bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

A Maize Practical Haplotype Graph Leverages Diverse NAM Assemblies

Jose A. Valdes Franco*[1] (0000-0002-0887-5827), Joseph L. Gage*[2] (0000-0001-5946-4414), Peter J. Bradbury (0000-0003-3825-8480)[3], Lynn C. Johnson[2] (0000-0001-8103-2722), Zachary R. Miller[2] (0000-0002-5454-4527), Edward S. Buckler[1,2,3] (0000-0002-3100-371X), M. Cinta Romay[2] (0000-0001-9309-1586)

1. Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA
2. Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA
3. United States Department of Agriculture-Agricultural Research Service, Ithaca, NY 14853, USA

Abstract

As a result of millions of years of transposon activity, multiple rounds of ancient polyploidization, and large populations that preserve diversity, maize has an extremely structurally diverse genome, evidenced by high-quality genome assemblies that capture substantial levels of both tropical and temperate diversity. We generated a pangenome representation (the Practical Haplotype Graph, PHG) of these assemblies in a database, representing the pangenome haplotype diversity and providing an initial estimate of structural diversity. We leveraged the pangenome to accurately impute haplotypes and genotypes of taxa using various kinds of sequence data, ranging from WGS to extremely-low coverage GBS. We imputed the genotypes of the recombinant inbred lines of the NAM population with over 99% mean accuracy, while unrelated germplasm attained a mean imputation accuracy of 92 or 95% when using GBS or WGS data, respectively. Most of the imputation errors occur in haplotypes within European or tropical germplasm, which have yet to be represented in the maize PHG database. Also, the PHG stores the imputation data in a 30,000-fold more space-efficient manner than a standard genotype file, which is a key improvement when dealing with large scale data.

Introduction

The functional diversity of maize (*Zea mays* ssp. *mays* L.) makes it one of the most important crops, enabling its adaptation across the world (Hake & Ross-Ibarra, 2015). Maize is also genomically diverse, having gone through two rounds of whole-genome duplication, one ~70M years ago when grasses diverged (Paterson et al., 2004), and a second one ~11M years ago, an allotetraploidization event that was followed by diploidization (Gaut & Doebley, 1997). In addition to whole-genome duplication events, tremendous activity by transposable elements has further contributed to its genome diversity (Oliver et al., 2013). Previous studies have found considerable variability in the presence and absence of transcribed genes due to transposon activity (Fu & Dooner, 2002)(Lai et al., 2005). More recent analyses have leveraged whole-genome assemblies, allowing a detailed view of the extent of the intraspecific changes between maize varieties. These have identified signals of gene reordering, copy number, and other structural variations (Sun et al., 2018), with some cases

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

accounting for 1.6Gb (equivalent to ~50% of the genome of B73) of variable transposable element sequences (Anderson et al., 2019).

Because maize is both a powerful system for genetics and evolutionary research, along with being a major crop worldwide, accounting for over 38% of the world's cereal production (*FAO*, 2018), it has been genotypically characterized through different approaches that suited each community's needs and the technology available at the time. However, there is a tremendous need to leverage knowledge across these disciplines to facilitate breeding and understand the molecular and evolutionary basis of diversity. This paper leverages the recently assembled NAM founders and various types of sequencing data from thousands of maize lines to call high-density SNP genotypes at a unified set of sites.

The maize Nested Association Mapping (NAM) population was created as a resource for capturing a large proportion of broad maize diversity in a single population. It represents a highly studied set of ~5,000 recombinant inbred lines, generated from the single seed descent of an F1 crossing between 25 diverse maize inbreds and into a common parent, B73 (McMullen et al., 2009), allowing for the dissection of the genetic components underlying the control of maize phenotypes. The NAM population mapping resource (Buckler et al., 2009) has been utilized to identify quantitative trait loci for various complex traits (Gage et al., 2020). A recent NSF funded project (*NAM Genomes Project*, 2020) has produced extremely high-quality chromosome level assemblies of the NAM founders by combining long-read sequencing and optical mapping technologies. This is the first time the maize community has had a large set of equal quality assemblies. Here we aim to leverage these assemblies through a pangenome graph to impute the NAM population's genotypes and other diverse germplasm.

Genotyping technologies vary in cost, accuracy, and number of sites and samples that can be genotyped in a single experiment (Romay, 2018). Specifically, the public maize community has used three major platforms: 1) Genotyping by Sequencing (GBS) has been used on tens of thousands of samples (Gouesnard et al., 2017; Rodgers-Melnick et al., 2015; Romay et al., 2013; Romero Navarro et al., 2017; Wu et al., 2016), but is challenged by short-read mapping issues and single-reference biases; 2) whole-genome sequencing (WGS) has been used in thousands of samples (Bukowski et al., 2018; Wang et al., 2020), with high variability in coverage and a similar reference mapping bias; and 3) SNP arrays, used to genotype thousands of samples over 55K and 600K sites (Unterseer et al., 2014; Xu et al., 2017), which by design have a predefined set of variant positions which are targeted to be genotyped. Additionally, new amplicon approaches (e.g., rhAmpSeq (Zou et al., 2020)) continue to be developed to increase the number of samples genotyped in a single experiment. All these distinct approaches represent a hurdle that needs to be addressed whenever there's an interest in analyzing across genotyping experiments. Whole-genome imputation with a pangenome allows each of these technologies' strengths and weaknesses to be complemented.

Imputation is the process of predicting genotypes that cannot be directly determined in a sample undergoing genotyping. In human studies, imputation approaches tend to leverage large reference panels (Browning et al., 2018; Das et al., 2016), composed of thousands to tens of thousands of samples with haplotypes identified through extremely dense SNP sets (Bycroft et al., 2018; McCarthy et al., 2016; Telenti et al., 2016). In the case of samples not represented by a reference panel, or when expecting some degree of relationship between the individuals in the sample, other imputation approaches attempt to identify identity-by-descent (IBD) segments from individuals that happen to

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

have genotypes with higher marker density than the rest in the sample. The presence and identification of IBD segments allow un-genotyped SNPs' imputation in lower density individuals by identifying their underlying haplotype. However, the human genome is an order of magnitude less diverse than plants like maize, and genotyping platforms for plants generally produce much less dense genotype marker sets. BEAGLE (Browning & Browning, 2013) is a leading tool in humans for within-sample imputation; it is also commonly used in crops, as it performs reasonably well in diverse and heterozygous populations with stable marker sets or high coverage (Chan et al., 2016; Pook et al., 2019). Other imputation approaches aim to leverage the peculiarities of populations within breeding programs; examples are FILLIN (Swarts et al., 2014), which utilizes breeding bottlenecks to capture libraries of haplotypes, and AlphaImpute (Hickey et al., 2012), which leverages the complex pedigrees of the samples under study to impute them.

To better capture the diversity of the plant genomes, some approaches represent this diversity as a collection of haplotypes in a graph, such as VG (Garrison et al., 2018). However, at present, VG cannot deal with the level of diversity in species such as maize. Another haplotype graph approach that can address this is the Practical Haplotype Graph (PHG) (Bradbury et al. in prep). In sorghum, a plant species more diverse than humans, the PHG was used to leverage haplotypes derived from parental samples, sequenced at high coverage, to impute progeny sequenced at very-low coverage (Jensen et al., 2020). Imputation in maize is challenging because of the high levels of divergence and repetitiveness result in poor read mapping. Also, its structural variation makes any single reference genome a poor model for the entire species. However, maize has an extensive collection of inbred varieties, where phasing of alleles becomes unnecessary. Through domestication, maize has gone through various selection bottlenecks, generating a modest subset of highly diverse haplotypes, as most of its diversity evolved in the hundreds of thousands of years before domestication. Here, by generating a database of haplotypes from the NAM founders, we try the first implementation of the PHG to impute samples within the structurally diverse maize species.

This paper asks whether the PHG, implemented as a database of haplotypes and an imputation platform, can address the issues of read mapping, haplotype library completeness, and suitability for genomic and breeding applications. The PHG pangenome representation and alignment processing should help deal with read mapping issues, which we tested with GBS and WGS data. Haplotype libraries are very useful in imputing across breeding programs (e.g., FILLIN), and here we test if the haplotype diversity from the NAM founders can be leveraged through the PHG for genotyping across the NAM RILs and a diverse population. Finally, we compare the PHG imputed genotype calls with known genotype benchmarks for each population to assess its utility in general breeding or genomic analysis.

Results

Representation of the pangenome in the PHG

A Practical Haplotype Graph database was constructed from 27 diverse inbred lines: the 26 parents of the Nested Association Mapping panel (McMullen et al., 2009; *NAM Genomes Project*, 2020) and B104 (reference TBA). The database consists of 71,354 reference ranges: regions from physical intervals of the B73 AGPv5 genome sequence. The edges for each reference range are defined by the starting and ending points of the gene annotations for the B73 assembly, resulting in alternating genic and intergenic reference ranges. These reference ranges allow for the identification of the haplotypes in the pangenome through the sequence aligned to them from each of the 26 other assemblies. In

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

some cases, reference range breakpoints (i.e., edges of genes) could not be aligned from the non-B73 to the B73 assembly, likely due to presence-absence variation.

On average, non-B73 assemblies had sequences aligned to ~87% of the B73 reference ranges (Fig.1). 80% and 69% of intergenic and genic ranges were present in all taxa (Fig.2). When comparing each NAM founder's background, it is apparent that tropical and sweet types have more missing haplotypes relative to the selected B73 reference. However, the sequence contained in the database does represent an average of 99% of each assembly (Fig.3), indicating that almost the complete sequence of them is represented in the stored haplotypes. Additionally, the genome coverage is not affected by the background. The assembly for Oh7B is a relative outlier due to a translocation of sequence between Chr9 and Chr10 (Albert et al., 2010). Our pipeline, which aligns pairs of equivalent chromosomes, misses those haplotypes. However, this translocation is not present in the Oh7B lineage used in breeding programs nor the creation of the NAM RILs. The next version of the PHG database will address this translocation to represent the NAM version of Oh7B.

Comparing each assembly to B73, we identified a median of 1M genic SNPs and 8M intergenic SNPs, for a combined total of 43.1M SNPs over all the assemblies in the database (Fig.4). B73 genic and intergenic divergence agrees with known pedigree backgrounds.

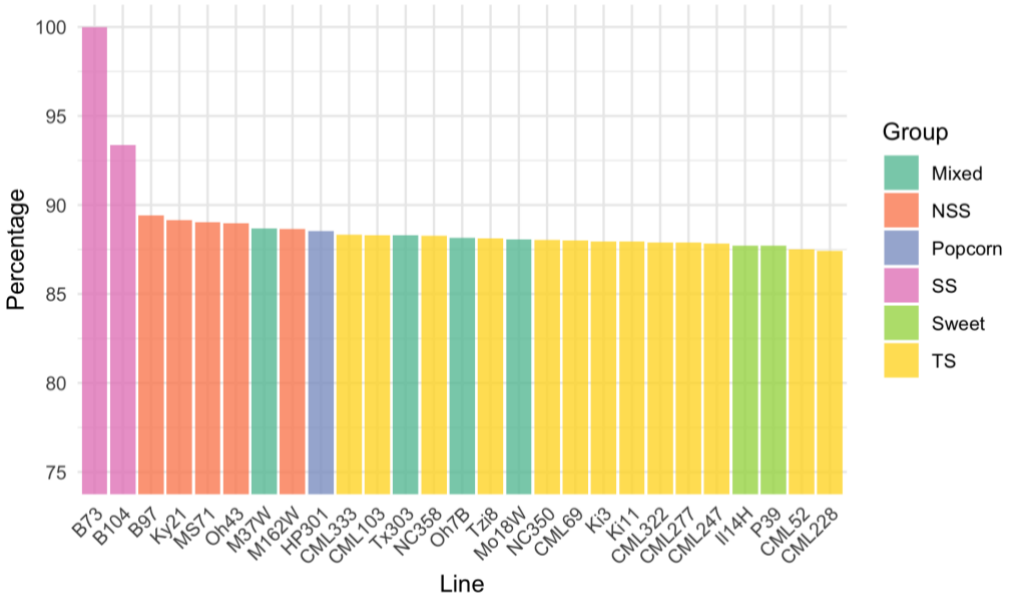


Fig.1. Percent of defined reference ranges with identified haplotypes in the assemblies stored in the database. B73 is at 100 as it is the assembly defining the haplotype regions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

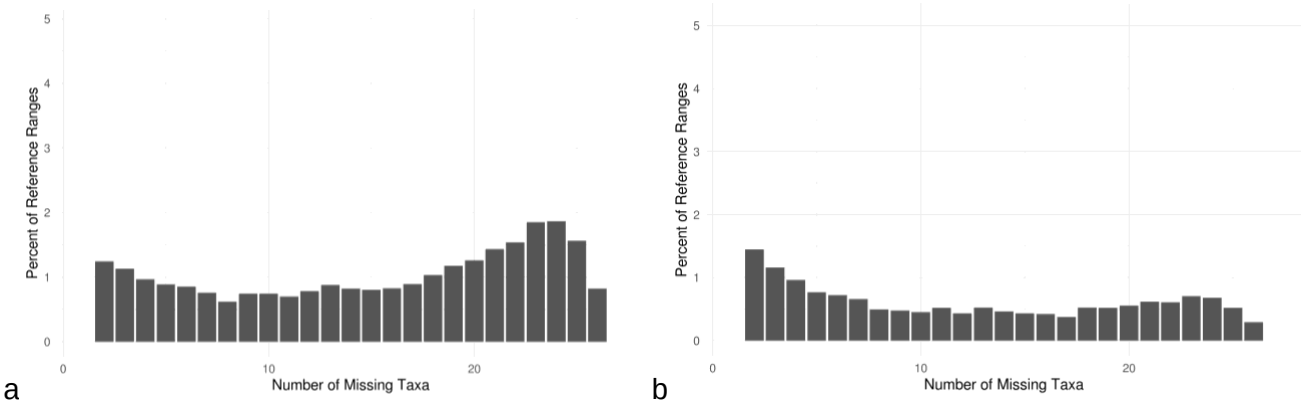


Fig.2. Reference range haplotype missingness. For genic (a) and intergenic (b) regions. A large portion of the reference ranges is found across all taxa, with a small portion of them being private to a subset. Bars at 0, not shown, represent 69% and 80% of genic and intergenic ranges being found across all taxa, respectively.

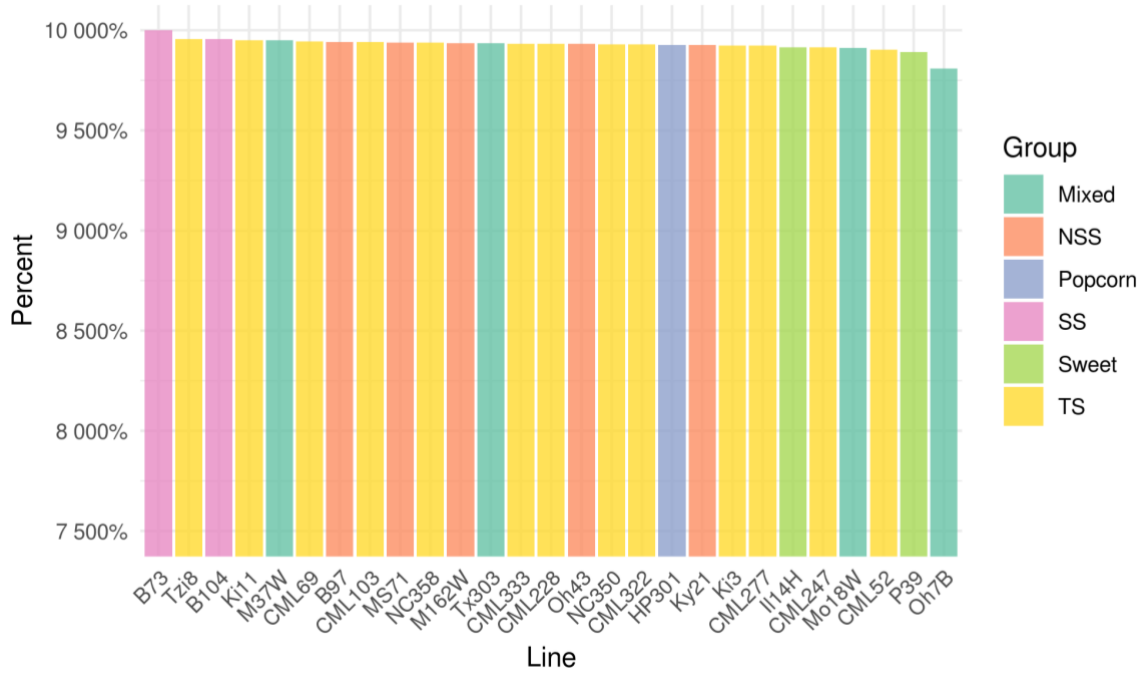


Fig.3. Percent of total assembly sequence contained in the PHG database for each assembly. Note that regardless of haplotypes not being identified for several reference ranges, the final length of the sequence contained is not severely impacted, as “novel” sequences potentially found surrounding the missing haplotypes are included in the adjacent reference range.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

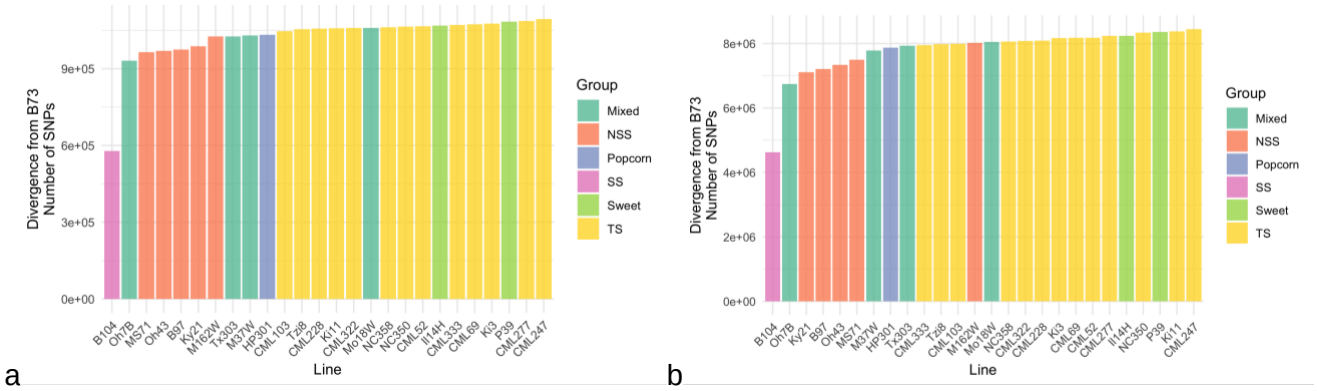


Fig.4. Number of SNPs identified on the haplotypes for each assembly's reference ranges, over genic (a) or intergenic (b) regions. Known genetically close varieties to B73 show the least number of SNPs, while more distant ones have higher numbers.

Genotyping and imputation of related lines using GBS

As an initial test for the maize PHG, we mapped GBS reads from 4705 accessions of the NAM RIL population. The GBS reads are generated with earlier Illumina/Solexa sequencers having ~70bp in length with extremely low coverage (more modern technologies would produce longer and more reads). These were mapped to the pangenome to impute haplotypes and generate SNP calls. To evaluate the imputation accuracy against existing results, we compared the imputed SNPs to 1,106 legacy SNPs (McMullen et al., 2009) and observed an average error rate of 0.8% (Fig.5a and 5b). The families with the larger error rates have known residual heterozygosity, CML52 at 1.4%, Tzi8 at 1.7%, and CML228 at 1.8%. The error rate by position appears evenly distributed throughout the chromosomes (Fig.Sup.1), except for three positions in chromosomes 4, 7, and 8, where all families have higher than average error rates for a subset of the taxa.

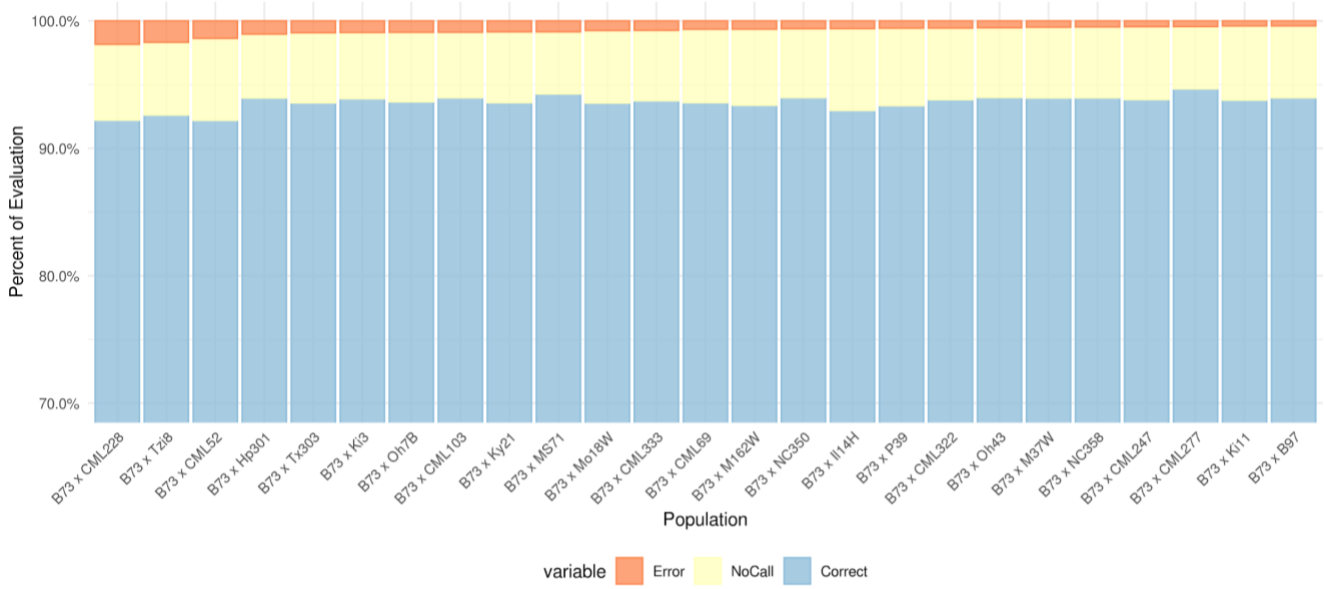


Fig.5a. Evaluation of imputed sites for each NAM RIL over 1107 uplifted legacy SNPs used as a benchmark. Evaluations are grouped by family for simplified viewing. Bars are sorted by percentage of errors. No calls are the product of residual heterozygosity, low density and missing data in benchmark SNPs, or small unresolved breakpoints between the GBS reads.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

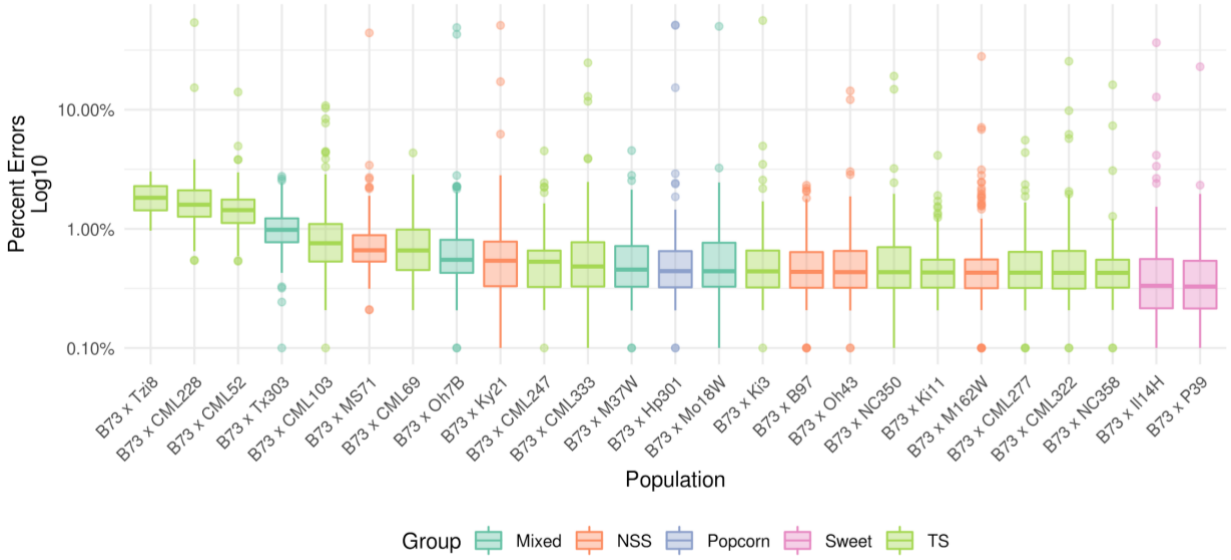


Fig.5b. Per family accession error. Colored by family background. Y-axis is log10 scaled. The presence of a few outliers in individual families does not increase the average family error rate. The background does not affect imputation accuracy.

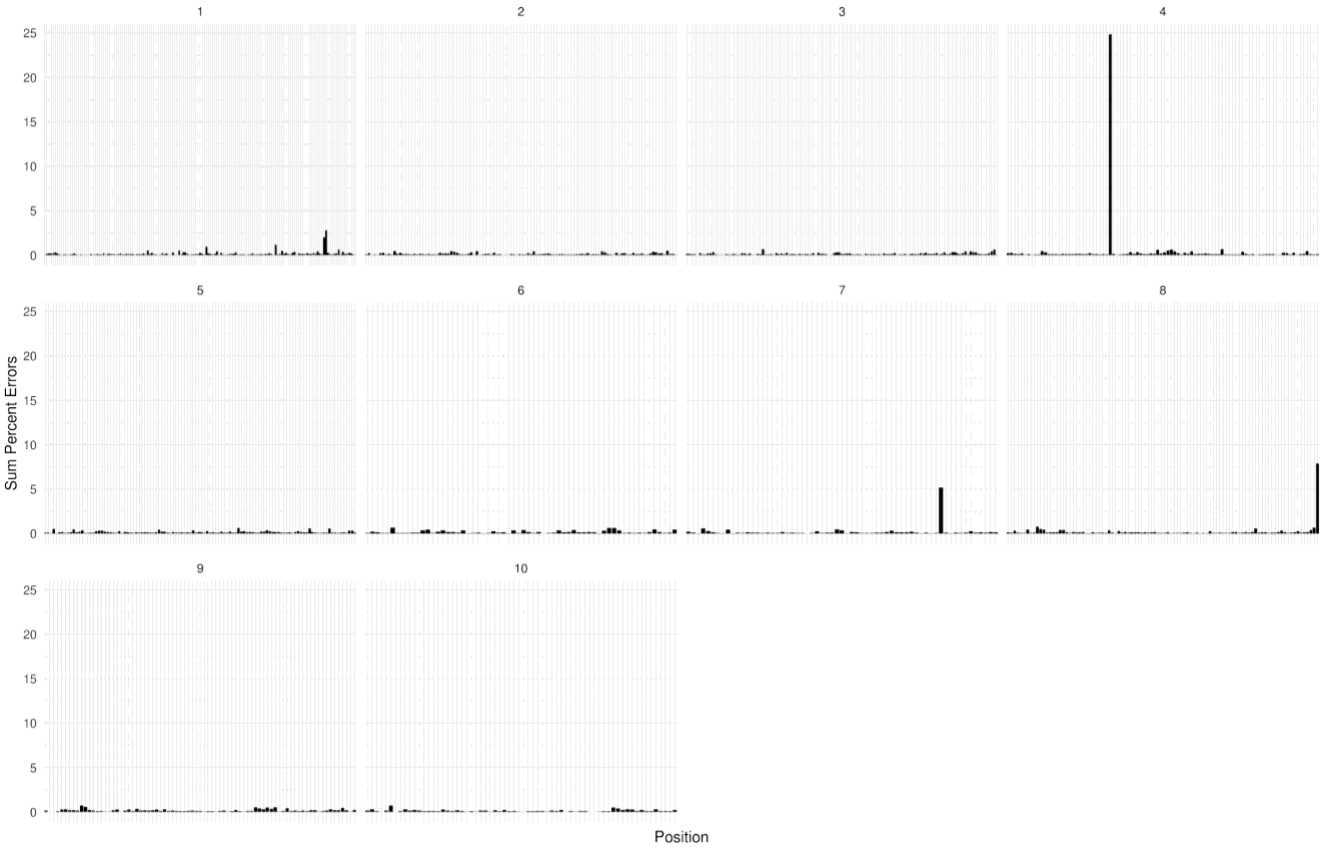


Fig.Sup.1. Sum of percent errors by chromosomal position. Some positions stack up a large proportion of the individually small errors across families (e.g., chromosomes 4, 7, and 8). Not shown here, some positions concentrate a majority of the errors of individual families (e.g., chromosome 6 for CML228, chromosome 10 for CML52)

Imputation of unrelated lines using GBS and WGS

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

To assess the PHG's ability to impute samples indirectly related to the included haplotypes, we tested it on the Goodman Association Panel population, which includes diverse global breeding lines (Flint-Garcia et al., 2005). After mapping both GBS reads and WGS reads for a subset of them, we imputed haplotype paths and called SNPs as with the NAM RILs. The uplifted 600K Axiom SNP array was used as a benchmark; however, this benchmark is biased toward temperate and European diversity (Unterseer et al., 2014).

We first compared the SNPs imputed by the PHG from GBS reads. We were able to identify an average error rate of 3.4% (Fig.6). For taxa represented in the PHG, we saw an average of 0.7% error rate, while for taxa not found in the database, the average error rate was 8.3%. F7 and EP1 had the highest error, at 10% and 12%, respectively. This was not unexpected, as they represent taxa with a European background that is not well represented in the current pangenome (NAM parents plus B104). We saw negligible differences in the accuracies between genic and intergenic regions, having overall average error rates of 3.4% and 3.5%, respectively (data not shown).

To assess the effect of having higher sequencing coverage and depth, we evaluated SNPs imputed by the PHG from WGS reads. We observed a decrease in the proportion of errors to an average of 2.2% (Fig.7), with average error rates of 0.7% and 5.3% for taxa represented or missing from the database, respectively. The average genic and intergenic error rates were 1.9% and 2.5%, respectively. The largest effect, albeit still small compared to when using GBS reads, was on the number of imputed sites; the average percentage of missing sites decreased from 4.5% to 2.2%.

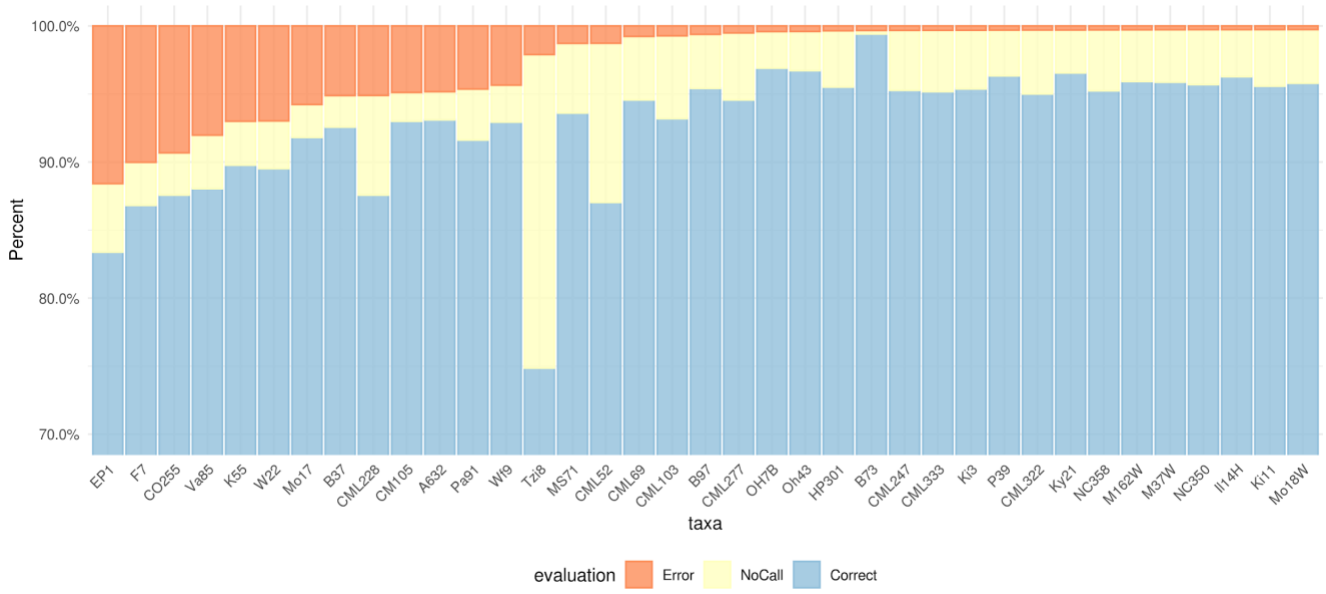


Fig.6. Evaluation of imputation of the Goodman Association Panel population using GBS data. Evaluated against the uplifted 600k Axiom SNP array. Blue and orange show the proportion of correct and erroneous calls, respectively. In yellow are the proportion of SNPs where the benchmark SNP was heterozygous and masked, or where the PHG does not impute that site. Bars are sorted by decreasing error-rate.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

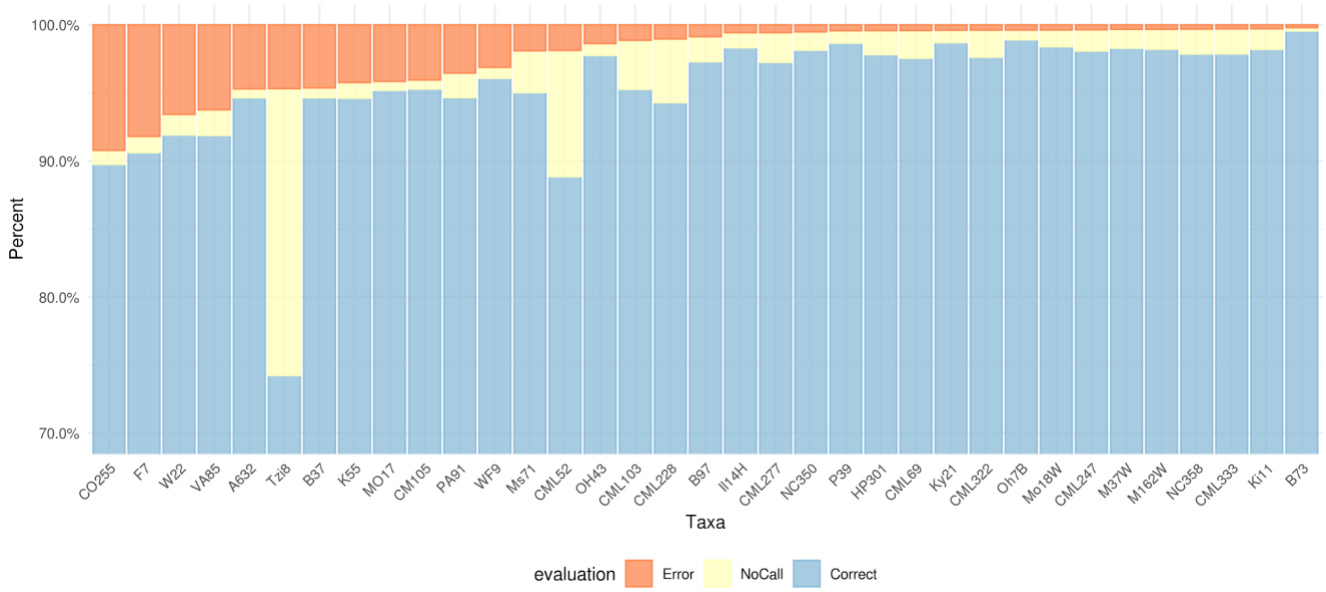


Fig.7. Evaluation of imputation for a subset of the Goodman Association Panel population through WGS data. Evaluated against the uplifted 600k Axiom SNP array. Blue and orange show the proportion of correct and erroneous calls, respectively. In yellow are the proportion of SNPs where the benchmark SNP was heterozygous and masked, or where the PHG had not imputed that site. Bars are sorted by decreasing error-rate.

Assessing causes that influence error rates

To identify the causes driving the errors, we analyzed four potential causes: missing haplotypes, recombination rate, minor allele frequency, and haplotype read counts.

Missing haplotypes

We compared the clustering of the errors across the genome to identify whether errors are due to missing haplotypes or rare alleles (Fig.8). When compared with a set of SNPs with randomized positions, we observed that errors are clustered in longer runs than expected at random. This effect is stronger when imputing taxa not represented in the database. This indicates that the current pangenome is missing haplotypes over a small set of reference ranges, which produce most of the errors. The inclusion of rare alleles not present in the database (~44,000 sites) has no effect (data not shown). The main source of error in the current pipeline is from the absence of important haplotypes, not rare alleles.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

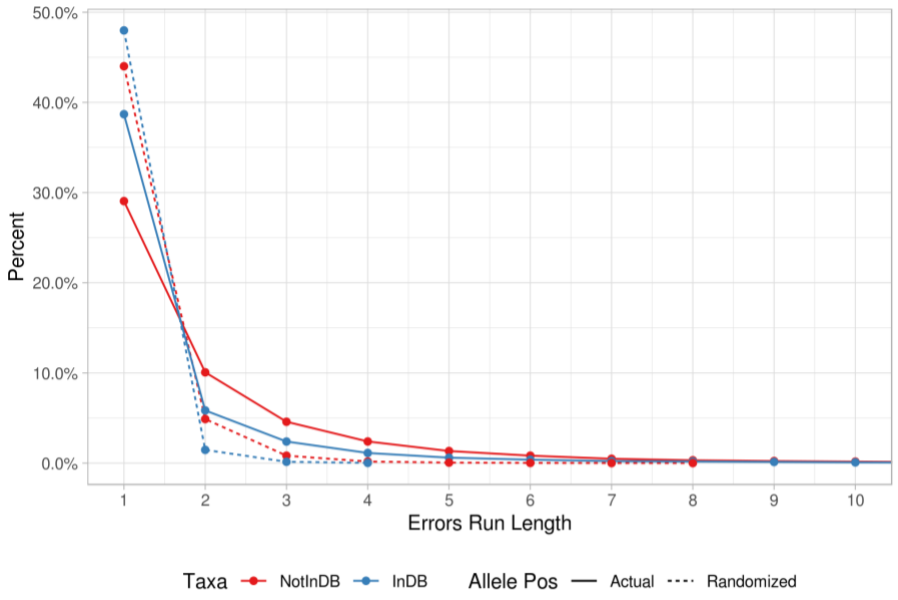


Fig.8. Runs of erroneously imputed sites. Imputing taxa present in the PHG haplotype database permits for longer runs of correctly imputed SNPs, with taxa not present in the PHG database having longer stretches of incorrectly imputed SNPs. Randomized allele positions are sampled randomly from 30% of the sites for each taxon and chromosome. Taxa were grouped by whether they were represented in the pangenome database (blue) or not (red).

Recombination rate

Another potential source of errors arises when samples have recombination within a reference range, resulting in two or more haplotypes. This is likely due to different rates depending on location in the genome, given that that recombination in maize occurs in a frequency that can vary within nearly two orders of magnitude along chromosomes (Rodgers-Melnick et al., 2015). Higher recombination rates should decrease the PHG's ability to represent haplotypes accurately. However, higher recombination rates occur near genic areas. Because gene boundaries were used as haplotype breakpoints, the effect of recombination rate on errors could be minimal. We tested if the recombination rate was correlated with the error rate in 100 equally sized bins of the recombination rate (Fig.9). Error rates for taxa not represented in the PHG database appear correlated with an increased recombination rate. As expected, taxa within the PHG database do not show the same pattern. The inclusion of rare alleles in the analysis did not change the effect. While minor allele frequency (MAF) correlates with error rate (Fig.Sup.2a), MAF is only weakly associated with recombination rates (Fig.Sup.2b). This shows that the errors are also partially driven by novel haplotypes derived from recombination. Increasing the haplotype sampling or decreasing the reference range lengths in high recombination areas could alleviate this problem.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

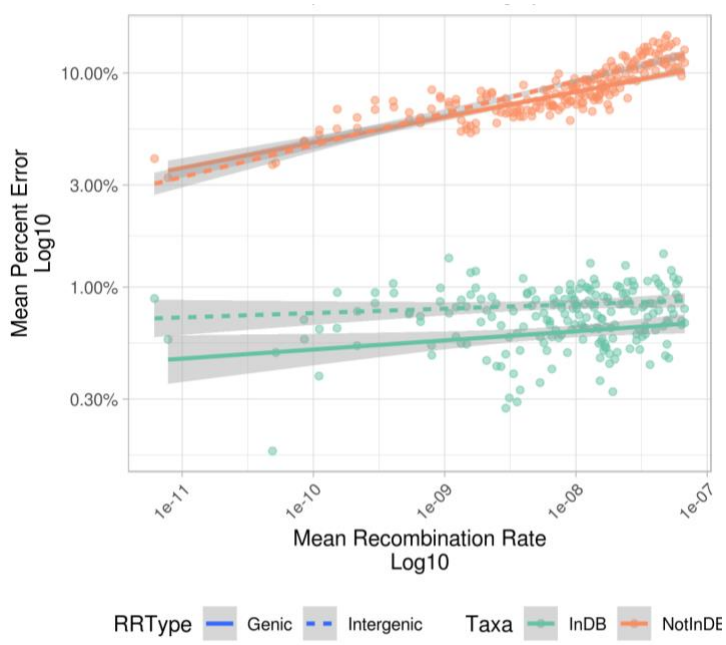
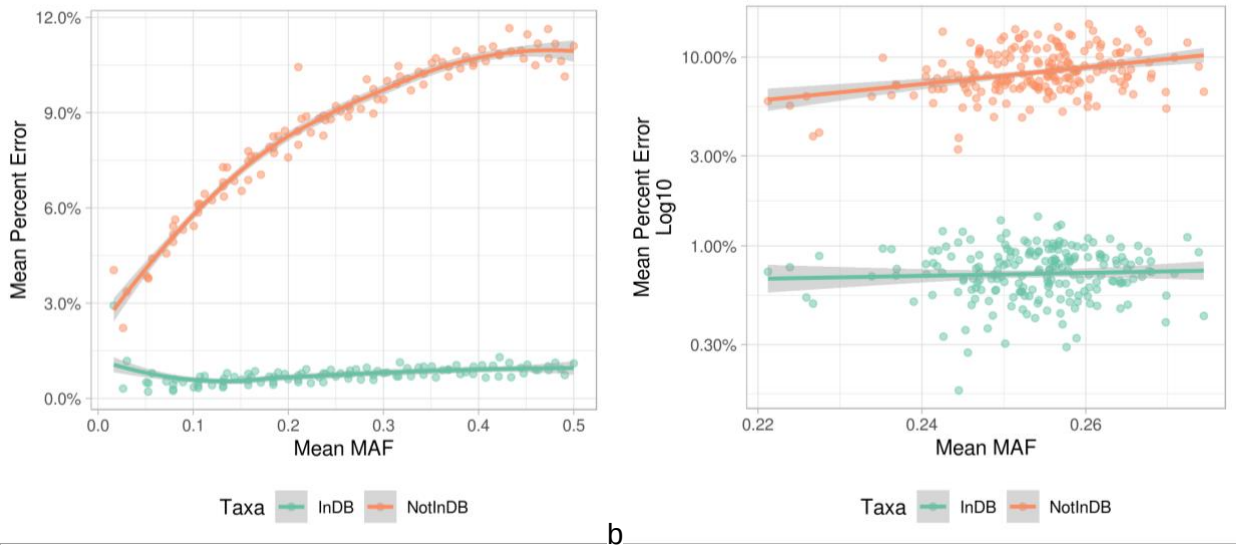


Fig.9. Binned analysis of per-site error rate by recombination rate. Averaged over 100 bins of equal size after the data was sorted by recombination rate. Taxa were grouped by whether they were represented in the pangenome database (green) or not (orange).



a b

Fig.Sup.2. Binned analysis of the per-site error rate. Effect of MAF over error rates. Averaged over 100 bins of equal size after the data was sorted by minor allele frequency (a) or recombination rate (b). Taxa were grouped by whether they were represented in the pangenome database (green) or not (orange).

Haplotype read counts

A large number of duplicated regions in the maize genome complicates accurate read mapping. The PHG pipeline addresses this by keeping read mappings only when they map within a single reference range. The identification of a specific haplotype within a reference range should increase with haplotype coverage. To assess the sensitivity of the PHG to read mapping coverage, we tested the effect of the number of reads mapped to the error rates on the imputed haplotypes over 100 bins of

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

increasing mean number of mappings (Fig.10). Error rates for taxa not in the PHG database decrease slightly with an increasing mean number of reads. As expected, this effect is smaller for taxa in the database. In both cases, the effect was higher in genic ranges.

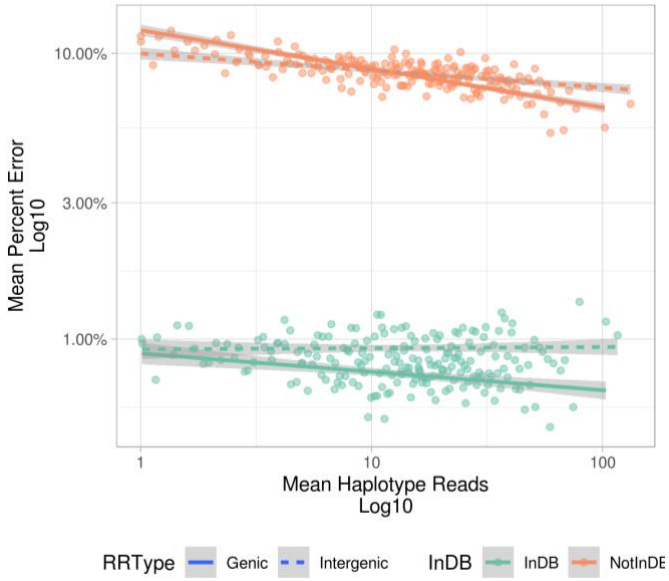


Fig.10. Per-haplotype error rate by mean haplotype read count. Binned haplotype error rates were calculated over 100 equally-sized bins of an increasing number of reads mapped. The analysis is shown separately by genic or intergenic reference range type and taxa grouped by whether they were represented in the pangenome database (green) or not (orange).

With the sequence context for over 40 million SNPs, this maize PHG database occupies a modest 83GB of disk space, being effectively only slightly larger than the genomes it represents. For comparison, a VCF file for the same taxa number, without haplotype data or mapping information, requires over 1TB of space. By leveraging the diverse NAM founders and their high-quality assemblies through the PHG, we have shown its utility in imputing related and unrelated samples. While the current database cannot perform as accurately on accessions with unrepresented haplotypes, we expect the future addition of a relatively small set of assemblies will address this issue. We expect the maize PHG will enable the community to more efficiently and effectively identify the genetic mechanisms underlying the control of many more phenotypes.

Table 1: Populations genotyped by the PHG

Population	Inbreeding	Diversity	Sequencing Depth
NAM RILs	High	Low	Low
Goodman Association Panel GBS	High	Moderate	Low
Goodman Association Panel WGS	High	Moderate	High

Table 2: Summary Main Genotyping results

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Population Panel		NAM - 1k legacy SNPs	Goodman Association Panel - 600k Axiom SNP chip	
Taxa in PHG database\Data type		GBS	GBS	WGS
Average PHG benchmark agreement*	In DB	99.2%	99.3%	99.3%
	Not in DB		91.7%	94.3%

*Considering only imputed and not heterozygous sites in the benchmark.

Discussion

Here we have presented a first maize PHG, a pangenome haplotype database, and a genotyping pipeline that allows the accurate imputation and reconstruction of whole-genome haplotypes and genotyping with minimal information. By parting from the concept of a closed pangenome, where a limited number of taxa can represent virtually all of the haplotypes found in a species, we leveraged the NAM assemblies in the PHG to create a novel genotyping and imputation tool. The maize PHG will thus facilitate community genotyping efforts by achieving higher accuracy of imputation and implementing a standardized approach that will return consistent results from low coverage sequencing.

Processing the assemblies through the PHG pipeline allows us to identify homologous haplotypes among the included assemblies for most defined reference ranges. Previous analyses of variation in intergenic (Anderson et al., 2019) and genic regions (Sun et al., 2018) show a relatively similar pattern to the results observed in this maize PHG database, having regions essentially shared at ~80% rate, with a relatively small subset of them being mostly missing in the majority of taxa. This is only a first approach, as potentially a more complete gene model set, resulting in smaller reference ranges, might better reflect the presence/absence variation of more refined haplotypes. Additionally, choosing a distinct base reference that more closely resembles the larger pangenome might also allow for the more direct dissection of otherwise less common haplotypes.

By imputing haplotypes known to be within our database, as shown with the NAM RILs, we demonstrate the pangenome's ability to act as a useful target for read mapping. Moreover, the pipeline functions as an effective tool to process these mappings and accurately identify the underlying haplotypes from which those reads are derived. This is of particular interest for the maize and general plant communities, due to the frequently high repetitive nature of their genomes, where the unambiguous mapping of short sequences continues to be a challenge for accurate and large scale genotyping.

Assessing the genotyping accuracy on the diverse Goodman Association Panel population, we also show that this current maize PHG database, while effective, is missing a relatively small but significant portion of the broader maize pangenome. As opposed to rare alleles, this lack of haplotypes was the main source of errors, as evidenced by the clustering of errors in a subset of the haplotypes and over those regions with a high recombination rate. This is particularly the case for samples of European backgrounds, which are not well represented in the implemented database. We expect an increase in haplotype diversity within the PHG database, by including additional diverse genomes such as those

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

from CIMMYT maize lines, will allow the PHG to be a one-stop solution for the accurate genotyping of most maize lines.

This first maize PHG database consists of a pangenome haplotype database from 27 high-quality maize assemblies. It includes the information on the mapping and imputation of over 5,000 NAM RILs and Goodman Association Panel accessions. We expect that this maize database, along with the broader PHG pipeline, through its ability to generate accurate genotype calls using either GBS or WGS reads, small computational footprint, and ease of transferability will help the maize community achieve accurate and inexpensive genotyping, enabling more analyses to discover the genetic basis of many more phenotypic traits.

Materials and Methods

Building the PHG

We populated a maize PHG database (Bradbury, 2020) to store the pangenome's information and keep track of reference ranges, haplotypes, mappings, and paths imputed. Briefly, the PHG is a relational database that divides the reference genome in ranges, subdivides other pangenomes in similar ranges, and allows for rapid and efficient storage and retrieval of information about the reference ranges, component haplotype IDs, and paths. A Java and R API have been developed for the PHG software package to implement and interact with this pangenome database (Bradbury, 2020; Bradbury et al., 2007).

To populate the maize PHG database, we generated a pangenome by leveraging the high-quality assemblies for the 26 diverse NAM parents (*NAM Genomes Project*, 2020) and B104 (reference TBA). To define the pangenome haplotypes, we took a three-stage approach. First, we selected the B73 RefGen_v5 (*MaizeGDB*, 2019) as the base reference. This allows for a direct comparison with existing genotyping platforms and results. Second, we defined the haplotype regions by choosing a set of reference ranges. We selected the B73 RefGen_v5 gene regions (Zm00001e.1) to allow for haplotype boundaries that are: clearly defined, of biological significance, and more likely to be conserved. We used the 35,677 genic regions as breakpoints for the definition of the 71,354 B73 haplotypes. Third, we used Mummer4 (Marçais et al., 2018) to identify the haplotypes in those regions on each assembly. Briefly, the assemblies were divided into individual chromosomes. Each chromosome was aligned against the B73 RefGen_v5 equivalent using the nucmer program of Mummer4. Testing values for the -c parameter between 150-500, we set the value to 250, to find a balance between speed and the length of maize exons (Haberer et al., 2005; *MaizeGDB*, 2020a). These alignments were then processed, with each reference range in each assembly having a haplotype ID assigned and stored in the database. Insertions or deletions found within each reference range are stored as part of the identified equivalent haplotype, regardless of their size. Reference ranges for taxa that do not produce an alignment are left empty (Bradbury, 2020).

Alignment to Pangenome

To evaluate the ability of the PHG to impute haplotypes, we utilized the GBS reads originally generated for the NAM (Rodgers-Melnick et al., 2015) and the Goodman Association Panel (Romay et al., 2013). Additionally, WGS paired-end reads (Bukowski et al., 2018) were obtained for a subset of the Goodman Association Panel samples (SRA study accession number SRP108889), which had also been genotyped with the Axiom 600k SNP array (Unterseer et al., 2014). This allows us to compare

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

the ability of the PHG to impute haplotypes to a distinct and diverse set of sequencing data. CO125 was removed from the analysis as its genotyping data suggests a sample mixup in its sequencing.

The reads were mapped to an index of the pangenome generated by minimap2 (Li, 2018) (Ver. 2.17-r941). We set parameters -k 21 -w 11 -I 90G to reflect the recommended short read alignment kmer size and the necessity of having the complete pangenome sequence loaded into memory at once. Failing to set -I large enough to fit the whole pangenome in RAM returns a multi-part index, which produces poor mapping processing results. The read mappings made use of the short read heuristics (-k21 -w11 --sr --frag=yes -A2 -B8 -O12,32 -E2,1 -r50 -p.5 -N25 -f1000,5000 -n2 -m20 -s40 -g200 -2K50m --heap-sort=yes). To maximize the likelihood of getting read mappings to all matching haplotypes, we modified the flag -N to produce 25 secondary mappings. The mappings were then processed through and stored in the PHG database. Briefly, only edit distance optimal alignments are considered, reads that map to multiple reference ranges are discarded, and reads that map to specific haplotypes are identified among the read mappings for the haplotypes within the reference range.

Imputation evaluation

Once the read mappings are loaded into the PHG database, imputation is done by finding a path through the haplotype graph using the BestHaplotypePathPlugin in the Java PHG API. Briefly, the read mappings are used to count the number of times reads mapped to each haplotype. Using these counts, and the transition probabilities between adjacent haplotypes, an HMM algorithm finds the most likely haplotype path through the graph. The parameter minReads is set to 0 so that the algorithm imputes haplotypes for all reference ranges, including those that have no reads mapping to them. The resulting imputed paths are stored in the PHG database. Finally, all SNPs within imputed haplotypes are exported as a VCF file (PathsToVCFPlugin). This generates a 1TB VCF with >42M sites for the 4,705 NAM accessions. A similar process was followed to impute and generate the SNP calls for the Goodman Association Panel's taxa.

An SNP benchmark set was defined for the NAM RILs and Goodman Association Panel population samples to evaluate imputation accuracy. For the NAM population, the 1,144 legacy SNP set (McMullen et al., 2009), NAM_map_and_genos-120731.zip, was obtained (*Panzea*, 2009) and uplifted from AGPv2 to v5 in a two-step approach. First, from the original v2 to v4, we used CrossMap (Zhao et al., 2014) and the corresponding chain file (*ENSEMBL*, 2020). These were then uplifted from v4 to v5 using the liftover_vcf pipeline (https://github.com/qisun2/liftover_vcf) and a v4 to v5 chain file (*MaizeGDB*, 2020b). This returned 1,106 variant sites uplifted to v5 coordinates, which we then used to compare our imputation results. For the Goodman Association Panel, the 600 K (Unterseer et al., 2014) Axiom SNP array was uplifted from AGPv4 to v5 coordinate equivalents through Crossmap and utilized as a benchmark.

Custom code was written to evaluate the accuracy of the imputed SNP calls by comparing them to the benchmarks. In short, the imputed VCF files were intersected with the Axiom SNP sites using bedtools (Quinlan & Hall, 2010). These VCF files were then loaded into R (R Core Team, 2018) through the SNPRelate package (Zheng et al., 2012). Heterozygous genotypes in the benchmark were set as missing, as the haplotypes are expected to be homozygous, and the current pipeline imputes homozygous SNPs. Then, the two sets of SNPs were matched to the equivalent taxa, and the correspondence of reference or alternate allele calls were evaluated as correct or incorrect if they agreed or not. Error rates are calculated as the number of incorrect allele calls divided by the total number of correct *and* incorrect calls. For sites on taxa where the PHG makes no allele call, or where

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

the benchmark has a heterozygous allele, the evaluation is set as NoCall. To identify the proportion of errors between genic and intergenic regions, the GenomicRanges package (Lawrence et al., 2013) was used to identify the SNPs as found within either set of regions, and the data.table (Dowle & Srinivasan, 2019) package was then used to summarize each region type's calls.

Assessing causes that influence the error rate

Evaluating runs of errors

SNP evaluations sorted by position were processed through the rle function (R Core Team, 2018) of R to get the run-length of error calls on each of the chromosomes for each taxon. These runs of errors were then analyzed by whether the taxa were represented in the pangenome database or not. A random set of 30% of the SNP evaluations for each taxon and chromosome were selected, effectively randomizing the SNP positions, on which we then calculated the run-length of errors. The frequency of each of the run-length of errors was then calculated for each category.

Comparison of error rate vs. recombination rate

Recombination rates on the NAM population were obtained from (Ramstein et al., 2020). These were uplifted from v4 to v5 through Crossmap. The uplifted values were then matched to the evaluated SNP calls. This combined data set was then sorted by the recombination rate. The mean recombination and error rate were calculated over 100 bins of equal size number of SNPs.

Comparison of error rate vs. minor allele frequency

Minor allele frequencies were obtained from the Axiom 600k SNP array from the taxa under analysis through the snpgdsSNPRateFreq of the SNPRelate package. These frequencies were matched to the evaluated SNP calls. One hundred bins of equal length were then created after sorting by increasing recombination rate or minor allele frequency. The means for error rate, recombination rate, and MAF were then calculated on each bin.

Comparison of error rate vs. read counts

Imputed haplotypes for each sample were obtained from the PHG database through the pathsForMethod function of the rPHG package (Monier et al., 2019). For each taxon, read mappings were obtained from the PHG database through the readMappingsForLineName, also of the rPHG package. The mappings were then matched to the haplotypes imputed for each taxon. The mean error rate and mean read count were calculated for the haplotypes over each reference range by the sample representation status in the pangenome database and by reference range type. The calculated values over these four categories were then sorted by the mean number of reads and analyzed over 100 bins of equal size. The mean number of reads and the mean error rate were calculated for each bin.

The maize PHG database is publicly available through the Buckler Lab webpage:
https://www.maizegenetics.net/post/the-first-maize-phg-database-now-available
The code for these analyzes is mostly written in R and made available at:
https://bitbucket.org/bucklerlab/p_maizephg

Acknowledgments

This material is based upon work supported by the USDA-ARS, NSF Research-PGR Grant No. IOS-1822330, NSF Postdoctoral Research Fellowship in Biology under Grant No. IOS-1906619, Bill and Melinda Gates Foundation, and a CONACYT-I2T2 scholarship for graduate studies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

References

Albert, P. S., Gao, Z., Danilova, T. V., & Birchler, J. A. (2010). Diversity of chromosomal karyotypes in maize and its relatives. *Cytogenetic and Genome Research*, *129*(1-3), 6–16. https://doi.org/10.1159/000314342

Anderson, S. N., Stitzer, M. C., Brohammer, A. B., Zhou, P., Noshay, J. M., O'Connor, C. H., Hirsch, C. D., Ross-Ibarra, J., Hirsch, C. N., & Springer, N. M. (2019). Transposable elements contribute to dynamic genome content in maize. *The Plant Journal: For Cell and Molecular Biology*, *100*(5), 1052–1065. https://doi.org/10.1111/tpj.14489

Bradbury, P. J. (2020, February 2). *PHG Wiki*. PHG Repository. https://bitbucket.org/bucklerlab/practicalhaplotypegraph/wiki/Home

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. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, *194*(2), 459–471. https://doi.org/10.1534/genetics.113.150029

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., … McMullen, M. D. (2009). The genetic architecture of maize flowering time. *Science* , *325*(5941), 714–718. https://doi.org/10.1126/science.1174276

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C., Fan, L., Gao, S., Xu, X., Zhang, G., Li, Y., Jiao, Y., Doebley, J. F., Ross-Ibarra, J., Lorant, A., … Xu, Y.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

(2018). Construction of the third-generation Zea mays haplotype map. *GigaScience*, *7*(4), 1–12. https://doi.org/10.1093/gigascience/gix134

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Chan, A. W., Hamblin, M. T., & Jannink, J. L. (2016). Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data. *PloS One*, *11*(8), 1–17. https://doi.org/10.1371/journal.pone.0160733

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Dowle, M., & Srinivasan, A. (2019). *data.table: Extension of `data.frame`*. https://CRAN.R-project.org/package=data.table

ENSEMBL. (2020, March 4). ftp://ftp.ensemblgenomes.org/pub/plants/release-47/assembly_chain/zea_mays/

FAO. (2018). Food and Agriculture Organization of the United Nations Agriculture Databases. http://www.fao.org/statistics/databases/en/

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., & Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal: For Cell and Molecular Biology*, *44*(6), 1054–1064. https://doi.org/10.1111/j.1365-313X.2005.02591.x

Fu, H., & Dooner, H. K. (2002). Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9573–9578. https://doi.org/10.1073/pnas.132259199

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Gage, J. L., Monier, B., Giri, A., & Buckler, E. S. (2020). Ten Years of the maize Nested Association Mapping Population: Impact, Limitations, and Future Directions. *The Plant Cell*. https://doi.org/10.1105/tpc.19.00951

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–879. https://doi.org/10.1038/nbt.4227

Gaut, B. S., & Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 6809–6814. https://doi.org/10.1073/pnas.94.13.6809

Gouesnard, B., Negro, S., Laffray, A., Glaubitz, J., Melchinger, A., Revilla, P., Moreno-Gonzalez, J., Madur, D., Combes, V., Tollon-Cordet, C., Laborde, J., Kermarrec, D., Bauland, C., Moreau, L., Charcosset, A., & Nicolas, S. (2017). Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *130*(10), 2165–2189. https://doi.org/10.1007/s00122-017-2949-6

Haberer, G., Young, S., Bharti, A. K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R. A., Rounsley, S., Birren, B., Nusbaum, C., Mayer, K. F. X., & Messing, J. (2005). Structure and architecture of the maize genome. *Plant Physiology*, *139*(4), 1612–1624. https://doi.org/10.1104/pp.105.068718

Hake, S., & Ross-Ibarra, J. (2015). Genetic, evolutionary and plant breeding insights from the domestication of maize. *eLife*, *4*. https://doi.org/10.7554/eLife.05861

Hickey, J. M., Kinghorn, B. P., Tier, B., van der Werf, J. H. J., & Cleveland, M. A. (2012). A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genetics, Selection, Evolution: GSE*, *44*, 9. https://doi.org/10.1186/1297-9686-44-9

Jensen, S. E., Charles, J. R., Muleta, K., Bradbury, P. J., Casstevens, T., Deshpande, S. P., Gore, M. A., Gupta, R., Ilut, D. C., Johnson, L., Lozano, R., Miller, Z., Ramu, P., Rathore, A., Romay, M. C.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Upadhyaya, H. D., Varshney, R. K., Morris, G. P., Pressoir, G., … Ramstein, G. P. (2020). A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *The Plant Genome*, *13*(1), 1687. https://doi.org/10.1002/tpg2.20009

Lai, J., Li, Y., Messing, J., & Dooner, H. K. (2005). Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(25), 9068–9073. https://doi.org/10.1073/pnas.0502923102

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., & Carey, V. (2013). Software for Computing and Annotating Genomic Ranges. In *PLoS Computational Biology* (Vol. 9). https://doi.org/10.1371/journal.pcbi.1003118

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

MaizeGDB. (2019, November 22). Maize Genetics and Genomics Database. https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/Zm-B73-REFERENCE-NAM-5.0.fa.gz

MaizeGDB. (2020a). Maize Genetics and Genomics Database. https://www.maizegdb.org/assembly

MaizeGDB. (2020b, May 2). Maize Genetics and Genomics Database. https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/chain_files/

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., … Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., … Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science*, *325*(5941), 737–740. https://doi.org/10.1126/science.1174320

Monier, B., Bradbury, P., Casstevens, T., Jannink, J.-L., & Buckler, E. (2019). *rPHG: R front-end for the practical haplotype graph*. https://bitbucket.org/bucklerlab/rphg/src/master/

NAM Genomes Project. (2020). Whole-Genome Assembly of the Maize NAM Founders. https://nam-genomes.org/

Oliver, K. R., McComb, J. A., & Greene, W. K. (2013). Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biology and Evolution*, *5*(10), 1886–1901. https://doi.org/10.1093/gbe/evt141

Panzea. (2009). www.panzea.org

Paterson, A. H., Bowers, J. E., & Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9903–9908. https://doi.org/10.1073/pnas.0307901101

Pook, T., Mayer, M., Geibel, J., Weigend, S., Cavero, D., Schoen, C. C., & Simianer, H. (2019). Improving Imputation Quality in BEAGLE for Crop and Livestock Data. *G3: Genes|Genomes|Genetics*, g3.400798.2019. https://doi.org/10.1534/g3.119.400798

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramstein, G. P., Larsson, S. J., Cook, J. P., Edwards, J. W., Ersoz, E. S., Flint-Garcia, S., Gardner, C. A., Holland, J. B., Lorenz, A. J., McMullen, M. D., Millard, M. J., Rocheford, T. R., Tuinstra, M. R., Bradbury, P. J., Buckler, E. S., & Romay, M. C. (2020). Dominance Effects and Functional Enrichments Improve Prediction of Agronomic Traits in Hybrid Maize. *Genetics*, *215*(1), 215–230. https://doi.org/10.1534/genetics.120.303025

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Statistical Computing. https://www.R-project.org/

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., Li, C., Li, Y., & Buckler, E. S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3823–3828. https://doi.org/10.1073/pnas.1413864112

Romay, M. C. (2018). Rapid, Affordable, and Scalable Genotyping for Germplasm Exploration in Maize. In J. Bennetzen, S. Flint-Garcia, C. Hirsch, & R. Tuberosa (Eds.), *The Maize Genome* (pp. 31–46). Springer International Publishing. https://doi.org/10.1007/978-3-319-97427-9_3

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., Elshire, R. J., Acharya, C. B., Mitchell, S. E., Flint-Garcia, S. A., McMullen, M. D., Holland, J. B., Buckler, E. S., & Gardner, C. A. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, *14*(6), R55. https://doi.org/10.1186/gb-2013-14-6-r55

Romero Navarro, J. A., Wilcox, M., Burgueño, J., Romay, C., Swarts, K., Trachsel, S., Preciado, E., Terron, A., Delgado, H. V., Vidal, V., Ortega, A., Banda, A. E., Montiel, N. O. G., Ortiz-Monasterio, I., Vicente, F. S., Espinoza, A. G., Atlin, G., Wenzl, P., Hearne, S., & Buckler, E. S. (2017). A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nature Genetics*, *2017*(April 2016). https://doi.org/10.1038/ng.3784

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., Song, W., Zhang, M., Cui, Y., Dong, X., Liu, H., Ma, X., Jiao, Y., Wang, B., Wei, X., Stein, J. C., Glaubitz, J. C., Lu, F., Yu, G., … Lai, J. (2018). Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nature Genetics*, *50*(9), 1289–1295. https://doi.org/10.1038/s41588-018-0182-0

Swarts, K., Li, H., Romero Navarro, J. A., An, D., Romay, M. C., Hearne, S., Acharya, C., Glaubitz, J. C., Mitchell, S., Elshire, R. J., Buckler, E. S., & Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome*, *7*(3), 0. https://doi.org/10.3835/plantgenome2014.05.0023

Telenti, A., Pierce, L. C. T., Biggs, W. H., di Iulio, J., Wong, E. H. M., Fabani, M. M., Kirkness, E. F., Moustafa, A., Shah, N., Xie, C., Brewerton, S. C., Bulsara, N., Garner, C., Metzker, G., Sandoval,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

E., Perkins, B. A., Och, F. J., Turpaz, Y., & Venter, J. C. (2016). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(42), 11901–11906. https://doi.org/10.1073/pnas.1613365113

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T. M., Fries, R., Pausch, H., Bertani, C., Davassi, A., Mayer, K. F., & Schön, C.-C. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*, *15*, 823. https://doi.org/10.1186/1471-2164-15-823

Wang, B., Lin, Z., Li, X., Zhao, Y., Zhao, B., Wu, G., Ma, X., Wang, H., Xie, Y., Li, Q., Song, G., Kong, D., Zheng, Z., Wei, H., Shen, R., Wu, H., Chen, C., Meng, Z., Wang, T., … Wang, H. (2020). Genome-wide selection and genetic improvement during modern maize breeding. *Nature Genetics*, *52*(6), 565–571. https://doi.org/10.1038/s41588-020-0616-3

Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., Sudha, N., Olsen, M., Prasanna, B. M., Zhang, X., & Babu, R. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *129*(4), 753–765. https://doi.org/10.1007/s00122-016-2664-8

Xu, C., Ren, Y., Jian, Y., Guo, Z., Zhang, Y., Xie, C., Fu, J., Wang, H., Wang, G., Xu, Y., Li, P., & Zou, C. (2017). Development of a maize 55 K SNP array with improved genome coverage for molecular breeding. *Molecular Breeding: New Strategies in Plant Improvement*, *37*(3), 20. https://doi.org/10.1007/s11032-017-0622-z

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., & Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* , *30*(7), 1006–1007. https://doi.org/10.1093/bioinformatics/btt730

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, C., Karn, A., Reisch, B., Nguyen, A., Sun, Y., Bao, Y., Campbell, M. S., Church, D., Williams, S.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.31.268425; this version posted September 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Xu, X., Ledbetter, C. A., Patel, S., Fennell, A., Glaubitz, J. C., Clark, M., Ware, D., Londo, J. P., Sun, Q., & Cadle-Davidson, L. (2020). Haplotyping the Vitis collinear core genome with rhAmpSeq improves marker transferability in a diverse genus. *Nature Communications*, *11*(1), 413. https://doi.org/10.1038/s41467-019-14280-1