

1 **Development of the Wheat Practical Haplotype Graph Database as a Resource for**
2 **Genotyping Data Storage and Genotype Imputation**

3 Katherine W. Jordan^{1,2#}, Peter J. Bradbury³, Zachary R. Miller⁴, Moses Nyine¹, Fei He¹, Max
4 Fraser⁵, Jim Anderson⁵, Esten Mason⁶, Andrew Katz⁶, Stephen Pearce⁶, Arron H. Carter⁷,
5 Samuel Prather⁷, Michael Pumphrey⁷, Jianli Chen⁸, Jason Cook⁹, Shuyu Liu¹⁰, Jackie C. Rudd¹⁰,
6 Zhen Wang¹⁰, Chenggen Chu¹⁰, Amir M. H. Ibrahim¹⁰, Jonathan Turkus¹¹, Eric Olson¹¹,
7 Ragupathi Nagarajan¹², Brett Carver¹², Liuling Yan¹², Ellie Taagen⁴, Mark Sorrells⁴, Brian
8 Ward¹³, Jie Ren^{1,14}, Alina Akhunova^{1,14}, Guihua Bai², Robert Bowden², Jason Fiedler¹⁵, Justin
9 Faris¹⁵, Jorge Dubcovsky¹⁶, Mary Guttieri², Gina Brown-Guedira¹³, Ed Buckler³, Jean-Luc
10 Jannink³, Eduard D. Akhunov^{1*}

11 ¹ Department of Plant Pathology, Kansas State University, Manhattan, KS, USA
12 ² USDA-ARS, Hard Winter Wheat Genetics Research Unit, Manhattan, KS, USA
13 ³ USDA-ARS, Plant Soil and Nutrition Research Unit, Ithaca, NY, USA
14 ⁴ Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA
15 ⁵ Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN, USA
16 ⁶ Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA
17 ⁷ Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA
18 ⁸ Department of Plant Sciences, University of Idaho, Aberdeen, ID, USA
19 ⁹ Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT,
20 USA
21 ¹⁰ Department of Soil and Crop Sciences, Texas A&M AgriLife Reseach, Amarillo, TX, USA
22 ¹¹ Department of Plan, Soil and Microbial Sciences, Michigan State University, East Lansing,
23 MI, USA
24 ¹² Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, OK, USA
25 ¹³ USDA-ARS, Plant Science Research Unit, Raleigh, NC, USA
26 ¹⁴ Integrative Genomics Facility, Kansas State University, Manhattan, KS, USA
27 ¹⁵ USDA-ARS, Cereal Crops Research Unit, Fargo, ND, USA
28 ¹⁶ Department of Plant Sciences, University of California-Davis, Davis, CA, USA
29

30 Running Title: Wheat Practical Haplotype Graph

31 Keywords: Wheat, Genotype Imputation, Practical Haplotype Graph, skim-seq, exome capture

32 Corresponding Author: Eduard Akhunov, Department of Plant Pathology, Kansas State
33 University, 1712 Claflin Rd, 4024 Throckmorton Plant Science Center, Manhattan, KS 66506,
34 eakhunov@ksu.edu

35 #KWJ is currently affiliated with USDA-ARS, Hard Winter Wheat Genetics Research Unit,
36 Manhattan, KS, USA.

37

38

39 **Abstract**

40 To improve the efficiency of high-density genotype data storage and imputation in bread wheat
41 (*Triticum aestivum* L.), we applied the Practical Haplotype Graph (PHG) tool. The wheat PHG
42 database was built using whole-exome capture sequencing data from a diverse set of 65 wheat
43 accessions. Population haplotypes were inferred for the reference genome intervals defined by
44 the boundaries of the high-quality gene models. Missing genotypes in the inference panels,
45 composed of wheat cultivars or recombinant inbred lines genotyped by exome capture,
46 genotyping-by-sequencing (GBS), or whole-genome skim-seq sequencing approaches, were
47 imputed using the wheat PHG database. Though imputation accuracy varied depending on the
48 method of sequencing and coverage depth, we found 93% imputation accuracy with 0.01x
49 sequence coverage, which was only slightly lower than the accuracy obtained using the 0.5x
50 sequence coverage (96.9%). Compared to Beagle, on average, PHG imputation was ~4% (p -
51 $value = 0.00027$) more accurate, and showed 27% higher accuracy at imputing a rare haplotype
52 introgressed from a wild relative into wheat. The reduced accuracy of imputation with GBS data
53 (90.4%) is likely associated with the small overlap between GBS markers and the exome capture
54 dataset, which was used for constructing PHG. The highest imputation accuracy was obtained
55 with exome capture for the wheat D genome, which also showed the highest levels of linkage
56 disequilibrium and proportion of identity-by-descent regions among accessions in our reference
57 panel. We demonstrate that genetic mapping based on genotypes imputed using PHG identifies
58 SNPs with a broader range of effect sizes that together explain a higher proportion of genetic
59 variance for heading date and meiotic crossover rate compared to previous studies.

60

61 **Introduction**

62 For the last 10,000 years, intensive selection of bread wheat, *Triticum aestivum*, created
63 varieties adapted to diverse environments and cultivation practices (Balfourier *et al.* 2019; He *et*
64 *al.* 2019; Walkowiak *et al.* 2020). Recent advances in crop genomics and the availability of
65 reference genomes have accelerated the adoption of sequence-based genotyping technologies for
66 studying the genetics of agronomic traits (Nyine *et al.* 2019) and local adaptation (He *et al.* 2019;
67 Juliana *et al.* 2019, 2020) and facilitated the introduction of genomics-assisted breeding
68 strategies into wheat improvement pipelines (Poland and Rife 2012; Isidro *et al.* 2014).
69 However, the limited genome coverage provided by these genotyping technologies does not
70 support exploration of the entire range of genetic effects conferred by all variants, limiting the
71 utility of the developed genomic diversity and functional genomics resources for understanding
72 genome-to-phenome connections.

73 The large size (17 Gb) and complexity of the wheat genome present a substantial
74 challenge for sequence-based analysis of genetic diversity. Alignment of short sequence reads to
75 the wheat genome is complicated by high levels of sequence redundancy resulting from two
76 rounds of recent whole genome duplication (IWGSC, 2018), and the recent propagation of
77 transposable elements (TEs) comprising nearly 90% of the genome (Wicker *et al.* 2018).
78 Therefore, the efforts of the wheat research community were focused primarily on sequencing
79 complexity-reduced genomic libraries produced by either enzymatic digests or by targeted
80 sequence capture. These efforts have resulted in a detailed description of the population-scale
81 haplotypic diversity in the low-copy genomic regions in large sets of genetically and
82 geographically diverse wheat lines and breeding populations (He *et al.* 2019; Juliana *et al.* 2019;
83 Pont *et al.* 2019). While these resources have been useful for genotype imputation in populations

84 genotyped using either SNP-based arrays or genotyping-by-sequencing (GBS) methods (Jordan
85 *et al.* 2015; Shi *et al.* 2017; Juliana *et al.* 2019; Nyine *et al.* 2019), the relatively small number of
86 shared markers between the reference and inference populations limits the number of imputed
87 genotypes, thus diminishing the utility of genotype imputation in wheat genetic studies and
88 breeding.

89 High-quality reference genomes and a reduction in the cost of sequencing presented
90 opportunities for the characterization of genetic diversity by direct sequencing of either whole
91 genomes or genomic regions targeted by sequence capture (Malmberg *et al.* 2018; He *et al.*
92 2019; Walkowiak *et al.* 2020). While these sequence-based genotyping approaches generate
93 unbiased information about the genetic variants of various frequency classes and genomic
94 locations, large-scale population sequencing of species with large genomes, including many
95 important agricultural crops, remains costly. This issue has been addressed by combining low-
96 coverage sequencing of whole genomes with the prediction of missing genotypes using
97 imputation tools, thereby increasing the power of association mapping and facilitating the
98 detection of causal variants (Davies *et al.* 2016; Das *et al.* 2018; Rubinacci *et al.* 2021).

99 Recently, a novel strategy referred to as Practical Haplotype Graph (PHG), was proposed
100 to improve the efficiency of sequence-based genotyping data storage and imputing genotypes in
101 low-coverage sequencing datasets (Jensen *et al.* 2020; Valdes Franco *et al.* 2020). The PHG is
102 capable of storing genotyping data generated using diverse genotyping technologies as a graph of
103 haplotypes of founder lines and is used for predicting missing genotypes in populations
104 characterized by various sequence- or array-based genotyping strategies. By reducing the
105 constraints associated with large-scale genotyping data storage, processing, and utilization, this
106 tool is another step towards leveraging the existing community-generated genomic diversity

107 resources in breeding and research applications. We used skim-seq, whole-exome capture,
108 genotyping-by-sequencing, and array-based genotyping datasets generated by the USDA-NIFA
109 WheatCAP to develop a wheat PHG database and evaluate its performance for genotype
110 imputation in wheat lines of different levels of relatedness and different depths of genome
111 coverage.

112

113 **Methods:**

114 *Library prep:* DNA was extracted from two-week old leaf tissue of germinated seedlings grown
115 under greenhouse conditions from breeding programs across the United States (Table S1). DNA
116 was extracted using Qiagen DNeasy kit following the manufacturer's protocol. DNA was
117 quantified with Picogreen (Sage Scientific) and wheat exome capture was performed on each
118 sample targeting the non-redundant low-copy portion of the genome. Briefly, wheat exome
119 captures designed in collaboration with Nimblegen targeted 170 Mb of sequence covering about
120 80,000 transcripts (Krasileva *et al.* 2017). The barcoded genomic libraries were pooled at 12- or
121 96-plex levels, and sequenced on NextSeq (Kansas State University Integrated Genomics
122 Facility) and NovaSeq (Kansas University Medical Center) platforms using 2 x 150 bp read runs
123 to produce sequence data providing about 30x coverage of the exome capture target space.

124 Genomic libraries for low-coverage sequencing were prepared for 18 samples from the
125 NAM18 family (Jordan *et al.* 2018) using Illumina DNA Prep Kit along with the Illumina's
126 Nextera CD adapters. Sequencing was performed on the Illumina NextSeq platform to produce
127 ~0.1x coverage.

128 *Data processing:* The quality of sequence reads was assessed using NGSQC toolkit v.2.3.3 (Patel
129 and Jain 2012). The sequence reads were aligned to the wheat reference genome RefSeq v.1.1 (
130 IWGSC, 2018) using HiSat2 (Kim *et al.* 2015) retaining only uniquely mapped reads. The
131 resulting alignments were processed using the GATK pipeline (McKenna *et al.* 2010) to generate
132 a genome variant call file (g.vcf) for each accession.

133 The raw variant calls generated by GATK for exome capture data were filtered using
134 *bcftools* (Danecek *et al.* 2021) to retain variants with minor allele frequency ≥ 0.015 and missing
135 data $< 10\%$. Filtered GATK variants were combined with 90K genotypic data (Wang *et al.*
136 2014), producing high quality filtered variants (henceforth, HQ-SNPs) that were used for
137 assessing the accuracy of the PHG-based imputation.

138 *Wheat PHG database construction:* The Wheat PHG database was built using PHG version
139 0.017. Instructions for creating the PHG along with source code are located with the PHG wiki:
140 <https://bitbucket.org/bucklerlab/practicalhaplotypegraph/wiki/Home>. The approaches and
141 parameters for constructing the Wheat PHG were discussed and developed during two PHG
142 workshops organized at Cornell University. The first step of the PHG database construction is to
143 create reference ranges for data storage and variant imputation (Fig. S1). In this case,
144 “informative” reference ranges were chosen by extending the high confidence gene model
145 coordinates from Chinese Spring RefSeq v.1.1 (IWGSC, 2018) 500 bp in each direction.
146 Adjacent ranges were merged if the boundaries lie within 500 bp from each other. This resulted
147 in a final set of 106,484 informative reference ranges across the genome from the Chinese Spring
148 accession, while the rest of the genome was deemed non-informative and represents intergenic
149 ranges across the genome of Chinese Spring (Fig. S1).

150 The second step in the PHG pipeline populates the database with sequence data from
151 diverse accessions across the reference ranges (Fig. S1). Pre-processed exome capture g.vcf files
152 for 65 accessions, including 58 *Triticum aestivum* accessions, 3 *Aegilops tauschii* accessions, 3
153 *Triticum turgidum* subsp. *durum* wheat cultivars, and one *Triticum turgidum* subsp. *dicoccum*
154 accession (Table S1) generated by GATK (McKenna *et al.* 2010) were loaded into the PHG,
155 creating a database of 6,705,472 haplotypes, which is representative of the diversity across the
156 wheat breeding programs within the US and breeding lines from the Great Plains region.

157 The third step of the PHG pipeline is to create consensus haplotypes for the reference
158 ranges, using the sequence information of the 65 accessions (Fig. S1). This step collapses the raw
159 haplotypes into consensus haplotypes using a user-defined maximum divergence (mxDiv)
160 parameter set to 0.0001. This translates into the clustering of raw haplotypes that contain less
161 than 1 bp divergence per 10,000 bp into a common haplotype. The value of the mxDiv parameter
162 was based on prior diversity estimates in wheat (Akhunov *et al.* 2010; Jordan *et al.* 2015), and
163 aimed at retaining a manageable number of haplotypes per reference range as described in Jensen
164 *et al.* (2020). In addition to the mxDiv parameter, we set minTaxa = 1, which retains haplotypes
165 present in only one accession and facilitates the imputation of rare haplotypes. Using these
166 parameters, a total of 712,733 consensus haplotypes were detected, which is approximately 6.7
167 haplotypes per informative reference range, similar to ~5 haplotypes per reference range reported
168 in the sorghum PHG (Jensen *et al.* 2020).

169 At the imputation step, the low coverage sequence data were aligned to the consensus
170 haplotypes stored in the PHG database (Fig S1), and a Hidden Markov model was used to infer
171 the paths through the practical haplotype graph that match the mapped reads while determining
172 the missing haplotypes. The variants were imputed using the haplotype structure stored in the

173 database, and exported as a vcf file. By using `minReads = 0` parameter, variant calls were
174 imputed for all variable positions in the wheat PHG database.

175 To assess the effect of genome coverage depth on imputation accuracy, we used *seqtk* (Li
176 2012) to generate down-sampled datasets from the 170Mb wheat exome capture data
177 representative of 0.01x (5,667 paired-end (PE) reads per accession), 0.1x (56,667 PE per
178 accession), and 0.5x (283,333 PE reads per accession) depth of coverage for 20 breeding lines
179 from the US Great Plains (Table S1). This set of 20 breeding lines included four lines (Duster,
180 Overlay, NuPlains, and Zenda), which were used to build the PHG database.

181 In addition, we assessed the accuracy of imputation in genotyping datasets generated
182 using GBS of genomic libraries prepared from MseI-PstI digested DNA (Saintenac *et al.* 2013)
183 and whole-genome skim sequencing (Malmberg *et al.* 2018). We used previously published GBS
184 data produced for 75 recombinant inbred lines from the wheat nested-association mapping
185 (NAM) population (Jordan *et al.* 2018) that included an average of 1.85 million (1x 100bp) reads
186 per accession. In the current study, we performed the whole-genome skim sequencing on a set of
187 18 recombinant inbred lines, using 2 x 150 bp sequencing reads providing on average 6.1 million
188 paired-end reads per accession, which represents ~0.1x genome coverage (Table S2).

189 *PHG Imputation Accuracy:* The accuracy of genotype calls for each accession was determined
190 by dividing the number of matching genotype calls between the HQ-SNPs and the PHG-imputed
191 SNP data by the total number of overlapped genotype calls. For down-sampled datasets
192 generated from the exome capture data, imputation accuracy was estimated using nearly 400,000
193 genotype calls per accession at each sequence coverage level. Imputation accuracy comparisons
194 by genome, and by MAF category were performed using ANOVA from *car* and *lme4* R

195 packages. The imputation accuracy estimates for the GBS and whole-genome skim-sequencing
196 data were based on approximately 5,000 HQ-SNPs genotype calls per accession.

197 *Comparison of accuracy between Beagle v.5.0 and PHG-based genotype imputation:* The
198 accuracy of PHG genotype imputation was compared to the accuracy of imputation with Beagle
199 v.5.0. (Browning and Browning 2013). For this purpose, we used the *reference panel* of 65
200 accessions that was also utilized to construct the wheat PHG. The genotyping data for a *target*
201 *panel* were generated by calling genotypes using the down-sampled sequence reads following the
202 same SNP calling procedures described above. The genotype calls were produced for the 20
203 Great Plains accessions at each genome coverage level. Imputation was performed using Beagle
204 v.5.0 with the default parameters. The genotype calls imputed with Beagle were compared to the
205 HQ-SNP dataset (see above) to assess the overall concordance and concordance of minor allele
206 calls. On average, the estimates of accuracy were based on about 323,000 genotype calls per
207 accession. Formal comparisons of the imputation accuracy between Beagle v5.0 and PHG
208 imputation methods by coverage level for 0.01x and 0.1x were performed using paired t-tests in
209 R. At each coverage level, imputation using PHG was statistically more accurate (0.01x: *p-value*
210 = 2.7×10^{-4} ; 0.1x: *p-value* = 2.2×10^{-7}).

211 *Diversity analysis:* Diversity statistics (π and Tajima's D) were estimated using TASSEL v5.2.65
212 (Bradbury *et al.* 2007) in sliding windows of 2,000 SNPs per window stepping 1,000 SNPs at a
213 time, and mean values per genome were calculated. The identity-by-descent (IBD) analysis was
214 determined using Beagle v.4.1 with the default parameters (Browning and Browning 2013), and
215 considered to be significant at $\text{LOD} \geq 3.0$. Overlap between the IBD segments was determined
216 using the MultiIntersectBed tool of the Bedtools suite v.2.26.0 (Quinlan and Hall 2010). Pairwise
217 linkage disequilibrium (LD) was determined using PLINK v.1.90b3.45 (Purcell *et al.* 2007) by

218 calculating the coefficients of determination (r^2) for all possible pairwise combinations of SNP
219 sites on the same chromosomes.

220 *Stepwise regression using the PHG imputed markers:* The parental lines of a family of 75
221 recombinant inbred lines (RILs) from the spring wheat NAM panel (Jordan *et al.* 2018) were
222 included into the panel of 65 accessions that were used to construct the wheat PHG. We ran PHG
223 imputation on the GBS data generated for 75 RILs, and imputed genotypes for 1.457 million
224 sites. These sites were filtered to retain variants that segregate between the parental lines, and
225 have allele frequencies between 0.35-0.65 in the RIL population. These variants were
226 subsequently thinned using PLINK (Purcell *et al.* 2007) to remove markers that had an $r^2 \geq 0.6$
227 within a 50 SNP window, stepping 10 SNPs at a time. The resulting set of 9,806 markers with no
228 missing data was used for stepwise regression mapping performed with the ICIM software
229 v.4.1.0.0 (Meng *et al.* 2015) with markers entering and exiting the model with $p\text{-value} < 0.0001$.
230 The estimates of the Total number of CrossOvers (TCO) and the distal CrossOvers (dCO) were
231 taken from the previous analyses of the spring wheat NAM population for family NAM1 (Jordan
232 *et al.* 2018). Heading dates were measured in three locations for two growing seasons (Montana,
233 South Dakota, Washington) for the 75 RILs and three checks. Best linear unbiased predictions
234 (BLUPs) for each line were estimated using the following linear mixed model with *lmer* package
235 in R:

$$236 \quad \mathbf{HD = year + location + line + year(location) + line*year}$$

237 where location, year, and location nested within year are fixed variables, and the line and line-
238 by-year interaction terms are random variables.

239

240 **Results:**

241 ***The Wheat PHG database development***

242 A wheat PHG database was created using whole-exome capture data from a set of 65
243 wheat accessions (Table S1) contributed by the major U.S. wheat breeding programs, as well as
244 the parental lines used for the genetic analyses of the yield component traits in WheatCAP
245 (www.triticeaecap.org). This set of accessions was selected from a larger diversity panel of
246 nearly 250 wheat cultivars assembled in coordination with the U.S. wheat breeding programs to
247 build a genomic resource to be used as a reference panel for genotype imputation. This diverse
248 set of 65 accessions is comprised of mostly spring and winter bread wheat cultivars, but it also
249 included three accessions of the diploid ancestor of the wheat D genome, *Aegilops tauschii*
250 (accessions TA1615, TA1718, and TA1662/PI603230), and four accessions of tetraploid wheat
251 (three *Triticum turgidum* subsp. *durum* wheat cultivars Langdon, Ben, and Mountrail and one
252 domesticated emmer, *Triticum turgidum* subsp. *dicoccum*, accession PI41025).

253 For constructing the wheat PHG, the wheat genome was split into a set of informative
254 reference ranges that represent the high confidence gene models in the IWGSC RefSeq v.1.1
255 (IWGSC, 2018). By using the predicted gene models to define reference ranges, we aimed to
256 reduce the impact of erroneous genotype calling associated with the misalignments of sequence
257 reads to the repetitive portion of the wheat genome (Wicker *et al.* 2018) on the estimation of
258 linkage disequilibrium (LD) and detecting haplotype blocks. A total of 106,484 reference ranges
259 spanning all 21 chromosomes were defined (Fig S1; Table S3), with an average of 5,070
260 reference ranges per chromosome; chromosome 4D contains the lowest (3,612 ranges) and
261 chromosome 2B harbors the highest (6,221 ranges) number of reference ranges.

262 Using the 65 accessions to populate the wheat PHG database, we discovered 1,473,670
263 variants across the 106,484 reference ranges, of which 1,457,321 are high quality, bi-allelic
264 SNPs (Table S3). The inclusion of three diploid *Ae. tauschii* accessions into the panel increased
265 the number of variable sites detected in the D genome lineage, which is the least polymorphic
266 genome in bread wheat (Wang *et al.* 2013; Jordan *et al.* 2015; He *et al.* 2019). Excluding the
267 variants from *Ae. tauschii*, we found that 161,226 (31%) sites in the D genome were
268 monomorphic among the bread wheat cultivars. Similarly, we found that 31,486 SNPs (7%) in
269 the A genome and 32,228 SNPs (6%) in the B genome are contributed by the domesticated
270 emmer and durum lines, and are monomorphic in hexaploid wheat. These private SNPs explain
271 the high levels of divergence between the domesticated emmer and *Ae. tauschii* accessions from
272 the hexaploid wheat lines (Fig. 1a). The overall patterns of genetic diversity and allele frequency
273 distribution in the D genome compared to those in the A and B genomes were consistent with the
274 population bottleneck (Table 1): 1) diversity mean estimates for the D genome were less than
275 2.3-fold that of the A and B genomes, ($\pi_D = 0.076$, $\pi_A = 0.175$, and $\pi_B = 0.182$; Table 1), 2) the
276 estimates of Tajima's D were lower in the D genome than in the A and B genomes (Tajima's
277 $D_D = -2.19$, Tajima's $D_A = -0.67$, and Tajima's $D_B = -0.55$, Table 1), 3) the mean minor allele
278 frequencies (MAF) were greater in the A and B genomes than in the D genome ($MAF_A = 0.12$,
279 $MAF_B = 0.12$, and $MAF_D = 0.05$), and 4) LD drops to half of its initial value ($r^2 \leq 0.33$) at 20 Mb
280 in the D genome, whereas in the A and B genomes LD drops to the same level at 12 and 10 Mb,
281 respectively (Table 1, Figure 1b).

282

283

284

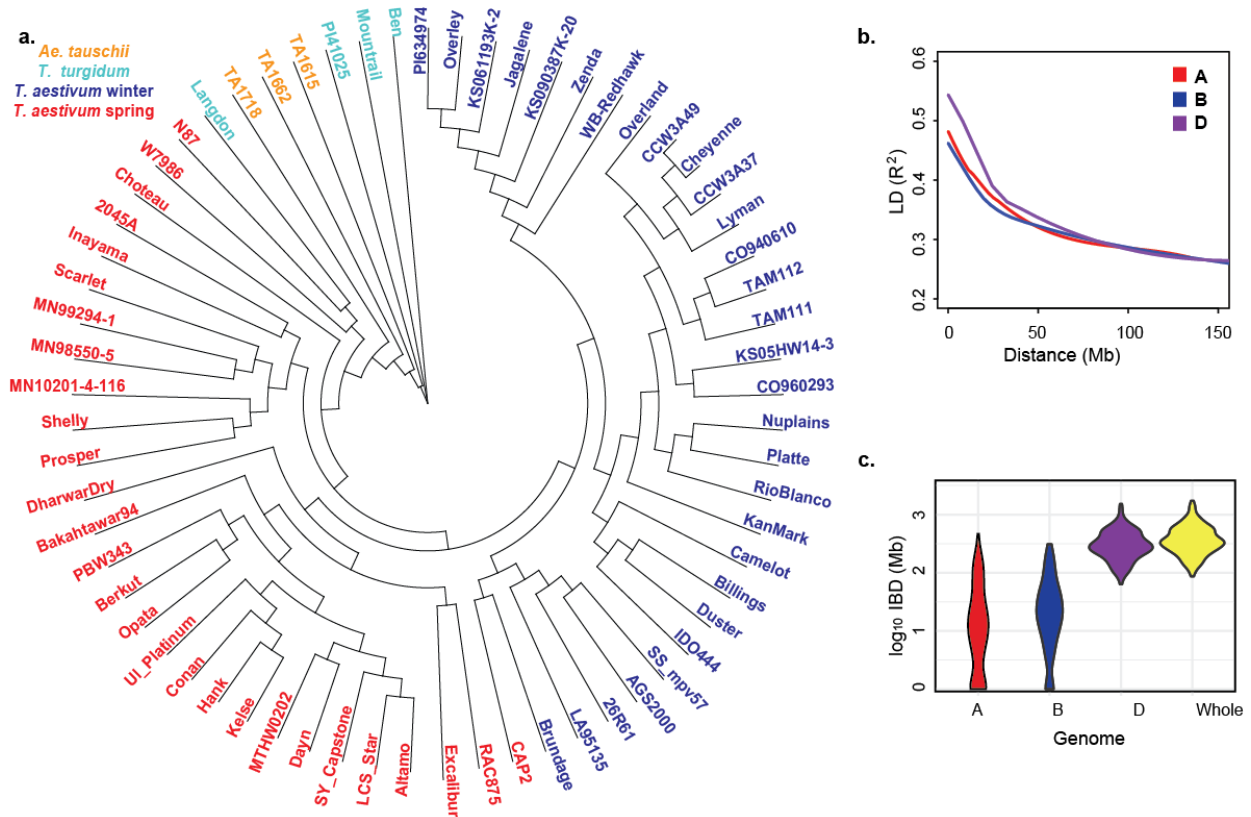
285 **Table 1.** Estimates of genetic diversity (π), minor allele frequency (MAF), Tajima's D and
286 linkage disequilibrium in the population used for constructing the Wheat PHG.

Diversity statistic	A genome	B genome	D genome
No. SNPs	430,050	504,260	523,011
MAF	0.116	0.122	0.050
π (per bp)	0.175	0.182	0.076
Tajima's D	-0.673	-0.552	-2.192
LD* ($r^2 \leq 0.33$)	12.2 Mb	9.8 Mb	20.0 Mb

287 *distance at which LD drops to half of its initial value ($r^2 \leq 0.33$).

288 The accuracy and the rate of genotype imputation are affected by the proportion of shared
289 genetic ancestry among individuals in a population (Browning and Browning 2013). For each
290 WheatCAP parental line included in the Wheat PHG, we estimated the length of genomic
291 segments sharing identity-by-descent (IBD) with other lines in the panel. On average, the pairs of
292 parents had 451 Mb (~3%) of IBD segments (Table S4), suggesting distant relationships among
293 the WheatCAP parental lines. However, the estimates of the total length of IBD segments among
294 cultivars were quite variable (Figure 1c). For example, in cultivars Prosper from North Dakota
295 and Shelly from Minnesota, the length of shared IBD segments was nearly 1.29 Gb (8.6%),
296 whereas hard winter wheat cultivars Lyman (South Dakota) and Overley (Kansas) shared only
297 128 Mb (0.85%) of IBD segments. The average length of IBD segments shared by the distantly
298 related durum wheat and domesticated emmer parents was only 57.6 Mb. Across all breeding
299 programs, we detected 556 regions sharing IBD, with an average IBD segment length of 12.2
300 Mb. Over half (53%) of the IBD segments overlapped with a segment from at least one other
301 breeding program, translating to more than 1.68 Gb of the genome shared between any two
302 wheat breeding programs. This estimate includes 1.49 Gb of shared IBD in the D genome (89%),
303 while only 86.4 Mb and 105.7 Mb of IBD with other breeding programs were detected in the A

304 and B genomes, respectively. The genomic segments sharing IBD with most of the wheat lines
 305 were located on chromosomes 7D (568 Mb - 571 Mb) and 3D (496.6 Mb - 505 Mb), which were
 306 common to seven breeding programs.



307
 308 **Figure 1. Genetic diversity of 65 accessions of wheat and its diploid and tetraploid relatives**
 309 **used for developing the Wheat PHG. a.** Neighbor-joining tree of accessions used for
 310 constructing the Wheat PHG. **b.** The rate of LD decay in the A, B and D genomes of wheat. **c.**
 311 The length of pair-wise IBD between the parental lines from different breeding programs used in
 312 WheatCAP.

313 In addition to the WheatCAP lines, we selected 21 hard red winter wheat cultivars from
 314 the U.S. Great Plains for constructing the PHG database (Table S1). Pairwise comparisons
 315 among these lines showed that, on average, they share 416 Mb of IBD segments, with an average
 316 IBD segment length of 13 Mb, and nearly 83% of all shared IBD regions are located in the D
 317 genome (Table S5). This finding is consistent with the lack of diversity among breeding lines in

318 the D genome (Chao *et al.* 2010) and the high levels of shared ancestry among the lines from the
319 U.S. Great Plains' breeding programs.

320

321 ***Genotype imputation using the Wheat PHG***

322 We used several low-coverage sequencing datasets to assess the imputation performance
323 of the wheat PHG. First, we used 20 spring and winter wheat lines (Table S1) from the U.S.
324 wheat breeding programs sequenced using the whole-exome capture approach (Krasileva *et al.*
325 2017; He *et al.* 2019) to mimic a low-coverage sequencing experiment. We down-sampled the
326 raw unmapped Illumina paired-end reads generated for each accession to create datasets with
327 three levels of sequence coverage depths (0.01x, 0.1x, and 0.5x) for the regions targeted by the
328 exome capture assay. The accuracy of imputation achieved using the Wheat PHG was estimated
329 by comparing the concordance of imputed genotype calls with the genotype calls from the HQ-
330 SNP set generated using the 90K iSelect array (Wang *et al.* 2014) and the high-coverage (20-30x
331 coverage) exome sequencing.

332 On average, using 0.5x coverage down-sampled exome capture data, we achieved 96.9%
333 imputation accuracy, ranging from 95% to 98% among lines (Figure 2a, Table 3). Five- and
334 fifty-fold reduction in the depth of read coverage for the inference panel did not result in a
335 substantial reduction in the accuracy of imputation. The mean accuracy of PHG imputation was
336 96% (94-98% range) with 0.1x coverage depth, and 93% (91-98% range) with as little as 0.01x
337 coverage depth (Figure 2a, Tables 2 and 3). These results suggest that the imputation method in
338 the PHG could effectively use 0.01x exome coverage data to adequately capture the haplotypic
339 diversity of the inference panel to achieve 93% imputation accuracy. The imputation accuracy

340 varied among the wheat genomes, likely due to genome-specific differences in the extent of LD
341 and haplotypic diversity (Jordan *et al.* 2015). At 0.01x coverage depth, the accuracy of genotype
342 imputation in the D genome was 95.5%, which was 3.2% and 4.3% more accurate (*p-value*
343 $(ANOVA) = 3.73 \times 10^{-5}$) than imputation in the A (92.3%), and the B genomes (91.2%), respectively
344 (Table 2; Figure 2b). The higher extent of LD in the D genome appears to contribute to more
345 accurate genotype imputation compared to that in the A and B genomes, which show faster rates
346 of LD decay and lower proportions of the genome sharing IBD segments in the panel used to
347 build the PHG database.

348 **Table 2. The accuracy of PHG imputation in different wheat genomes.**

Wheat genomes	Exome capture* (0.01x), %	GBS*, %	skim-seq* (0.1x), %
Total	92.6	90.4	85.3
A	92.3	90.9	86.5
B	91.2	91.1	84.9
D	95.5	87.4	82.9

349 *Accuracies by approach are comprised of different germplasm, EC: n=20, GBS: n=75, skim-seq: n=18

350 We compared the performance of the wheat PHG to one of the commonly used low-
351 coverage imputation methods implemented in Beagle v5.0 (Browning and Browning 2013). For
352 this purpose, the panel of 65 accessions included into the wheat PHG database was used as the
353 reference panel, and an independent set of 20 wheat cultivars from the U.S. wheat breeding
354 programs was used as the inference panel. Overall, Beagle imputed missing genotypes with
355 89.2% accuracy for this set of 20 lines at 0.01x coverage (ranging from 76% to 94%), and 92.6%
356 (ranging from 84% to 95%) at 0.1x coverage (Figure 2a, Table 3). Direct comparisons of
357 imputation methods show PHG imputation statistically outperformed Beagle imputation by 4%
358 at both coverage levels (*p-value* $_{0.1x (t-test)} = 2.0 \times 10^{-7}$; *p-value* $_{0.01x (t-test)} = 2.7 \times 10^{-4}$).

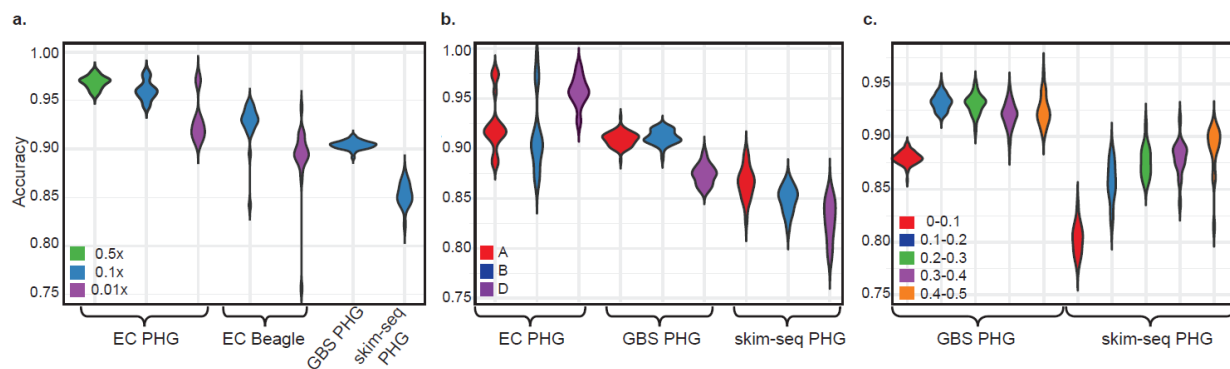
359 **Table 3. Comparison of imputation accuracy between PHG and Beagle using exome**
 360 **capture data.**

Lines	PHG 0.5x	PHG 0.1x	PHG 0.01x	Beagle 0.1x	Beagle 0.01x
Bolles	97.2%	96.0%	92.2%	93.3%	88.6%
Duster*	97.8%	97.6%	97.1%	93.0%	89.3%
Expedition	97.1%	96.1%	93.1%	93.3%	91.0%
Forefront	96.5%	95.3%	91.0%	91.7%	88.0%
Goodstreak	97.3%	96.2%	91.3%	93.8%	90.3%
Ideal	96.4%	95.7%	92.6%	92.6%	89.0%
Jagger	95.9%	94.4%	90.6%	84.2%	75.6%
Linkert	96.9%	96.1%	93.3%	93.8%	90.1%
McGill	96.6%	95.3%	91.4%	92.6%	89.6%
Mott	97.1%	96.3%	91.8%	93.2%	89.6%
NuPlains*	98.0%	98.0%	96.7%	94.5%	91.4%
Overley*	97.1%	97.3%	97.1%	92.9%	89.4%
Panhandle	96.3%	95.4%	91.8%	92.0%	89.2%
Prevail	96.8%	95.8%	90.7%	91.8%	89.7%
Robidoux	97.2%	96.7%	92.7%	94.0%	91.7%
TAM303	95.6%	94.4%	91.4%	89.5%	87.0%
Traverse	97.0%	95.6%	91.9%	93.0%	90.3%
Wesley	97.2%	96.2%	92.9%	94.6%	91.8%
Yellowstone	95.9%	94.8%	92.0%	94.7%	94.3%
Zenda*	97.7%	97.7%	97.6%	92.7%	88.9%
Average	96.9%	96.0%	93.0%	92.6%	89.2%

* represents cultivars used in PHG database construction

361 We conducted the analyses of PHG performance in the datasets down-sampled from
 362 exome capture data generated for four cultivars Duster, Overley, NuPlains, and Zenda, that were
 363 included in the wheat PHG construction. The accuracy of PHG-based imputation for these four
 364 cultivars was statistically higher (ANOVA for different levels of sequence coverage: *p-value*_{0.5x}
 365 = 0.004; *p-value*_{0.1x} = 2.0 x 10⁻⁵; *p-value*_{0.01x} = 7.4 x 10⁻¹⁰) than for other cultivars at all levels of
 366 sequence coverage (Fig. S2a). No similar relationship between the presence of specific
 367 haplotypes in the reference panel and imputation accuracy was observed for Beagle. We further
 368 explored this relationship by analyzing genotype imputation results in cultivar Jagger, which
 369 showed a substantial reduction in imputation accuracy in the low sequence coverage datasets
 370 (0.1x and 0.01x coverage) imputed using Beagle (Fig. S2a). We assumed that one of the likely

371 factors contributing to the decreased imputation performance of Beagle in cultivar Jagger was
372 the presence of wild-relative introgression from *Ae. ventricosa* on chromosome 2A (Cruz et al.
373 2016). Because cultivar Overley, which was used to build the PHG database, also carries this *Ae.*
374 *ventricosa* introgression (Cruz et al. 2016), we could evaluate the impact of the presence of the
375 rare introgressed haplotype in both the PHG database and the Beagle's reference panel on
376 imputation accuracy. The chromosome-by-chromosome assessment of imputation accuracy for
377 cv. Jagger in the 0.01x coverage dataset showed modest accuracy (90%) for chromosome 2A
378 using PHG. However, for the same chromosome, the imputation accuracy of Beagle reached
379 only 63% (Fig. S2b). The accuracy of Beagle imputation was also low for other chromosomes
380 (2D, 6A, 7A) (Fig. S2b), which suggests that cv. Jagger likely carries other regions with unique
381 haplotypes (Kippes *et al.* 2018; Walkowiak *et al.* 2020) poorly represented in the reference set
382 used for Beagle imputation. For the same three chromosomes, the accuracy of PHG imputation
383 was higher than that obtained using Beagle, indicating that PHG is more effective at utilizing the
384 rare haplotypes in the reference panel for imputation than Beagle.



385
386 **Figure 2. The accuracy of imputation using the wheat PHG. a.** The impact of sequence
387 coverage and the method of imputation on accuracy, (EC: n=20; GBS: n=75; skim-seq: n=18) **b.**
388 Accuracy of imputation in the A, B and D genomes of wheat using exome capture (EC), GBS
389 and whole genome skim-seq data. **c.** Accuracy of imputation for alleles with different minor
390 allele frequency.

391

392 ***Imputation accuracy with reduced coverage sequencing data***

393 To this point, we tested the imputation accuracy using the same type of genomic data
394 (whole-exome capture) as was used to populate the database. We also evaluated the utility of the
395 developed PHG database for imputing genotypes in the inference panels genotyped using two
396 cost-effective complexity-reduced sequencing approaches, genotyping-by-sequencing (GBS)
397 (Elshire *et al.* 2011; Saintenac *et al.* 2013) and whole-genome skim-seq (Malmberg *et al.* 2018).
398 First, utilizing GBS reads generated for a set of recombinant inbred lines (RILs) from the spring
399 wheat NAM panel (Jordan *et al.* 2018), we performed genotype imputation at 1.4 million
400 variable sites. The parents of these NAM RILs were included into the wheat PHG construction.
401 The mean accuracy of imputation across the 75 RILs was 90.4%, ranging from 89 - 91.4% across
402 individual lines (Figure 2a, Table S6). These estimates of accuracy were only slightly lower than
403 those observed for the imputed genotypes in the down-sampled exome capture data, and likely
404 explained by the relatively small overlap (~5%) between the sites in the GBS and exome capture
405 datasets (Jordan *et al.* 2015). Overall, this result indicates that the PHG based on the panel of
406 wheat lines re-sequenced by exome capture assay provides accurate imputation on the inference
407 population characterized by complexity-reduced sequencing approaches similar to the GBS
408 method.

409 We also evaluated the wheat PHG in a set of NAM RILs genotyped using the whole-
410 genome skim-seq approach. The genomic libraries generated for a set of RILs from the spring
411 wheat NAM population (Jordan *et al.* 2018; Blake *et al.* 2019) were sequenced on an Illumina
412 sequencer (2 x 150 bp run) to provide ~0.1x genome coverage. The accuracy of PHG-imputed
413 genotypes in the skim-seq dataset (85.3%) was lower than that obtained for genotypes in either
414 the exome capture or GBS datasets. This lower accuracy likely is associated with a lower

415 proportion of skim-seq reads, mostly represented by reads from the repetitive regions, uniquely
 416 mapped to the wheat genome compared to the proportion of uniquely mapped reads from the
 417 exome capture and GBS datasets, which are enriched for the low-copy genomic regions
 418 (Saintenac *et al.* 2013; Jordan *et al.* 2015). The accuracy of imputation varied across different
 419 SNP frequency classes. For SNPs with MAF > 0.1, the accuracy of imputation improved by at
 420 least 5% for both GBS and skim-seq genotypes. The accuracy reached nearly 90% for skim-seq
 421 and 93% for GBS datasets when the MAF were ≥ 0.2 (Table 4, Figure 2c).

422 **Table 4. Relationship between minor allele frequency and the accuracy of imputation.**
 423

	Minor Allele Frequency (MAF)				
	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5
No. Sites	1,029,330	156,251	97,013	73,001	66,296
GBS Accuracy	0.8798	0.9326	0.9309	0.9209	0.9249
skim-seq Accuracy	0.8026	0.8570	0.8764	0.8798	0.8886

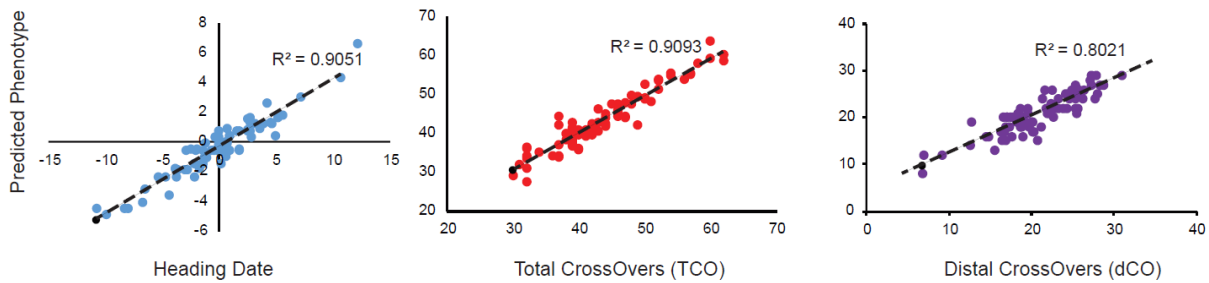
424

425 *Genetic analyses of trait variation using the imputed genotypes*

426 The ability to accurately impute genotypes across the genome in low-coverage
 427 sequencing datasets provides a cost-effective means for advancing the genetic dissection of trait
 428 variation. We used the imputed genotypes to assess the genetic contribution to heading date
 429 (HD) variation in the nested association mapping (NAM) family previously used for studying the
 430 genetics of recombination rate variation in wheat (Jordan *et al.* 2018). A stepwise regression
 431 (SR) was applied to identify variants associated with phenotypic variation. Before mapping, co-
 432 segregating redundant markers were removed, resulting in nearly 10,000 markers with no
 433 missing data. The SR method identified 11 SNPs together explaining 90% of the variance in
 434 heading date, which was measured over two years at three locations (Fig 3, Table S7). Among

435 these SNPs are loci with modest effect sizes located on the long arms of chromosomes 5A and
436 5D, within 10 Mb from the *Vrn-A1* and *Vrn-D1* loci, which play a major role in the regulation of
437 flowering in wheat (Distelfeld *et al.* 2009). In addition, significant SNPs on chromosomes 1B
438 and 1D were mapped to the regions within 50 Mb of the *Elf-3* gene, which is associated with the
439 transition from vegetative to reproductive growth in wheat (Alvarez *et al.* 2016; Zikhali *et al.*
440 2016).

441 We also used the imputed genotypes to revisit the genetic analysis of meiotic crossover
442 rate variation in the wheat NAM population (Jordan *et al.* 2018; Blake *et al.* 2019). In the
443 previous study, using a limited number of SNPs genotyped using the 90K iSelect array and GBS,
444 we performed SR analysis and identified 15 and 12 SNPs associated with variation in the total
445 number of crossovers (TCO) and the number of distal crossovers (dCO), respectively (Jordan *et*
446 *al.* 2018). The identified SNPs explained 48.6% of the variation for TCO and 41% of the
447 variation for dCO. Using the PHG imputed genotypes, we mapped 16 SNPs that together
448 explained 91% of the variance for TCO per line and 12 SNPs explaining 80% of the variance for
449 dCO (Fig. 3, Table S7). Compared to the previous study, SR analyses based on the PHG imputed
450 SNPs detected additional loci with smaller effects on crossover rate (Jordan *et al.* 2018). As a
451 result, the average effect size estimates for TCO and dCO were 2.5 COs and 1.5 COs,
452 respectively. These estimates were lower than the previously reported average effect sizes of
453 3.36 COs for TCO and 2.3 COs for dCO (Jordan *et al.* 2018). Taken together, these results
454 indicate that the increase in marker density after imputation using the wheat PHG helped to
455 identify new loci with a broader range of effect sizes that together explain a higher proportion of
456 genetic variance compared to the previous study (Jordan *et al.* 2018).



457

458 **Figure 3.** Relationship between the true and predicted phenotypes. Significant markers were
459 identified by stepwise regression on heading date, total number of crossovers per line (TCO), and
460 total number of distal crossovers per line (dCO) phenotypes.

461

462 **Discussion:**

463 We constructed a wheat PHG database using wheat lines from the major U.S. breeding
464 programs and demonstrated that PHG combined with inexpensive low-coverage genome
465 sequencing could be used to impute genotypes with high accuracy, sufficient to identify variants
466 with smaller effects and support high-resolution mapping studies. Our analyses suggest that the
467 wheat PHG has the potential to effectively utilize community-generated whole-exome capture
468 datasets, currently including thousands of diverse wheat accessions from different geographic
469 regions (Molero *et al.* 2018; He *et al.* 2019; Pont *et al.* 2019), to create a global resource for
470 imputing genotypes. The imputation accuracy provided by the PHG in populations genotyped
471 using the skim-seq, GBS, as well as low-coverage exome sequencing approaches varied, but
472 overall were comparable, indicating that the marker density in the large populations of wheat
473 lines previously genotyped using these methods could be substantially increased by imputation
474 with this newly developed wheat PHG tool.

475 The accuracy of PHG imputation compared favorably with the commonly used
476 imputation tool Beagle v.5.0 (Browning and Browning 2013), which imputed genotypes with 4%

477 lower accuracy at 0.01x and 0.1x genome coverage than the wheat PHG. In previous studies,
478 imputation of exome capture data with Beagle in populations genotyped using the 90K SNP
479 array and GBS was 93-97% (Jordan *et al.* 2015) and 98% (Nyine *et al.* 2019), respectively.
480 These estimates of accuracy are slightly higher than those obtained in our current study, but
481 overall are comparable, and likely associated with filtering applied to reduce the proportion of
482 missing data in the imputed datasets (Nyine *et al.* 2019), and with the inclusion of more common
483 variants from the array-based genotyping methods. Compared to the imputation accuracy of
484 sorghum (94.1%) and maize (92-95%) PHGs (Jensen *et al.* 2020; Valdes Franco *et al.* 2020), our
485 estimates of accuracy were slightly lower and likely caused by genotyping errors associated with
486 the misalignment of short reads to the more complex, highly repetitive, allopolyploid wheat
487 genome. The higher imputation accuracy in the low-coverage datasets down-sampled from the
488 whole exome capture compared to the accuracy of whole genome skim-seq datasets, which are
489 mostly composed of reads from the repetitive regions of the wheat genome, supports this
490 explanation.

491 The imputation accuracy among different allele frequency classes improves with an
492 increase in the allele frequency and is higher for a reference allele than for an alternative allele.
493 Consistent with these expectations, the accuracy of imputation in the GBS dataset improved from
494 87.9% for SNPs with $MAF < 0.1$ to 92.5% for SNPs with $MAF > 0.4$, and in the skim-seq
495 dataset from 80.3% for SNPs with $MAF < 0.1$ to 88.9% for SNPs with $MAF > 0.4$. Previous
496 studies showed that an increase in the reference population size also increases the probability of
497 capturing rare alleles and substantially improves the imputation accuracy of rare variants (Shi *et*
498 *al.* 2017; Das *et al.* 2018). Our results suggest that the wheat PHG appear to be more effective at
499 utilizing rare haplotypes included into the reference panel for genotype imputation than the

500 commonly used low-coverage imputation method from Beagle. This was demonstrated by
501 imputing genotypes on chromosome 2A, which carries introgression from *Ae. ventricosa* in
502 cultivar Jagger (Cruz *et al.* 2016). The inclusion of genotyping data from cultivar Overley, which
503 also carries this *Ae. ventricosa* introgression, into the PHG database was sufficient for accurate
504 imputation in Jagger. In spite of including genotyping data from cultivar Overley into the
505 reference panel, Beagle imputation of chromosome 2A genotypes in cultivar Jagger was lower
506 compared to PHG. Further efforts aimed at broadening the diversity of accessions in the wheat
507 PHG, including wheat lines carrying known introgressions from wild relatives, will be needed to
508 improve the utility PHG tool for genotype imputation in wheat germplasm.

509 The application of imputed genotypes to the genetic analyses of trait variation in the
510 wheat NAM population showed that an increase in marker density increases the number of loci
511 associated with trait variation and detects alleles that have smaller effects on phenotypes (*e.g.*,
512 recombination rate) than those previously detected using lower density marker sets. The increase
513 in the number of significant loci also resulted in a higher proportion of genetic variance (80-
514 91%) in recombination rate and heading date being explained, suggesting that the imputed
515 genotypes are better at capturing the genetic architecture of these traits, and have the potential to
516 identify more adaptive and beneficial genetic targets in breeding programs.

517

518 **Acknowledgements**

519 This project is supported by the Agriculture and Food Research Initiative Competitive
520 Grant 2017–67007-25939 (WheatCAP) from the USDA National Institute of Food and
521 Agriculture, and by the International Wheat Yield Partnership (IWYP). Mention of trade names
522 or commercial products in this publication is solely for the purpose of providing specific

523 information and does not imply recommendation or endorsement by the US Department of
524 Agriculture. USDA is an equal opportunity provider and employee.

525

526 **Data availability**

527 The raw sequence data for previously published accessions can be accessed from the NCBI
528 Short-Read Archive database (BioProject SUB2540330 and PRJNA381058). Newly generated
529 exome capture data can be accessed from NCBI Short-Read Archive database (BioProject
530 PRJNA732645). Phenotypic datasets for NAM family 1 associated with the paper can be
531 downloaded from the wheat NAM project website:

532 <http://wheatgenomics.plantpath.ksu.edu/nam/>.

533

534 **References:**

535 Akhunov, E. D., A. R. Akhunova, O. D. Anderson, J. a Anderson, N. Blake *et al.*, 2010
536 Nucleotide diversity maps reveal variation in diversity among wheat genomes and
537 chromosomes. *BMC Genomics* 11: 702.

538 Alvarez, M. A., G. Tranquilli, S. Lewis, N. Kippes, and J. Dubcovsky, 2016 Genetic and
539 physical mapping of the earliness per se locus Eps-A m 1 in *Triticum monococcum*
540 identifies EARLY FLOWERING 3 (ELF3) as a candidate gene. *Funct. Integr. Genomics*
541 16: 365–382.

542 Balfourier, F., S. Bouchet, S. Robert, R. DeOliveira, H. Rimbart *et al.*, 2019 Worldwide
543 phylogeography and history of wheat genetic diversity. *Sci. Adv.* 5:.

544 Blake, N. K., M. Pumphrey, K. Glover, S. Chao, K. Jordan *et al.*, 2019 Registration of the
545 Triticeae-CAP Spring Wheat Nested Association Mapping Population. *J. Plant Regist.* 0: 0.

- 546 Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al.*, 2007 TASSEL:
547 software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:
548 2633–5.
- 549 Browning, B. L., and S. R. Browning, 2013 Improving the accuracy and efficiency of identity-
550 by-descent detection in population data. *Genetics* 194: 459–71.
- 551 Chao, S., J. Dubcovsky, J. Dvorak, M.-C. Luo, S. P. Baenziger *et al.*, 2010 Population- and
552 genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter
553 wheat (*Triticum aestivum* L.). *BMC Genomics* 11:.
- 554 Cruz, C. D., G. L. Peterson, W. W. Bockus, P. Kankanala, J. Dubcovsky *et al.*, 2016 The 2NS
555 translocation from *Aegilops ventricosa* confers resistance to the *Triticum* pathotype of
556 *Magnaporthe oryzae*. *Crop Sci.* 56:.
- 557 Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan *et al.*, 2021 Twelve years of
558 SAMtools and BCFtools. *Gigascience* 10: 1–4.
- 559 Das, S., G. R. Abecasis, and B. L. Browning, 2018 Genotype Imputation from Large Reference
560 Panels. *Annu. Rev. Genomics Hum. Genet.* 19: 73–96.
- 561 Davies, R. W., J. Flint, S. Myers, and R. Mott, 2016 Rapid genotype imputation from sequence
562 without reference panels. *Nat. Genet.* 48: 965–969.
- 563 Distelfeld, A., C. Li, and J. Dubcovsky, 2009 Regulation of flowering in temperate cereals. *Curr.*
564 *Opin. Plant Biol.* 12: 178–84.
- 565 Elshire, R. J., J. C. Glaubitz, Q. Sun, J. Poland, K. Kawamoto *et al.*, 2011 A robust, simple
566 genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379.
- 567 He, F., R. Pasam, F. Shi, S. Kant, G. Keeble-Gagnere *et al.*, 2019 Exome sequencing highlights
568 the role of wild relative introgression in shaping the adaptive landscape of the wheat
569 genome. *Nat. Genet.* 51: 896–904.
- 570 Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2014 Training set optimization
571 under population structure in genomic selection. *Theor. Appl. Genet.* 128: 145–58.
- 572 Jensen, S. E., J. R. Charles, K. Muleta, P. J. Bradbury, T. Casstevens *et al.*, 2020 A sorghum

- 573 Practical Haplotype Graph facilitates genome-wide imputation and cost- effective genomic
574 prediction. *Plant Genome* 13: 1–15.
- 575 Jordan, K. W., S. Wang, F. He, S. Chao, Y. Lun *et al.*, 2018 The genetic architecture of genome-
576 wide recombination rate variation in allopolyploid wheat revealed by nested association
577 mapping. *Plant J.* 95: 1039–1054.
- 578 Jordan, K., S. Wang, Y. Lun, L. Gardiner, R. MacLachlan *et al.*, 2015 A haplotype map of
579 allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes.
580 *Genome Biol.* 16: 48.
- 581 Juliana, P., J. Poland, J. Huerta-espino, S. Shrestha, J. Crossa *et al.*, 2019 Improving grain yield,
582 stress resilience and quality of bread wheat using large-scale genomics. *Nat. Genet.*
- 583 Juliana, P., R. P. Singh, J. H. Espino, S. Bhavani, M. S. Randhawa *et al.*, 2020 Genome - wide
584 mapping and allelic fingerprinting provide insights into the genetics of resistance to wheat
585 stripe rust in India , Kenya and Mexico. *Sci. Rep.* 1–16.
- 586 Kim, D., B. Langmead, and S. L. Salzberg, 2015 HISAT: a fast spliced aligner with low memory
587 requirements. *Nat. Methods* 12: 357–60.
- 588 Kippes, N., M. Guedira, L. Lin, M. A. Alvarez, G. L. Brown-Guedira *et al.*, 2018 Single
589 nucleotide polymorphisms in a regulatory site of VRN-A1 first intron are associated with
590 differences in vernalization requirement in winter wheat. *Mol. Genet. Genomics* 293: 1231–
591 1243.
- 592 Krasileva, K. V., H. A. Vasquez-Gross, T. Howell, P. Bailey, F. Paraiso *et al.*, 2017 Uncovering
593 hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U. S. A.* 114: E913–E921.
- 594 Li, H., 2012 seqtk, Toolkit for processing sequences in FASTA/Q formats.
- 595 Malmberg, M. M., D. M. Barbulescu, M. C. Drayton, M. Shinozuka, P. Thakur *et al.*, 2018
596 Evaluation and recommendations for routine genotyping using skim whole genome re-
597 sequencing in canola. *Front. Plant Sci.* 871: 1–15.
- 598 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome
599 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing

- 600 data. *Genome Res.* 20: 1297–303.
- 601 Meng, L., H. Li, L. Zhang, and J. Wang, 2015 QTL IciMapping: Integrated software for genetic
602 linkage map construction and quantitative trait locus mapping in biparental populations.
603 *Crop J.* 3: 269–283.
- 604 Molero, G., R. Joynson, F. J. Pinera-Chavez, L. Gardiner, C. Rivera-Amado *et al.*, 2018
605 Elucidating the genetic basis of biomass accumulation and radiation use efficiency in spring
606 wheat and its role in yield potential. *Plant Biotechnol. J.* 1–13.
- 607 Nyine, M., S. Wang, K. Kiani, K. Jordan, S. Liu *et al.*, 2019 Genotype imputation in winter
608 wheat using first-generation haplotype map SNPs improves genome-wide association
609 mapping and genomic prediction of traits. *G3 Genes, Genomes, Genet.* 9:.
- 610 Patel, R. K., and M. Jain, 2012 NGS QC Toolkit: a toolkit for quality control of next generation
611 sequencing data. *PLoS One* 7: e30619.
- 612 Poland, J. A., and T. W. Rife, 2012 Genotyping-by-Sequencing for Plant Breeding and Genetics.
613 *Plant Genome* 5:.
- 614 Pont, C., T. Leroy, M. Seidel, A. Tondelli, W. Duchemin *et al.*, 2019 Tracing the ancestry of
615 modern bread wheats. *Nat. Genet.* 51: 905–911.
- 616 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. a R. Ferreira *et al.*, 2007 PLINK: a tool set
617 for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*
618 81: 559–75.
- 619 Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing
620 genomic features. *Bioinformatics* 26: 841–842.
- 621 Rubinacci, S., D. M. Ribeiro, R. J. Hofmeister, and O. Delaneau, 2021 Efficient phasing and
622 imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* 53:
623 120–126.
- 624 Saintenac, C., D. Jiang, S. Wang, and E. Akhunov, 2013 Sequence-based mapping of the
625 polyploid wheat genome. *G3 (Bethesda)*. 3: 1105–14.
- 626 Shi, F., J. Tibbits, R. K. Pasam, P. Kay, D. Wong *et al.*, 2017 Exome sequence genotype

627 imputation in globally diverse hexaploid wheat accessions. *Theor. Appl. Genet.* 130: 1393–
628 1404.

629 The International Wheat Genome Sequencing Consortium (IWGSC), 2018 Shifting the limits in
630 wheat research and breeding using a fully annotated reference genome. *Science* 361:
631 eaar7191.

632 Valdes Franco, J. A., J. L. Gage, P. J. Bradbury, L. C. Johnson, Z. R. Miller *et al.*, 2020 A Maize
633 Practical Haplotype Graph Leverages Diverse NAM Assemblies. *bioRxiv* 2: 0.

634 Walkowiak, S., L. Gao, C. Monat, G. Haberer, M. T. Kassa *et al.*, 2020 Multiple wheat genomes
635 reveal global variation in modern breeding. *Nature* 588: 277–283.

636 Wang, J., M.-C. Luo, Z. Chen, F. M. You, Y. Wei *et al.*, 2013 *Aegilops tauschii* single
637 nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity
638 and pinpoint the geographic origin of hexaploid wheat. *New Phytol.* 198: 925–937.

639 Wang, S., D. Wong, K. Forrest, A. Allen, S. Chao *et al.*, 2014 Characterization of polyploid
640 wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array.
641 *Plant Biotechnol. J.* 12: 787–96.

642 Wicker, T., H. Gundlach, M. Spannagl, C. Uauy, P. Borrill *et al.*, 2018 Impact of transposable
643 elements on genome structure and evolution in bread wheat. *Genome Biol.* 19: 1–18.

644 Zikhali, M., L. U. Wingen, and S. Griffiths, 2016 Delimitation of the Earliness per se D1 (Eps-
645 D1) flowering gene to a subtelomeric chromosomal deletion in bread wheat (*Triticum*
646 *aestivum*). *J. Exp. Bot.* 67: 287–299.

647