieon DNaseq, 2018) and the resulting genomic VCF (gVCF) files were added to the PHG using the CreateHaplotypesFromGVCF pipeline. The Sentieon pipeline was chosen for computational

efficiency. As an alternative, the Genome Analysis Toolkit (GATK) HaplotypeCaller pipeline offers a similar, but slower, open-source pipeline. The same process was used to make a smaller PHG database with only the 24 founder individuals from the Chibas breeding program.

The 398 or 24 individuals added to the PHG were condensed into common "consensus haplotypes" using the CreateConsensus pipeline. Taxa were clustered with the unweighted pair group method with arithmetic mean clustering method in Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL; Bradbury et al., 2007) at 10 possible divergence levels: .001, .00075, .0005, .00025, .0001, .000075, .00005, .000025, .00001, and 0. Unique divergent haplotypes were retained as consensus haplotypes ($\text{minTaxa} = 1$), and all other parameters were left at their

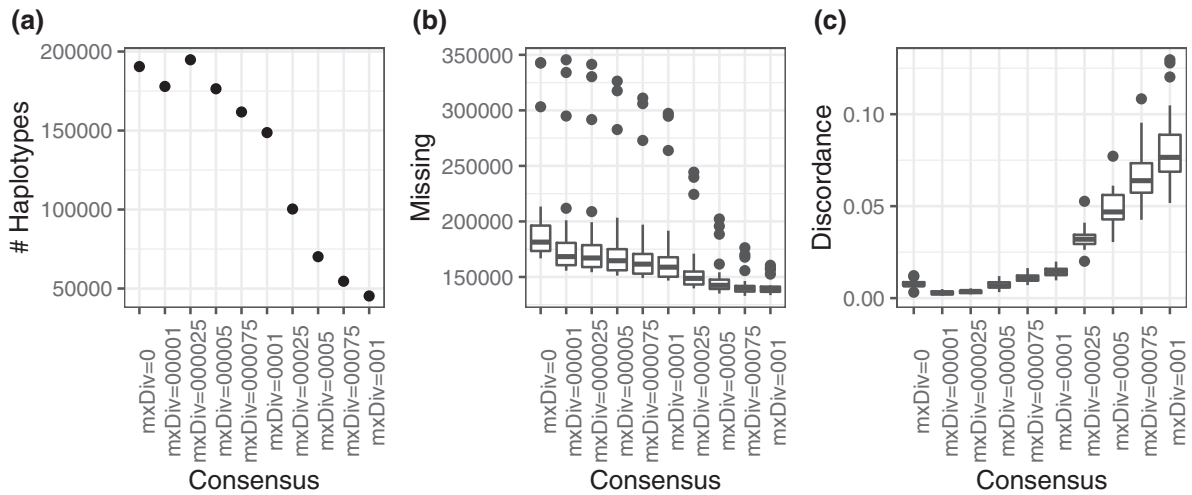


FIGURE 3 Metrics used to choose a consensus level for condensing haplotypes in the founder practical haplotype group (PHG). (a) Number of haplotypes at each consensus level. (b) Box plot summarizing missing counts at each consensus level. The number of missing base pairs for each individual across all reference ranges was counted at each level of haplotype collapse to determine how collapsing haplotypes reduces missingness in the sequence data. (c) Box plot of discordance between single nucleotide polymorphisms (SNPs) called by the PHG and a set high-confidence whole genome sequence SNPs. The PHG SNPs were compared to WGS data and the proportion of SNPs that did not match between PHG and WGS calls was recorded for each individual. Ultimately, a divergence level of $\text{mxDiv} = .00025$ was selected as the level of haplotype collapse that balances the tradeoff between amount of missing information in the database and discordance from condensing haplotypes too much

default values. The Chibas founder taxa were used as a baseline to select the best level of haplotype collapse for this dataset, and a consensus level of .00025 (1 in 4,000 bp differences) was chosen. This level balances SNP count and concordance with other sequencing methods, while minimizing the number of haplotypes at each reference range (Figure 3). Subsequent genotyping and imputation analyses were run using this consensus level.

2.4.3 | Predicting new genotypes

After creating consensus haplotypes, the findPaths pipeline maps reads to the consensus pangenome and uses a hidden Markov model (HMM) and the Viterbi algorithm (Rabiner, 1989) to predict the best path through all consensus haplotypes in the graph. In each analysis we subset the pangenome graph to only the consensus haplotypes containing one or more of the 24 taxa that founded the Chibas breeding program. Paired-end, 150-bp reads with varied sequence coverage from random sequence data were used as input to predict the best path through the graph, both for individuals in the database and for new taxa that are not already represented. Paths were exported to a VCF file and used to evaluate accuracy of imputation and genomic prediction with the PHG.

All steps described above were run using a Docker image containing necessary software and scripts available for download from the PHG Bitbucket site. We chose to run scripts through a Docker container because the PHG processing

pipeline involves steps from multiple software applications. The Docker container implementation makes it easier to keep compatible versions of all software together as a single download. Specific parameters used to run each script can be found in the config.txt file on bitbucket. Additional scripts and files for creating the sorghum PHG are available at https://bitbucket.org/bucklerlab/p_sorghumphg/src/master.

2.5 | Imputation accuracy

2.5.1 | PHG imputation accuracy for WGS

WGS data for the Chibas founder taxa were downsampled with seqtk (Li, 2013) to 1x, 0.1x, and 0.01x coverage. Sequences were produced with three separate seed integers to create three unique sets of reads at each level of coverage. The full WGS data and each set of down-sampled sequencing reads were run through the PHG findPaths pipeline using a PHG database with nodes built from the Chibas founders, $\text{minReads} = 0$, $\text{minTaxa} = 1$, and all other parameters left at default values. Setting the minReads parameter to 0 means that the HMM will attempt to find a path through the entire genome, even when there is no sequence data observed at a particular reference range. Setting the minTaxa parameter to 1 means that all haplotypes are kept, even if taxa are too divergent to group with other individuals in the database. The SNPs were written at all variant sites in the graph, as well as all positions in the sorghum hapmap (Lozano et al., 2019). The

SNP calling accuracy was assessed by comparing PHG SNP calls to a set of 3,468 GBS SNPs (Muleta et al., unpublished data, 2019). The SNPs with minor allele frequency $<.05$ or call rate $<.8$ were removed before comparing PHG and GBS SNP calls. Haplotype calling accuracy was evaluated by running low-coverage sequence through the database and counting the number of times that the selected node in the graph contained the taxon being imputed.

While error rates for most taxa were consistent with the overall error, BF-95-11-195 stood out as having a five-fold higher error than expected in calling SNPs, although its haplotype calling error was not abnormally high. We suspect this sample was mixed up or contaminated with DNA from another sample during sequencing but kept BF-95-11-195 in the database and included it in all analyses.

2.5.2 | Beagle 5.0 imputation accuracy

Because the PHG is expected to be useful when only skim sequence data is available for an individual, we compared PHG imputation accuracy to Beagle 5.0 (Browning & Browning, 2016) imputation accuracy from low-coverage sequence. The WGS data for each taxon was down-sampled as described above. Each down-sampled dataset and the full-coverage ($\sim 8x$) WGS data from 24 founders of the Chibas sorghum breeding program was aligned to the sorghum v3.0 reference genome with BWA MEM (Li & Durbin, 2009; McCormick et al., 2017) and variants were called with the Sentieon DNaseq variant calling pipeline (Sentieon DNaseq, 2018). The VCF files for each founder were merged using bcftools (Li et al., 2009). When variant sites did not line up in the full coverage WGS (i.e., a variant was called for one individual but not for another such that merging variant calls across taxa would produce a missing call in some taxa and an alternate allele call in others), the unobserved site was assumed to be the reference call. To simplify both the Beagle and PHG imputation pipelines and because individuals used in the database construction were expected to be inbred lines, all heterozygous calls were assumed to come from sequencing and genotyping errors rather than residual heterozygosity and were removed. For the down-sampled datasets, unobserved sites were left as missing. A reference panel created from full-coverage WGS was used to impute SNPs in the down-sampled VCF files. No sites in the down-sampled data were masked; instead, missing information was imputed directly using the reference panel. In the full-coverage dataset, 1% of all sites were masked and re-imputed. Imputation accuracy at all levels of sequence coverage was evaluated by comparing Beagle calls to a set of 3,849 GBS SNPs.

2.5.3 | Designing amplicon sequencing markers and using the PHG to impute variants

Amplicon sequencing technologies like rhAmpSeq use PCR amplification to identify SNPs at targeted sites in the genome (Fresnedo-Ramírez et al., 2019). Genomic prediction with custom rhAmpSeq markers was tested with and without PHG imputation. These rhAmpSeq markers were developed using 100 taxa from the ICRISAT mini-core collection, which are also used in the diversity PHG (Supplemental Table 1). Paired SNP variants between 10 and 100 bp apart were identified in this panel of 100 taxa and designated as potential haplotype regions. Each potential haplotype region was expanded on either side of the SNP pair to generate 104-bp segments centered on the initial pair of SNPs. This identified 336,082 potential haplotype regions, and the polymorphic information content (PIC) score was calculated for each haplotype using the 100-taxa panel.

The sorghum reference genome annotation (Sbicolor 313, annotation v3.1) and sequence (Sbicolor 312, assembly v3.0) were used to divide the chromosome-level assembly into 2,904 genomic regions. Each region contained equal numbers of non-overlapping gene models; overlapping gene models were collapsed into a single gene model. Of these regions, 2,892 contained at least one SNP-pair haplotype. For each region, the SNP-pair haplotype with the highest PIC score was selected as a representative marker locus. These genome-wide candidates, along with 148 target marker regions of interest provided by the sorghum breeding community, were used by the rhAmpSeq team at Integrated DNA Technologies to design and test rhAmpSeq genotyping markers. After design and testing, markers for 1,974 genome-wide haplotype targets and 138 community-identified targets were selected as the rhAmpSeq amplicon set.

The rhAmpSeq sequence data was processed through the PHG findPaths pipeline in the same way as the random skim sequence data described above. To determine how many markers the PHG needs for imputation, we randomly sampled 500 and 1,000 loci from the original set of 2,112 haplotype targets and used the PHG findPaths pipeline to impute SNPs across the rest of the genome. Results were written to a VCF file and used for genomic prediction.

2.6 | Genomic prediction

The PHG SNP performance in genomic prediction was evaluated using a set of 207 individuals in the Chibas training population for which GBS (Elshire et al., 2011) and rhAmpSeq SNP data was also available. The PHG genotypes were predicted with the findPaths pipeline of the PHG using either random skim sequence data at approximately 0.1x or 0.01x

coverage, or rhAmpSeq reads for 2,112, 1,000, or 500 loci (corresponding to 4,854, 1,453, and 700 SNPs, respectively) as inputs. Paths were determined by using an HMM to extrapolate across all reference ranges ($\text{minReads} = 0$, $\text{removeEqual} = \text{false}$). Genomic relationship matrices based on PHG-imputed SNPs were created with the “EIGMIX” option in the SNPRelate R package (Zheng et al., 2012). A haplotype relationship matrix using PHG consensus haplotype IDs was created as described in Equation 2 of Jiang, Schmidt, and Reif (2018), using the tcrossprod function in base R. For GBS markers, markers with more than 80% missing or minor allele frequency $\leq .05$ were removed from the dataset and missing markers were imputed with mean imputation, and a genomic relationship matrix was computed as described in Endelman et al., (2011). Genomic prediction accuracies were Pearson’s correlation coefficients between observed and predicted genotype means, calculated with 10 iterations of five-fold cross validation. The GBS and rhAmpSeq SNP data without PHG imputation were used as a baseline to determine prediction accuracy. To see if the PHG could impute WGS starting from rhAmpSeq amplicons, genomic prediction accuracies using the PHG with rhAmpSeq-targeted loci were compared to prediction accuracies using rhAmpSeq data alone.

3 | RESULTS

We developed two sorghum PHG databases. One contains only the original founder haplotypes of the Chibas breeding population (“founder PHG”, 24 genotypes), while the other PHG contains both the Chibas founders and WGS from an additional 374 taxa that reflect the overall diversity within sorghum (“diversity PHG,” 398 genotypes). We determined how much sequence coverage is needed for the PHG and how genomic prediction with PHG-imputed markers compares to genomic prediction with GBS and rhAmpSeq markers. Data was processed through the founder PHG and the diversity PHG in the same way.

3.1 | A PHG for global sorghum diversity

The sorghum diversity PHG stores sequence information for 398 diverse inbred lines at 19,539 reference ranges covering all genic regions of the genome and is built from WGS data with coverage ranging from 4 to 40x, although most individuals have 10x coverage or less. The founder PHG contains WGS at $\sim 8x$ coverage for 24 founders of the Chibas breeding program. A gVCF file is made by calling variants between WGS and the reference genome, and variants from the gVCF are added to the PHG database in all genic reference ranges. The reference ranges in both versions of the sorghum PHG

are centered around gene regions. At each reference range, haplotypes are collapsed into consensus haplotypes to combine similar taxa and fill in missing sequence across the graph. There is a tradeoff when choosing a divergence cutoff for consensus haplotypes: a low divergence level will retain lower-frequency SNPs, but not fill in gaps and missing data as well as a high divergence level. In both the diversity PHG and the founder PHG, consensus haplotypes were created by collapsing haplotypes that had fewer than 1 in 4,000-bp differences ($\text{mxDiv} = .00025$), which is a slightly lower density of variants than the GBS SNP density reported by Morris et al. (2013). This level was chosen because it marks an inflection point in the number of consensus haplotypes that are created (Figure 3a), with an average of five haplotypes per reference range in the founder PHG and intermediate levels of missingness and discordance with WGS calls made with the Sentieon pipeline (Figure 3b, 3c). The consensus haplotypes produced at this divergence level were used to evaluate PHG SNP-calling and genomic prediction accuracy.

3.2 | PHG SNP-calling accuracy is minimally affected by read count

The PHG haplotype and SNP calling accuracies are minimally affected by decreasing amounts of sequence data. The PHG was evaluated to determine the lower boundary of sequence coverage before imputation accuracy decreased substantially. For each founder in the Chibas breeding program, WGS was subset down to 2,433,333, 243,333, and 24,333 reads, corresponding to 1x, 0.1x, and 0.01x genome coverage, respectively. Sequencing reads were randomly selected from the original WGS fastq files and used to predict SNPs or haplotypes with the PHG, and PHG-predicted SNPs and haplotypes at each level of sequence coverage were evaluated for accuracy. Haplotypes were considered correct if the imputed haplotype node for a given taxon also contained that taxon in the PHG. Single nucleotide polymorphisms were considered correct if they matched GBS calls at 3,369 loci for which GBS data had a minor allele frequency $> .05$ and a call rate $> .8$.

Haplotype error was higher than SNP calling error in both the founder PHG database (24 taxa) and the diversity PHG database (398 taxa), and accuracy increased in both databases with increasing sequence coverage. Both haplotype and SNP error rates were lower with PHG imputation than with a naïve imputation that always imputes the major allele. Haplotype error ranged from 11.5–12.1% in the founder database to 18.6–23.5% in the diversity database. The SNP error ranged from 2.9 to 5.9% and 4.3 to 15.2% in the founder and diversity PHG databases, respectively (Figure 4). Higher haplotype error rates are likely due to similarity among haplotypes that leads the HMM to call an incorrect haplotype even if most of the SNPs within that haplotype are correct. We also

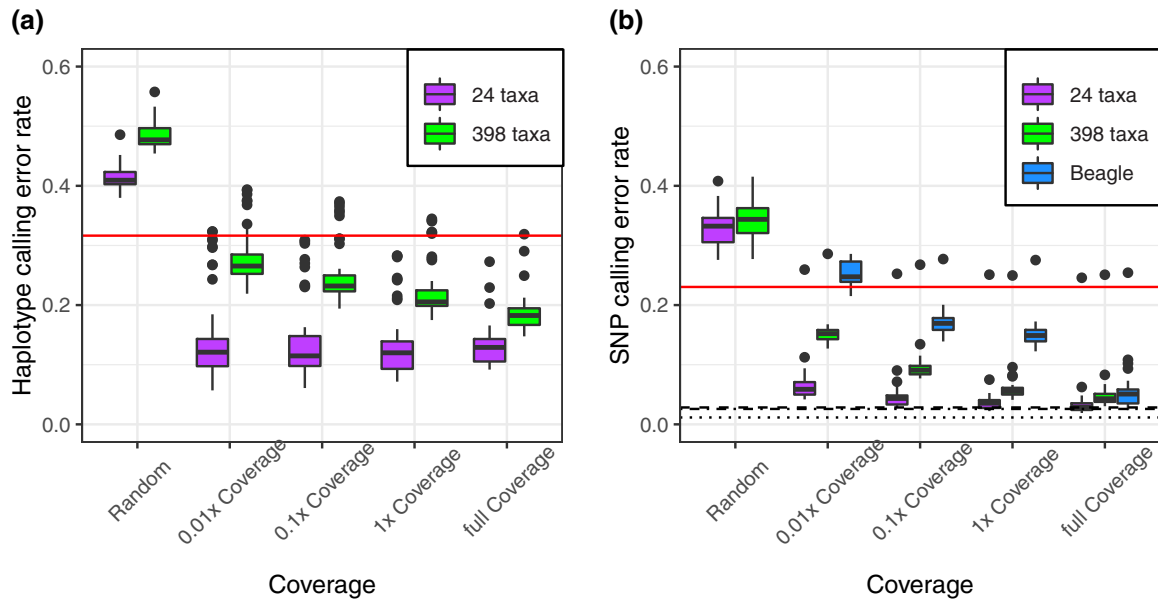


FIGURE 4 Practical haplotype graph (PHG) haplotype (a) and single nucleotide polymorphism (SNP) (b) error rates compared to genotyping-by-sequencing (GBS) data with the 24-taxa founder PHG database (purple) and the 398-taxa diversity PHG database (green) for a random path through the PHG and a range of sequence coverages. Beagle SNP calling accuracy (blue) at each coverage level is included in (b). Solid red horizontal lines in both plots represent the error rate for a naïve imputation where the algorithm always imputes the major allele for the Chibas taxa. Black horizontal lines in (b) represent the best-case imputation results: minimum achievable error rate of the PHG (dotted), GBS (dot-dash), and Beagle (dashed). Points outside of boxes represent outlier taxa with abnormally high error rates

compared imputation accuracies with the founder PHG for a set of unrelated individuals and found SNP error ranging from 2 to 32% depending on sequence coverage (Supplemental Figure 1). Increasing accuracy with coverage suggests that the correct haplotypes are in the founder PHG database, but the recombination break points of the new individuals are not captured in the existing consensus haplotypes.

The PHG-imputed SNP calls were better than Beagle-imputed SNP calls when using low-coverage skim sequence, and about equal to Beagle calls at full coverage ($\sim 8x$) WGS. Notably, while SNP calling accuracy did improve with higher sequence coverage for all imputation methods, calling SNPs with the founder PHG resulted in only a 3% difference in SNP-calling accuracy between the highest coverage tested (8x coverage, 2.9% error) and the lowest coverage tested (0.01x coverage, 5.9% error; Figure 4b). In contrast, there was a 23% difference in SNP-calling accuracy between low and high coverage sequence when using Beagle for imputation. These results indicate that the PHG is a useful imputation tool that can impute SNPs more accurately than Beagle when starting from low-coverage skim sequence.

To determine the PHG baseline error rate, we looked at the intersection of PHG, Beagle, and GBS SNP calls at 3,363 loci in 24 taxa. These loci were chosen because they represented biallelic SNPs called with the GBS pipeline that also had genotype calls made by both the PHG and Beagle imputation methods. The baseline error was calculated as the pro-

portion of SNPs where genotype calls from one of the three methods did not match the other two. Using this metric, baseline error for Beagle imputation, GBS SNP calls, and PHG imputation were calculated to be 2.83%, 2.58%, and 1.15%, respectively (Figure 4b, dashed and dotted lines). To investigate the source of the 1.15% PHG error, we compared the SNP calls from a model path through the PHG (i.e., the calls that the PHG would make if it called the correct haplotype for every taxon at every reference range) to the incorrect PHG SNP calls. Allele calls that were correct in the model SNP set but not called in the genotypes predicted by the findPaths pipeline were counted as an error in the pathfinding step, which is caused by the HMM incorrectly calling the haplotype at a reference range. Allele calls that were not present in the model SNP set were counted as an error in the consensus step. Consensus errors are due to alleles being merged in the createConsensus pipeline because of similarity in haplotypes. Our analysis found that 25% of the PHG baseline error comes from incorrectly calling the haplotype at a given reference range (pathfinding error), while 75% comes from merging SNP calls when creating consensus haplotypes (consensus error). Haplotype and SNP calls from the founder PHG were more accurate than calls with the diversity PHG at all levels of sequence coverage. Therefore, subsequent analyses were done with the founder PHG.

We compared accuracy in calling minor alleles between PHG and Beagle SNP calls. Beagle accuracy drops when

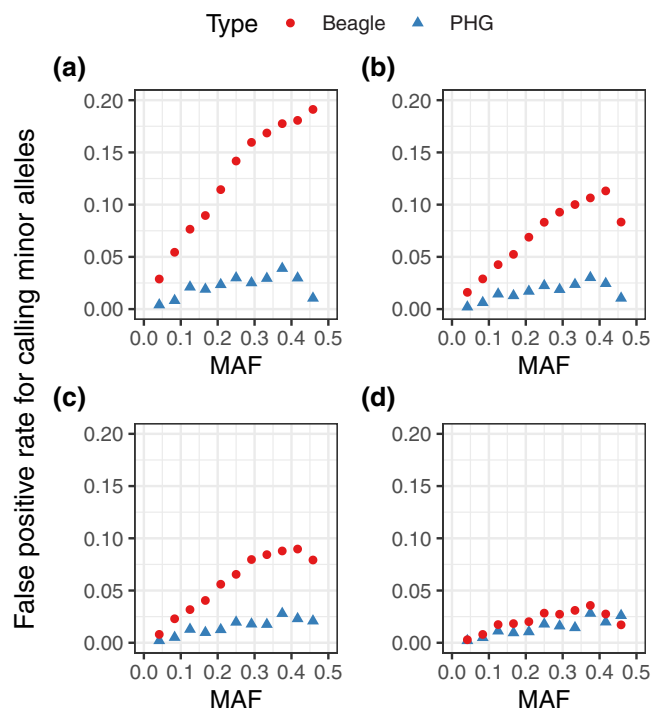


FIGURE 5 False positive rate when calling minor alleles at each minor allele frequency (MAF). Plots show the average number of incorrectly called minor alleles divided by the total number of minor alleles at each allele frequency for Beagle (red circles) and the practical haplotype graph (PHG; blue triangles) at each level of sequence coverage. (a) 0.01x coverage, (b) 0.1x coverage, (c) 1x coverage, (d) full (8x) coverage

dealing with datasets where 90–99% of sites are missing (0.1 or 0.01x coverage) because it makes more errors when calling minor alleles (Figure 5, red circles). When imputing from 0.01x coverage sequence, the PHG calls minor alleles correctly 73% of the time, whereas Beagle calls minor alleles correctly only 43% of the time. The difference between PHG and Beagle minor allele calling accuracy decreases as sequence coverage increases. At 8x sequence coverage, both methods perform similarly, with minor alleles being called correctly 90% of the time. The PHG accuracy in calling minor alleles was consistent regardless of minor allele frequency (Figure 5, blue triangles).

3.3 | PHG genomic prediction accuracies match genomic prediction accuracies from GBS

To test whether PHG haplotype and SNP calls predicted from low-coverage sequence are accurate enough to use for genomic selection in a breeding program, we compared prediction accuracies with PHG-imputed data to prediction accuracies with GBS or rhAmpSeq markers. We predicted breeding values for 207 individuals from the Chibas train-

ing population for which GBS, rhAmpSeq, and random skim sequencing data was available. Haplotype IDs from PHG consensus haplotypes were also tested to evaluate prediction accuracy from haplotypes instead of SNPs (Jiang et al., 2018). The five-fold cross-validation results suggest that prediction accuracies for SNPs imputed with the PHG from random skim sequences are similar to prediction accuracies from GBS SNP data for multiple phenotypes, regardless of sequence coverage for the PHG input. Haplotypes can be used with equal success; prediction accuracies using PHG haplotype IDs were similar to prediction accuracies using PHG or GBS SNP markers (Figure 6a). Results are similar with the diversity PHG database (Supplemental Figure 2). With rhAmpSeq markers, adding PHG-imputed SNPs matched, but did not improve, prediction accuracies relative to accuracy with rhAmpSeq markers alone (Figure 6b). Using the PHG to impute from random low-coverage sequence can, therefore, produce genotype calls that are just as effective as GBS or rhAmpSeq marker data, and SNP and haplotype calls predicted with the findPaths pipeline and the PHG are accurate enough to use for genomic selection in a breeding program.

4 | DISCUSSION

4.1 | SNP calling accuracy

The PHG is a cost-effective genotyping tool that combines WGS data in a database to capture the main haplotype groups in a breeding program or species. Sequences and consensus haplotypes stored in the PHG can be used for genomic prediction. We built a diversity PHG with 398 individuals to capture sorghum-wide diversity and a second, smaller database with only the 24 breeding program founders. In general, the 24-taxa founder PHG database had higher SNP and haplotype calling accuracy, but both databases produced genotypes that could be used effectively for genomic prediction.

When testing the accuracy of the PHG, we find that random skim sequence data can be imputed for SNPs across the PHG reference ranges with high accuracy. Based on the levels tested, 0.01x coverage is the most cost-effective level of sequence coverage with 94.1% SNP calling accuracy—only a 3% drop in SNP calling accuracy relative to accuracy at 8x-coverage WGS. For the sorghum genome, 0.01x coverage corresponds to ~25,000 completely random paired-end 150-bp reads. The sequence reads tested here were selected randomly and are unlikely to cover all reference ranges, which shows that the PHG can impute across reference ranges even when sequence can only be aligned to a portion of the ranges in the database. Long-read sequence data, which creates fewer reads, therefore, could also be used as input for the PHG path-finding algorithm (findPaths pipeline). A few long reads spaced randomly across the genome would likely identify

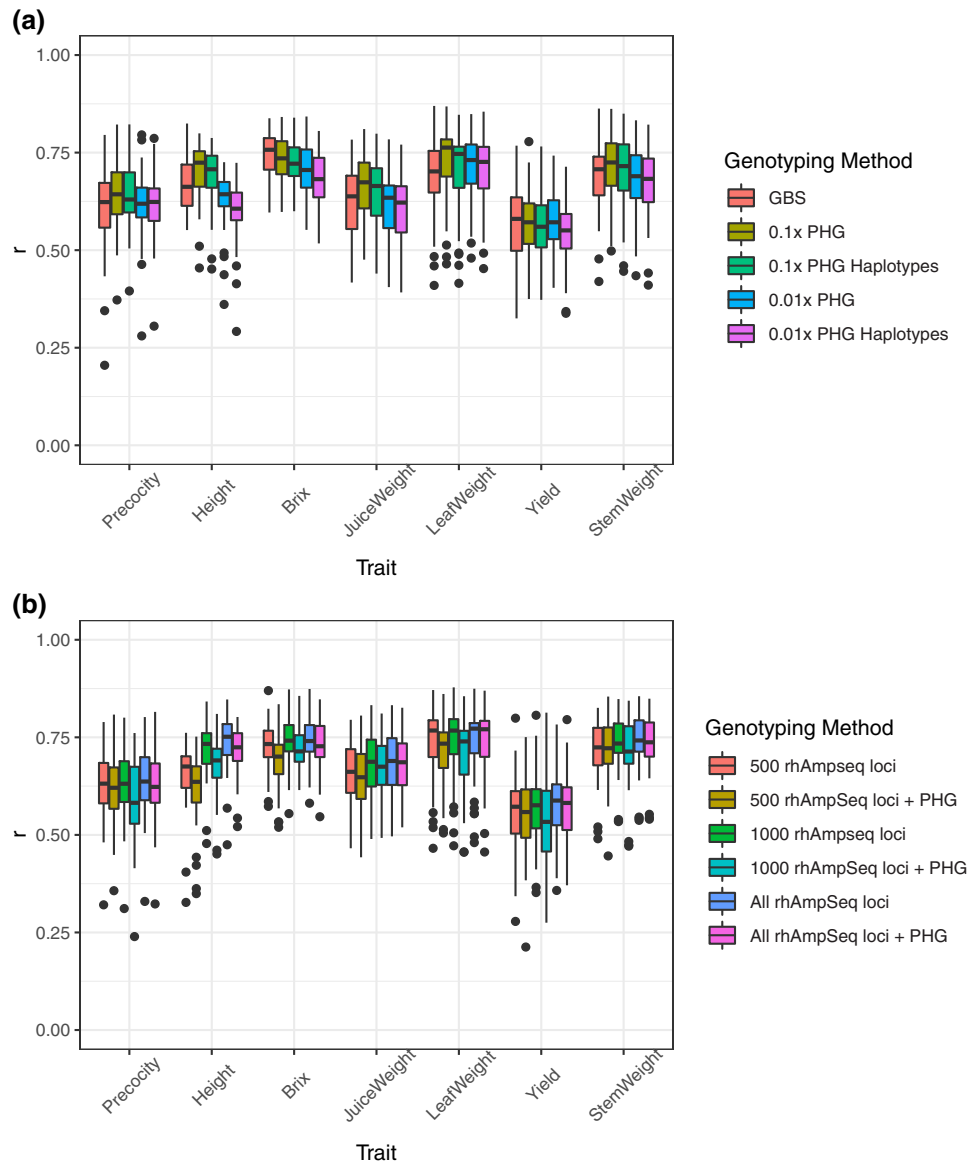


FIGURE 6 Five-fold cross validation prediction accuracies for the Chibas training population ($n = 207$; 10 iterations) are similar when using (a) genotyping-by-sequencing (GBS), practical haplotype graph (PHG) single nucleotide polymorphisms (SNPs) with 0.1x and 0.01x sequence coverage, or PHG haplotypes. (b) Prediction accuracies do not change when using rhAmpSeq markers alone or using rhAmpSeq markers with additional markers from the PHG, regardless of how many rhAmpSeq markers are included

haplotypes with similar levels of accuracy as 0.01x coverage short-read sequence data.

The imputation accuracies reported here used a set of founder taxa from the Chibas breeding program to build the PHG and reported imputation accuracies for imputing SNPs in these same taxa, which is similar to the genotyping needs that might be encountered in a breeding program. In this case, important parent lines would be used to build the PHG, and then genotypes calculated for a derived (and similar) progeny population. As with genomic prediction, the imputation accuracy is expected to decay as individuals being genotyped diverge from the core set of genotypes used in the PHG database (Muleta et al., 2019). To maintain high imputation

accuracies, the PHG works best when the program founders or important parents are sequenced and included in the database when creating consensus haplotypes.

The PHG can be updated to capture new information as new data are generated or new germplasm is added to a breeding program. For example, in a breeding program, new individuals can be periodically added to the PHG database to update genotypes as the breeding program progresses, or a smaller subset of target individuals can be used to predict genotypes if founders are removed from the breeding pool. If the PHG is built with the full genome, the list of reference ranges can be adjusted and intervals between reference ranges can also be included in the set of reference ranges. The PHG may be

useful for other applications in population genetics, or diversity and evolution studies if a more diverse set of individuals is used to build the database.

4.2 | Genomic prediction accuracy

Both 0.01x and 0.1x coverage sequence imputed with the PHG, as well as haplotype IDs from the PHG, can be used for genomic prediction with prediction accuracies similar to those produced by GBS markers. In the training dataset comprising 207 individuals, there was no difference in using a haplotype relationship matrix instead of genomic relationship matrix built from PHG SNPs. However, in larger datasets with more individuals, using haplotype IDs instead of SNP markers may improve computational efficiency without a cost in terms of prediction accuracy. Using the PHG with rhAmpSeq markers worked as well as using rhAmpSeq markers alone for complex traits, but prediction accuracies dropped slightly for some traits (e.g., height, juice weight) if only 500 rhAmpSeq markers were used with PHG imputation. This could be related to trait genetic architecture; height is an oligogenic trait in sorghum, while traits like grain yield and precocity would be expected to be more polygenic (Girma et al., 2019; Pereira & Lee, 1995).

Low-coverage sequence in conjunction with the PHG allows large-scale imputation of progeny genotypes across the genome. We show that the PHG can be used with random skim sequence and rhAmpSeq sequences, suggesting that the PHG can also be used to unify sequence data from diverse genotyping platforms. With only 0.01x genome coverage needed for accurate calls, per-taxon sequencing costs become negligible compared to the cost of DNA extraction and library preparation (both of which are also continuing to decrease; see Anderson et al., 2018; Baym et al., 2015; Zou et al., 2017). These new technologies, together with the PHG, provide breeders with an effective, inexpensive, and fast tool for imputation and genomic selection that works across genotyping platforms. Since the founder PHG produces more accurate SNP and haplotype calls than the diversity PHG, we also suggest that small PHG databases built for specific breeding programs can be more effective than a single species-wide haplotype graph.

It is surprising that the diversity PHG had lower SNP and haplotype calling accuracy than the founder PHG because it contains more information about the species diversity overall. We hypothesize that the difference in performance is due to differences in allele frequency or LD patterns between the Chibas breeding program and the taxa in the diversity PHG. As expected for a breeding program, the Chibas breeding material captures much less diversity than is present in the species as a whole. If common alleles in the Chibas founder individuals are rare relative to the taxa included in the diversity PHG, then the Chibas founders may have been pulled into

consensus haplotypes with alleles that are not common in the Chibas breeding program. Using these consensus haplotypes to impute from skim sequence could have added alleles that are not present in the founder PHG which would increase SNP calling error relative to GBS. Linkage disequilibrium patterns between sorghum races may also differ. Therefore, transition probabilities between reference ranges may be estimated less accurately in a diverse PHG database comprising multiple races. For now, if working in a breeding context where diversity is limited, a founder PHG with program-specific haplotypes appears to work best. In the future, this issue could be solved by building a diverse PHG but using only a specific subset of taxa (in this case, the Chibas founders) to build consensus haplotypes. Sequence from other taxa in the database would be disregarded when creating consensus haplotypes but could be used to fill in gaps in the sequence data for particular taxa if still needed after building consensus haplotypes. The resulting haplotypes would have less missing information than the current founder PHG but would maintain the allele frequencies and haplotype patterns of the original set of anchor taxa. The entire PHG database could then be useful both to represent diversity within the species and diversity within a specific program.

4.3 | Decreasing genotyping costs

The cost of building a PHG depends on the number of individuals for which WGS or de novo assemblies need to be produced. Relying on existing resequencing data when possible can significantly reduce the overall cost. For the sorghum PHGs produced here, all taxa in the Chibas sorghum breeding program were multiplexed in a single sequencing lane, resulting in approximately 8x coverage for each individual and low levels of missing data. The initial sequencing investment for the founder PHG was \$5,283. The additional 374 taxa added to the diversity PHG were produced for other research purposes and no additional sequence data were produced for these individuals. Thus, the total upfront cost for building the sorghum PHGs was under \$6,000—less than the genotyping costs for a round of genomic selection.

The PHG aims to make genotyping and genomic selection marker agnostic, that is, all marker systems should produce similar results. We see the PHG likely to be used with five current and future platforms. The most expensive is GBS at approximately \$15 per sample, which is substantially driven by expensive DNA preparation and uneven library coverage. A Tn5-based skim sequencing can use simple DNA extraction protocols, has the same procedure for any species, and costs ~\$10 per sample. Targeted amplicon-based sequencing can use very inexpensive sample preparation protocols and provides 500–2,000 loci for \$3.50–10 per sample. It does, however, require significant upfront investment to develop

amplicons for each species, and the per-sample cost is dependent on the number of samples processed annually. The cost of random prime sequencing with simple DNA extractions is similar to targeted amplicon sequencing at \$5–10 per sample, but the price for random prime sequencing does not depend on sample number. Long-read sequence data can also be used with the PHG and the price is likely to drop as long-read technologies are developed further. The PHG is designed to work with any of these sequence types, making it possible to unify sequence data from multiple genotyping platforms. Because the per-sample read count can vary with skim-sequencing methods (e.g., read count varied from 12 to 480k reads per sample in skim sequence of the Chibas training set), reliable imputation is also necessary to allow fair comparison across genotypes.

4.4 | Turnaround time

Continuous GS cycles require rapid data turnaround so that there is enough time to identify the best progeny and plan crosses before flowering. In the Chibas breeding program, tissue samples can be collected about 1 mo after planting. For continuous GS cycles to be possible, all sample processing, sequencing, and variant calling must be complete 6 wk after samples are collected. Using a DNA extraction service took 2 wk, with another 1.5 wk needed for library preparation and sequencing. Sequencing data took 48 h to run through the founder PHG, giving us about a 4-wk turnaround time from sample collection to genotype data and leaving 2 wk to make selections and plan crosses. Additional time could potentially be saved with cheap, on-site DNA extractions and library preparations (Anderson et al., 2018; Zou et al., 2017). The extra time gained with on-site methods could make a second round of tissue collection and sequencing possible if samples are contaminated or if more data is needed for a particular breeding line. With these improvements, genotyping would no longer be the factor limiting turnaround time, and techniques like speed breeding could be used to further reduce cycle time in sorghum breeding (Watson et al., 2018).

5 | CONCLUSIONS

We developed two PHGs: a diversity PHG with 398 individuals consisting of all five sorghum races and a founder PHG with individuals specific to the Chibas sorghum breeding program, which consists mostly of breeding lines derived from east African caudatum sorghums (Muleta et al., unpublished data, 2019). Our study demonstrated how the PHG can accurately impute SNPs and haplotypes from low-coverage random sequence data and predict genotypes for a breeding population. The PHG was able to produce bet-

ter imputed genotype calls than Beagle 5.0 when using low-coverage random sequence data, and prediction accuracies from PHG-imputed SNPs matched those produced with higher-cost GBS marker data for traits with a range of genetic architectures when starting from similar levels of sequence coverage. These results demonstrate that the PHG is a useful low-cost genotyping tool for breeders around the globe.

ACKNOWLEDGMENTS


This study is made possible by the support of the American People provided to the Feed the Future Innovation Lab for Collaborative Research on Sorghum and Millet through the United States Agency for International Development (USAID) under Associate Award No. AID-OAA-LA-16-00003. The information, data, or work presented herein was also funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000598. The contents are the responsibility of the authors and do not necessarily state or reflect those of USAID, the United States Government, or any agency thereof. We thank Integrated DNA Technologies for a custom rhAmpSeq panel in sorghum. Additional support comes from the United States Department of Agriculture, Agricultural Research Service and the Bill and Melinda Gates Foundation.

DATA AVAILABILITY

Raw sequence data for this project is available at NCBI BioProject ID PRJNA575060 and PRJNA598805. All processed data can be found on CyVerse at <https://doi.org/10.25739/nccb-0k78>. Scripts to replicate the analyses can be found on bitbucket at https://bitbucket.org/bucklerlab/p_sorghumphg/src/master/.

ORCID

Sarah E. Jensen  <https://orcid.org/0000-0002-4432-4715>


Jean Rigaud Charles 

<https://orcid.org/0000-0001-9271-9909>

Kebede Muleta  <https://orcid.org/0000-0001-5403-5098>

Peter J. Bradbury  <https://orcid.org/0000-0003-3825-8480>

Terry Casstevens  <https://orcid.org/0000-0001-7602-0487>

Santosh P. Deshpande 

<https://orcid.org/0000-0002-7182-3456>

Michael A. Gore  <https://orcid.org/0000-0001-6896-8024>

Rajeev Gupta  <https://orcid.org/0000-0002-1451-215X>

Daniel C. Ilut  <https://orcid.org/0000-0001-7497-9109>

Lynn Johnson  <https://orcid.org/0000-0001-8103-2722>

Roberto Lozano  <https://orcid.org/0000-0003-0760-4977>

Zachary Miller  <https://orcid.org/0000-0002-5454-4527>

Punna Ramu  <https://orcid.org/0000-0003-3998-568X>

Abhishek Rathore  <https://orcid.org/0000-0001-6887-4095>

M. Cinta Romay  <https://orcid.org/0000-0001-9309-1586>

Hari D. Upadhyaya 
<https://orcid.org/0000-0002-5166-6844>
 Rajeev K. Varshney 
<https://orcid.org/0000-0002-4562-9131>
 Geoffrey P. Morris 
<https://orcid.org/0000-0002-3067-3359>
 Gael Pressoir 
<https://orcid.org/0000-0001-9062-6530>
 Edward S. Buckler 
<https://orcid.org/0000-0002-3100-371X>
 Guillaume P. Ramstein 
<https://orcid.org/0000-0003-1547-3268>

REFERENCES

- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., ... Zhou, X. (2016). 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, *166*(2), 481–491. <https://doi.org/10.1016/j.cell.2016.05.063>
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>
- Anderson, C. B., Franzmayr, B. K., Hong, S. W., Larking, A. C., Stijn, T. C., Tan, R., ... Griffiths, A. G. (2018). Protocol: A versatile, inexpensive, high-throughput plant genomic DNA extraction method suitable for genotyping-by-sequencing. *Plant Methods*, *14*(1). <https://doi.org/10.1186/s13007-018-0336-1>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, *67*(1). <https://doi.org/10.18637/jss.v067.i01>
- Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. (2015). Inexpensive multiplexed library preparation for megabase-sized genomes. *PLOS ONE*, *10*(5). <https://doi.org/10.1371/journal.pone.0128036>
- Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., ... Devos, K. (2012). Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology*, *30*(6), 555–561. <https://doi.org/10.1038/nbt.2196>
- Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082–1090. <https://doi.org/10.2135/cropsci2006.11.0690>
- Bouchet, S., Olatoye, M. O., Marla, S. R., Perumal, R., Tesso, T., Yu, J., ... Morris, G. P. (2017). Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. *Genetics*, *206*(2), 573–585. <https://doi.org/10.1534/genetics.116.198499>
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics*, *98*(1), 116–126. <https://doi.org/10.1016/j.ajhg.2015.11.020>
- Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., ... Xu, Y. (2018). Construction of the third-generation *Zea mays* haplotype map. *Gigascience*, *7*(4). <https://doi.org/10.1093/gigascience/gix134>
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE*, *6*(5). <https://doi.org/10.1371/journal.pone.0019379>
- Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome Journal*, *4*(3), 250–255. <https://doi.org/10.3835/plantgenome2011.08.0024>
- Fresnedo-Ramírez, J., Yang, S., Sun, Q., Karn, A., Reisch, B. I., & Cadle-Davidson, L. (2019). Computational analysis of ampseq data for targeted, high-throughput genotyping of amplicons. *Frontiers in Plant Science*, *10*. <https://doi.org/10.3389/fpls.2019.00599>
- Girma, G., Nida, H., Seyoum, A., Mekonen, M., Nega, A., Lule, D., ... Mengiste, T. (2019). A large-scale genome-wide association analyses of Ethiopian sorghum landrace collection reveal loci associated with important traits. *Frontiers in Plant Science*, *10*. <https://doi.org/10.3389/fpls.2019.00691>
- Heffner, E. L., Lorenz, A. J., Jannink, J. L., & Sorrells, M. E. (2010). Plant breeding with Genomic selection: Gain per unit time and cost. *Crop Science*, *50*(5), 1681–1690. <https://doi.org/10.2135/cropsci2009.11.0662>
- Heslot, N., Jannink, J., & Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science*, *55*(1), 1–12. <https://doi.org/10.2135/cropsci2014.03.0249>
- Jiang, Y., Schmidt, R. H., & Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3: Genes, Genomes, Genetics*, *8*(5), 1687–1699. <https://doi.org/10.1534/g3.117.300548>
- Li, H. (2013). seqtk: Toolkit for processing sequences in FASTA/Q formats. Retrieved from <https://github.com/lh3/seqtk>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., & Homer, N. ... 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lozano, R., Gazave, E., dos Santos, J. P. R., Stetter, M., Valluru, R., Bandillo, N., ... Gore, M. A. (2019). Comparative evolutionary analysis and prediction of deleterious mutation patterns between sorghum and maize. *Biorxiv*, <https://doi.org/10.1101/777623>
- Mace, E. S., Tai, S., Gilding, E. K., Li, Y., Prentis, P. J., Bian, L., ... Wang, J. (2013). Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communications*, *4*. <https://doi.org/10.1038/ncomms3320>
- Mace, E. S., Rami, J. F., Bouchet, S., Klein, P. E., Klein, R. R., Kilian, A., ... Jordan, D. R. (2009). A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array Technology (DArT) markers. *BMC Plant Biology*, *9*. <https://doi.org/10.1186/1471-2229-9-13>
- McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., ... Mullet, J. E. (2017). The Sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, *93*(2), 338–354. <https://doi.org/10.1111/tpj.13781>
- Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.
- Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., ... Kresovich, S. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences*, *110*(2), 453–458. <https://doi.org/10.1073/pnas.1215985110>

- Muleta, K. T., Pressoir, G., & Morris, G. P. (2019). Optimizing genomic selection for a sorghum breeding program in Haiti: A simulation study. *G3: Genes, Genomes, Genetics*, *9*(2), 391–401. <https://doi.org/10.1534/g3.118.200932>
- Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., ... Buell, C. R. (2007). The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Research*, *35*, D883–D887. <https://doi.org/10.1093/nar/gkl976>
- Pereira, M. G., & Lee, M. (1995). Identification of genomic regions affecting plant height in sorghum and maize. *Theoretical and Applied Genetics*, *90*, 380–388. <https://doi.org/10.1007/BF00221980>
- Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., ... Jannink, J.-L. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome*, *5*(3), 103. <https://doi.org/10.3835/plantgenome2012.06.0006>
- Rabiner, L. R. (1989). A tutorial on HMM and selected Applications in Speech Recognition. *Proceedings of the Institute of Electrical and Electronics Engineers*, *77*(2), 257–286. Retrieved from <https://courses.physics.illinois.edu/ece417/fa2017/rabiner89.pdf>
- Ramstein, G. P., Jensen, S. E., & Buckler, E. S. (2018). Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theoretical and Applied Genetics*, *132*, 559–567. <https://doi.org/10.1007/s00122-018-3267-3>
- Ribaut, J. M., de Vicente, M. C., & Delannay, X. (2010). Molecular breeding in developing countries: Challenges and perspectives. *Current Opinion in Plant Biology*, *13*(2), 213–218. <https://doi.org/10.1016/j.pbi.2009.12.011>
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S. S., ... Wilson, R. K. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, *326*, 1112–1115. <https://doi.org/10.1126/science.1178534>
- Sentieon DNASeq, “Sentieon,” (2018). [Online]. Retrieved from <https://www.sentieon.com/products/>
- Shakoor, N., Nair, R., Crasta, O., Morris, G., Feltus, A., & Kresovich, S. (2014). A Sorghum bicolor expression atlas reveals dynamic genotype-specific expression profiles for vegetative tissues of grain, sweet and bioenergy sorghums. *BMC Plant Biology*, *14*(1), 35. <https://doi.org/10.1186/1471-2229-14-35>
- Valluru, R., Gazave, E. E., Fernandes, S. B., Ferguson, J. N., Lozano, R., Hirannaiah, P., ... Bandillo, N. (2019). Deleterious mutation burden and its association with complex traits in sorghum (*Sorghum bicolor*). *Genetics*, *211*(3), 1075 LP – 1087. <https://doi.org/10.1534/genetics.118.301742>
- Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., & Bevan, M. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, *463*(7282), 763–768. <https://doi.org/10.1038/nature08747>
- Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M.-D., ... Hickey, L. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature Plants*, *4*(1), 23–29. <https://doi.org/10.1038/s41477-017-0083-8>
- Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. <https://doi.org/10.1093/bioinformatics/bts606>
- Zou, Y., Mason, M. G., Wang, Y., Wee, E., Turni, C., Blackall, P. J., ... Botella, J. R. (2017). Nucleic acid purification from plants, animals and microbes in under 30 seconds. *PLoS Biology*, *15*(11). <https://doi.org/10.1371/journal.pbio.2003916>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Jensen SE, Charles JR, Muleta K, et al. A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *Plant Genome*. 2020;13:e20009. <https://doi.org/10.1002/tpg2.20009>