Received: 31 March 2020 | Accepted: 23 June 2020 | Published online: 17 September 2020

DOI: 10.1002/csc2.20255

Crop Science

ORIGINAL RESEARCH ARTICLE

Crop Breeding & Genetics

Single nucleotide polymorphism calling and imputation strategies for cost-effective genotyping in a tropical maize breeding program

Amanda Avelar de Oliveira[1] | Lauro José Moreira Guimarães[2] | Claudia Teixeira Guimarães[2] | Paulo Evaristo de Oliveira Guimarães[2] | Marcos de Oliveira Pinto[2] | Maria Marta Pastina[2] | Gabriel Rodrigues Alves Margarido[1]

[1] "Luiz de Queiroz" College of Agriculture, Department of Genetics, University of São Paulo, Piracicaba, SP 13418–900, Brazil

[2] Embrapa Milho e Sorgo, Sete Lagoas, MG 35701–970, Brazil

Correspondence

Gabriel Rodrigues Alves Margarido, University of São Paulo, "Luiz de Queiroz" College of Agriculture, Department of Genetics, Piracicaba, SP 13418–900, Brazil.
Email: gramarga@usp.br

Maria Marta Pastina, Embrapa Milho e Sorgo, Sete Lagoas, MG 35701–970, Brazil.
Email: marta.pastina@embrapa.br

Assigned to Associate Editor Timothy Beissinger.

Funding information

Embrapa (Brazilian Agricultural Research Corporation); CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior); CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico); FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais)

Abstract

Genotyping-by-sequencing (GBS) datasets typically feature high rates of missingness and heterozygote undercalling, prompting the use of data imputation. We compared the accuracy of four imputation methods—NPUTE, Beagle, *k*-nearest neighbors imputation (KNNI), and fast inbreed line library imputation (FILLIN)—using GBS data of maize (*Zea mays* L.) inbred lines, genotyped using different multiplexing levels. Two strategies for SNP-calling and genotype imputation were evaluated. First, only lines genotyped through 96-plex were used for single nucleotide polymorphism (SNP) discovery, whereas both 96- and 384-plex were simultaneously used in the second strategy. In the first genotype imputation strategy, only the 96-plex lines were imputed, then the remaining lines were appended (96-plex-imputed plus 384-plex) and then imputed. In the second imputation strategy, we jointly imputed both datasets. We also investigated the impacts of including heterozygous genotypes and distinct rates of missing genotypes per locus. The different SNP-calling strategies and percentage of missing data did not substantially affect the imputation accuracy. However, the different imputation strategies showed a substantial effect. Generally, imputations were less accurate for heterozygotes. The scenario 96-plex-imputed plus 384-plex showed accuracies similar to the 96-plex scenario. Beagle and NPUTE produced the highest accuracies. Our results indicate that combining SNP-calling and imputation strategies can enhance genotyping in a cost-effective manner, resulting in higher imputation accuracies.

1 | INTRODUCTION

The emergence of next-generation sequencing technology presented the possibility of obtaining molecular

Abbreviations: FILLIN, fast inbreed line library imputation; GBS, genotyping-by-sequencing; InDel, insertion–deletion; KNNI, *k*-nearest neighbors imputation; MAF, minor allele frequency; SNP, single nucleotide polymorphism.

markers densely distributed across the genome using high-throughput techniques such as GBS (Elshire et al., 2011). Making use of these genome-wide genotyping platforms, genomic selection and genome-wide association studies offer great potential to accelerate and enhance selection efficiency of plant breeding programs (Chang, Toghiani, Ling, Aggrey, & Rekaya, 2018; Desta & Ortiz, 2014; Dias et al., 2018; Faville et al., 2018; Gerard, Kobiljski, Lohwasser, Börner, & Simón, 2018; Haile et al., 2018; Kayondo et al., 2018). However, the costs of high-density genotyping for large numbers of individuals are still infeasible, representing a barrier to a more widespread adoption of these tools. Because the accuracy of genomic selection and power of association studies usually increase with increasing numbers of individuals and density of markers, low-cost genotyping strategies have to be adopted to address resource limitations (Cericola et al., 2018; Han et al., 2018; Jacobson, Lian, Zhong, & Bernardo, 2015).

The adoption of genomic selection in a maize breeding program allows breeders to genotype elite lines and to predict the performance of all possible hybrids even if they are not phenotypically evaluated. This strategy reduces the costs and labor involved in field trials and can increase genetic gains. In any case, cost-effective genotyping is crucial. Key features of GBS data are the high rates of missingness, heterozygote undercalling, and nonuniform distribution of sequence reads, which vary according to the kind of population and multiplexing level (Beissinger et al., 2013). Several studies have reported the efficiency of imputing missing data using different methods and strategies (Bouwman, Hickey, Calus, & Veerkamp, 2014; Cleveland, Hickey, & Kinghorn, 2011; Gonen et al., 2018; Hickey, Crossa, Babu, & De Los Campos, 2012; Howie, Donnelly, & Marchini, 2009; Nazzicari, Biscarini, Cozzi, Brummer, & Annicchiarico, 2016; Swarts et al., 2014). An effective strategy involves genotyping some of the individuals at higher marker density, then using these high-density data to impute larger numbers of individuals genotyped at lower marker density. Genomic selection studies adopting this approach reported increases in the predictive accuracy (Gorjanc et al., 2017a; 2017b; Jacobson et al., 2015). For species for which genotyping chips are available, as is the case of economically important animals and some crops, combining data from high- and low-density SNP arrays is a cost-effective strategy (Gorjanc et al., 2017b; Hickey, Gorjanc, Varshney, & Nettelblad, 2015; Jacobson et al., 2015). When genotyping chips are not available or their use is prohibitive, the GBS technology allows breeders to adjust the amount of retrieved information and its cost in different ways (Gorjanc et al., 2017b). For instance, choosing different restriction enzymes, regulating sequencing depth, and the level of multiplexing (Deschamps, Llaca, & May, 2012; Elshire et al., 2011; Poland & Rife, 2012). However, to the best of our knowledge, no studies have yet empirically investigated the combined use of SNP-calling and imputation strategies to improve GBS data quality.

Core ideas

- Combining SNP-calling and imputation strategies can enhance cost-effective genotyping
- SNP-calling strategies and percentage of missing data did not affect the imputation accuracy
- Beagle and NPUTE produced the highest accuracies

There are several imputation methods available but most of them were developed for humans (Browning & Yu, 2009; Howie et al., 2009; Liu, Li, Wang, & Li, 2013). However, humans are highly heterozygous, obligate outcrossers, show little inbreeding, and much less structural variation than that observed in crop plants. These factors make the imputation methods designed for humans not necessarily optimized for use in crop systems. For this reason, it is worthwhile to compare different imputation methods, which may or may not allow for heterozygous genotypes. Situations in breeding programs where there are genotypic datasets with varying levels of multiplexing and heterozygosity are increasingly common. There is therefore scientific and practical interest in gaining knowledge about how to better explore such datasets in order to achieve high imputation accuracies.

In this paper, we compared the imputation accuracy of four imputation methods: NPUTE (Roberts et al., 2007), Beagle (Browning & Browning, 2007), KNNI (Troyanskaya et al., 2001), and FILLIN (Swarts et al., 2014), which are well known algorithms implemented in freely available software libraries. We imputed missing genotypes from GBS data of maize inbred lines genotyped using different levels of multiplexing per sequencing lane. We evaluated different SNP-calling strategies in order to better explore the low and high multiplexing levels of our dataset. Because this dataset represents a panel of maize inbred lines mostly in final stages of the breeding program, we expected that these lines were homozygous for the majority of loci. However, a few of those lines were in initial stages of the breeding program and could thus have higher heterozygosity rates. Hence, we also evaluated the impact of including heterozygous genotypes on imputation accuracy. The main objective of this study was to evaluate different SNP-calling and imputation strategies in a real maize breeding program scenario. By doing so, we aimed to better explore the possibility of using both low and high multiplexing levels to deliver cost-effective genotyping without compromising the imputation accuracy.

2 | MATERIALS AND METHODS

2.1 | Experimental data

Data used in this study came from a collection of 1,060 maize inbred lines from the Embrapa Maize and Sorghum breeding program. These lines represent dent (34%) and flint (51%) heterotic groups, as well as another group—here called Group C (15% of the lines), which is unrelated to both dent and flint sources. We performed DNA extraction from young leaves based on the cetyl trimethylammonium bromide method (Saghai-Maroof, Soliman, Jorgensen, & Allard, 1984). The DNA samples were quantified using the Fluorometer Qubit 2.0 following the manufacturer's instructions (Life TechnologiesTM). Samples were also evaluated on 1% agarose gel in Tris acetate-EDTA buffer, stained with GelRedTM (Biotium) and recorded under UV light in the Imager Gel Doc L-PIX (Loccus Biotecnologia). Genotyping-by-sequencing was carried out at the Genomic Diversity Facility at Cornell University (Ithaca, NY, USA) using the standard GBS protocol (Elshire et al., 2011) with the restriction enzyme ApeKI. The inbred lines were split into two groups: (a) 680 lines genotyped using 96 samples per sequencing lane (HiSeq2500, 1 × 100 bp) and (b) 380 lines genotyped with 384 samples per lane (NextSeq500, 1 × 90 bp). We thus expected a larger number of reads per sample in the first group. Tags were aligned to the B73 reference genome (AGPv3) (Law et al., 2015) using the Bowtie2 aligner (Langmead & Salzberg, 2012). Then, SNPs were called using the GBSv2 Discovery Pipeline, available in the software TASSEL v. 5.2.28 (Glaubitz et al., 2014), using different strategies as shown below.

2.2 | Single nucleotide polymorphism calling strategies

We evaluated different SNP-calling and imputation strategies as summarized in Figure 1. In our first strategy, denoted as SNP-calling strategy I, we ran the discovery SNP caller plugin using only the 680 lines genotyped with 96 samples per sequencing lane. In this scenario, only the lines with higher depth of coverage, which are likely to have less missing data and lower genotyping error probability, were used for SNP discovery. We thus expected this subset of the data to provide a set of high quality SNPs with greater power of detection of polymorphic sites and less false positives. Next, we ran the production SNP caller plugin with all 1,060 lines. By doing so, all lines were effectively genotyped but only for the loci detected in the first set (Figure 1a, blue color). For SNP-calling strategy II, we ran the discovery and production SNP caller plugins combining both the high- and low-multiplexing sets of lines. This strategy was likely to affect the number and quality of discovered SNPs because we included lines genotyped with 384 samples per sequencing lane throughout the SNP discovery step (Figure 1a, gray color). Finally, we evaluated the descriptive statistics generated for each discovered marker and applied filters for minor allele frequency (MAF) <5% and inbreeding coefficient <0.8. Only SNPs that passed both filters were used for further analyses.

2.3 | Imputation scenarios

We initially cleaned the two datasets by removing insertion–deletion (InDel) and nonbiallelic markers, because calling of InDel and multiallelic variants from low-coverage sequencing can be unreliable. Moreover, some imputation methods used do not allow InDel and multiallelic markers, and removing them guaranteed that all comparisons were valid. Next, we used two contrasting approaches to compare the influence of heterozygous genotypes by either keeping or removing any nonhomozygous genotype calls. Because our dataset contains a collection of maize inbred lines mostly in final stages of the breeding program (F_6–F_7), we expected that these lines were homozygous for the majority of locus. However, a few of those lines were in initial stages of inbreeding (F_3–F_4) and could thus have higher heterozygosity rates (up to 25 and 12.5%, respectively).

For each of these scenarios, we then evaluated two different imputation strategies to leverage the varying levels of multiplexing (Figure 1b). Toward this end, we first imputed only the 680 lines genotyped with 96 samples per sequencing lane. We expected that the imputation accuracy of this dataset would be higher because these lines have higher depth of coverage. Later we appended to these imputed data the remaining 380 lines, which were genotyped with 384 samples per sequencing lane and finally performed the imputation of the remaining missing data. The competing strategy consisted of jointly imputing the high and low multiplexing datasets in a single step.

In addition to the imputation strategies, we aimed to evaluate the impacts of the rate of missing data per marker on the imputation accuracies. Then, we filtered the SNP data to have a maximum of 10, 20, 50 or 80% missing data per marker, generating four subdatasets. We used these four subdatasets, in addition to the unfiltered dataset,

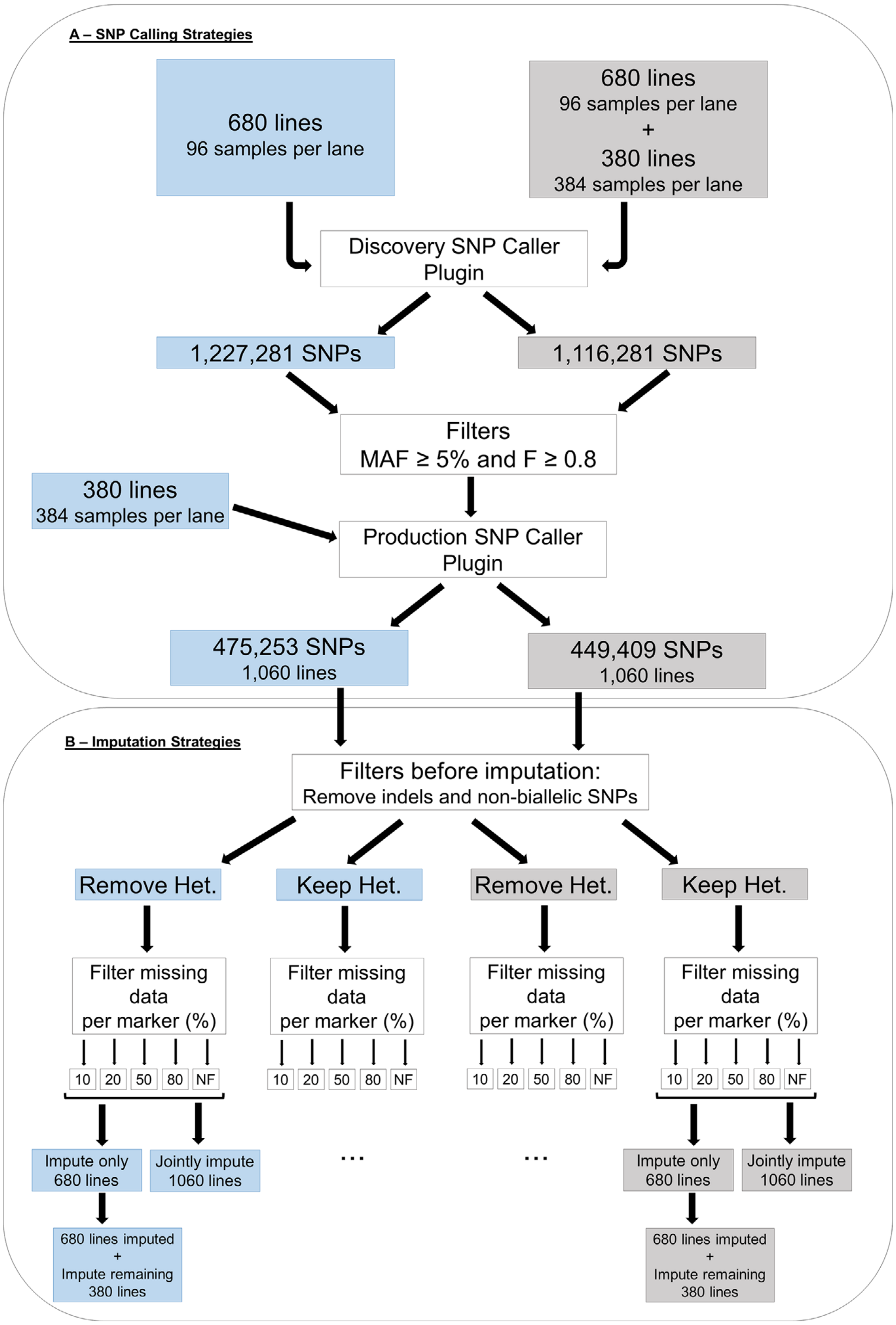

FIGURE 1 Summary of (a) single nucleotide polymorphism (SNP)-calling and (b) imputation strategies. Blue color indicates SNP-calling strategy I, whereas gray color indicates SNP-calling strategy II. The imputation strategies showed in (b) were applied to all the four imputation methods evaluated. MAF, minor allele frequency; F, inbreeding coefficient; Het, heterozygous; NF, not filtered

to perform the imputation analyses. For each of these datasets, we randomly introduced an additional 10% missing genotypes based on which imputation accuracy could be measured. All the SNP-calling and imputation strategies are summarized in Figure 1.

2.4 | Imputation methods and software

We used the software TASSEL v.5.2.28 (Glaubitz et al., 2014) and the open-source environment for statistical programming R (R Core Team, 2018) for data handling, editing, summarizing results, and figure design. We performed the subsequent imputation analyses using the filtered datasets from the two competing SNP-calling strategies. We evaluated four imputation methods: (a) NPUTE software v.1 (Roberts et al., 2007), which is based on the similarities between haplotypes of different individuals for the same genomic region in which different window imputation sizes were tested for each chromosome, and the windows with higher accuracies were chosen; (b) Beagle (version 4.1 using default parameters [Browning & Browning, 2016]; Browning & Browning, 2007), which was originally developed for human genetic studies but also presents a wide application in animal and plant genetics (Law et al., 2015; Nazzicari et al., 2016), infers haplotypes and imputes missing alleles both with known and unknown linkage phase using a stochastic procedure based on hidden Markov models to find the most likely haplotype pair for each individual, a method that works iteratively using an expectation–maximization approach; (c) KNNI (Troyanskaya et al., 2001), which is a method based on the weighted average of the *k* closest markers, used in R using the function *KNNIcatimpute* from the R package *Scrime* (Schwender & Fritsch, 2015); and (d) FILLIN (Swarts et al., 2014), an imputation method optimized for inbred populations implemented in the software TASSEL v. 5 (Glaubitz et al., 2014), which is based on haplotype reconstruction around recombination break points. With FILLIN, imputation is carried out in two steps: first, the inference of haplotype takes place (FILLINFindHaplotypesPlugin), which is followed by imputation of missing data based on the resulting haplotypes (FILLINImputationPlugin). We used the TASSEL v.5.2.28 (Glaubitz et al., 2014) plugin FILLINFindHaplotypesPlugin followed by FILLINImputationPlugin to perform the imputation procedure, considering the options –accuracy and –proSitesMask to calculate the accuracy.

2.5 | Imputation accuracy and computational time

For each imputation scenario, we used the artificial missing genotypes to measure the overall imputation accuracy and the accuracy for each genotype class. The imputation accuracy was computed as the proportion of correct imputation measured as the number of correctly imputed missing data divided by the total number of artificially missing data points.

We also measured the computational time required for imputation to be completed in each analysis as an indicator of the software relative performance. To ensure consistency, all jobs were separately submitted to the same computing platform, a multinode server with two Intel Xeon E5-2650 v4 at 2.20 GHz CPUs with a total of 48 threads and 256 GB of RAM.

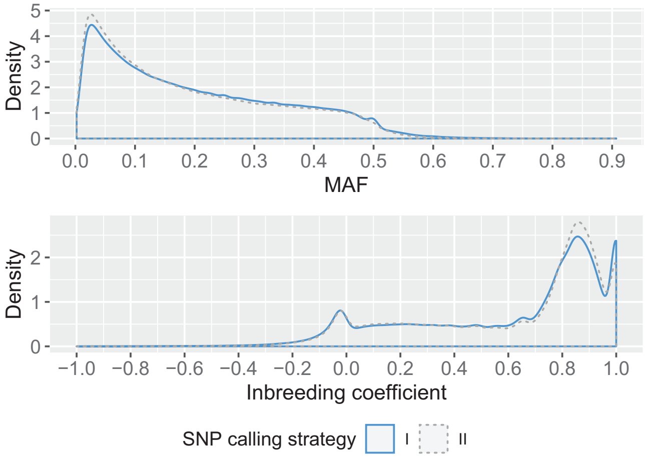

FIGURE 2 (a) Minor allele frequency (MAF) and (b) inbreeding coefficient distributions for two alternative single nucleotide polymorphism (SNP) calling strategies. Solid blue line corresponds to SNP-calling strategy I, that is, SNP identification using only the 680 lines genotyped with 96 samples per sequencing lane, followed by genotype calling with all 1060 lines. Dashed gray line corresponds to SNP-calling strategy II, that is, SNP identification and genotype calling using all 1060 samples. A total of 1,227,281 and 1,116,281 markers are represented in SNP-calling strategies I and II, respectively

3 | RESULTS

3.1 | Genotypic data

Using our SNP-calling strategy I, that is, with only the 680 lines with higher depth of coverage for SNP discovery, we found 1,227,281 SNPs. We initially removed SNPs that did not pass the MAF and inbreeding coefficient filters, generating 475,253 SNPs (Figure 2). Comparatively, we found 1,116,281 SNPs with our SNP-calling strategy II, that is, using all the 1060 lines with high and low depths of coverage for SNP discovery. We again removed SNPs with the same filtering criteria, resulting in 444,409 SNPs (Figure 2). As expected, the number of SNPs found with SNP-calling strategy II was slightly smaller, possibly a result of the lower detection power with this higher multiplexing dataset. The mean depth of coverage was 3.04 and 0.77 reads per locus per sample for the lines genotyped using 96 and 384 samples per sequencing lane, respectively. The number of markers found per chromosome followed similar patterns in both SNP-calling strategies (Supplemental Figure S1). In addition to the higher number of markers found with SNP-calling strategy I, the number of missing data per locus was also higher (Supplemental Figure S2). After removing InDel and nonbiallelic markers,

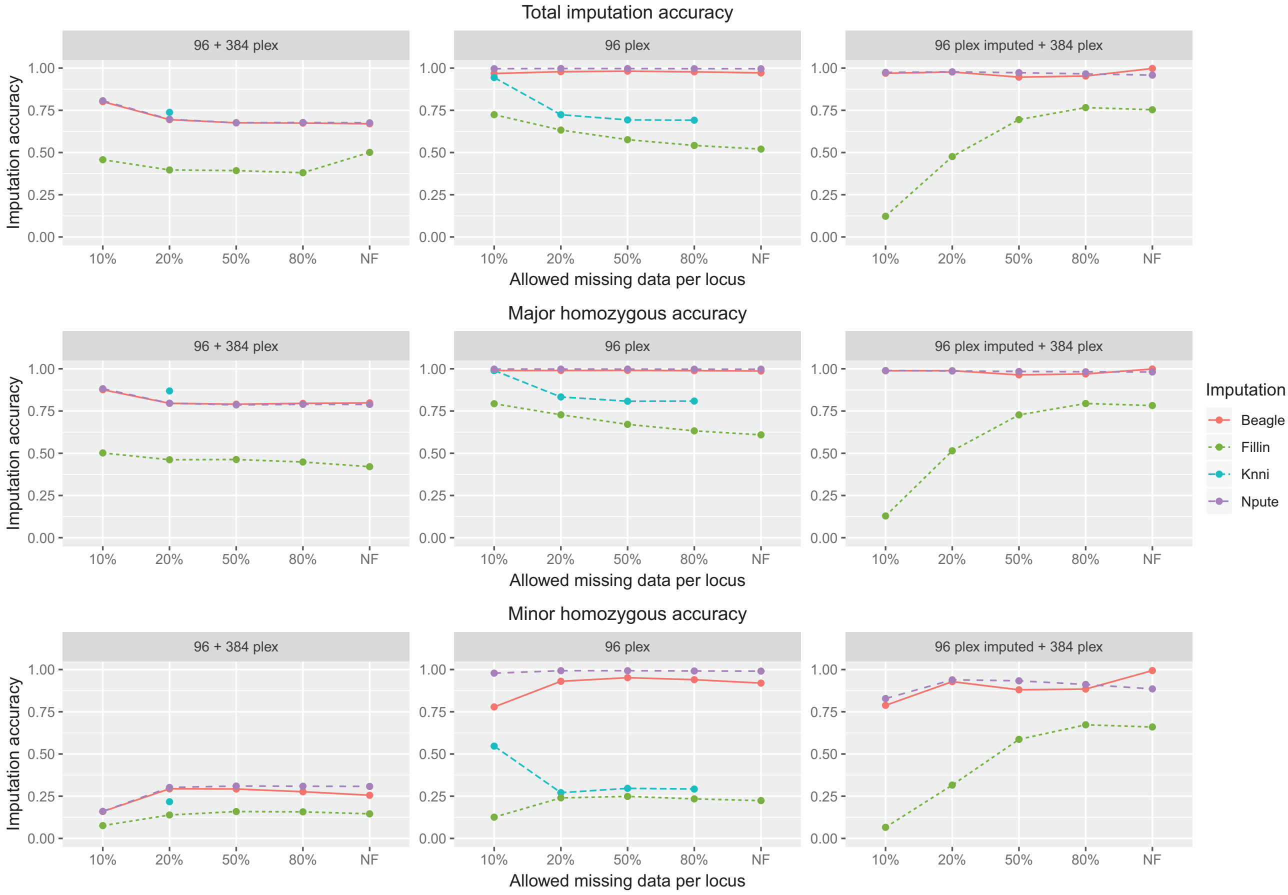

FIGURE 3 Imputation accuracy using single nucleotide polymorphism (SNP) calling strategy I, that is, SNP identification using only the 680 lines genotyped with 96 samples per sequencing lane, followed by genotype calling with all 1060 lines, removing heterozygous genotypes. Each row represents imputation accuracy for different genotypic classes: total imputation accuracy, major homozygous accuracy, and minor homozygous accuracy. Each column represents an imputation strategy: 96- plus 384-plex, 96-plex, and 96-plex-imputed plus 384-plex. The *x*-axis represents the different filters of allowed missing data per locus (10, 20, 50, and 80% and not filtered). Line colors represent the four imputation methods: Beagle (solid red), FILLIN (dotted green), KNNI (dashed blue), and NPUTE (dashed purple)

the different filters of allowed missing data (10, 20, 50, and 80% and no filter) generated 12,957; 42,053; 173,328; 368,351; and 474,367 markers, respectively, for SNP-calling strategy I and 17,508; 50,793; 187,440; 380,955; and 443,940 markers, respectively, for SNP-calling strategy II.

3.2 | Imputation accuracy

Accuracies are reported for each combination of SNP-calling and imputation strategies (Figures 3 through 6). Comparing the imputation accuracies between the SNP-calling strategies I and II, we did not observe pronounced differences. When removing heterozygous markers, we observed that Beagle and NPUTE outperformed all other imputation methods in most scenarios evaluated (Figures 3 and 4). The KNNI method presented a computational limitation and in most evaluated scenarios did not run to completion. Interestingly, however, in some cases it outperformed all other methods. For example, in the joint imputation of the whole dataset, the total and major homozygous accuracies of the KNNI method were slightly higher than all other methods when the rate of allowed missing data per locus was 20%. The FILLIN method resulted in considerably smaller accuracies in all scenarios evaluated.

Contrary to our expectations, the allowed missing data per locus did not substantially adversely affect the imputation accuracy, with most methods showing a (nearly) flat response to increased missing data (Figures 3 and 4). The KNNI method showed decreasing imputation accuracy with increasing missing data. On the other hand, FILLIN showed increasing imputation accuracy with increasing missing data in the imputation strategy 96-plex imputed plus 384-plex and, particularly in the SNP-calling strategy II, for the 96 plus 384-plex scenario (Figure 4).

When not removing the heterozygous genotypes, we assessed the imputation accuracy with only the Beagle

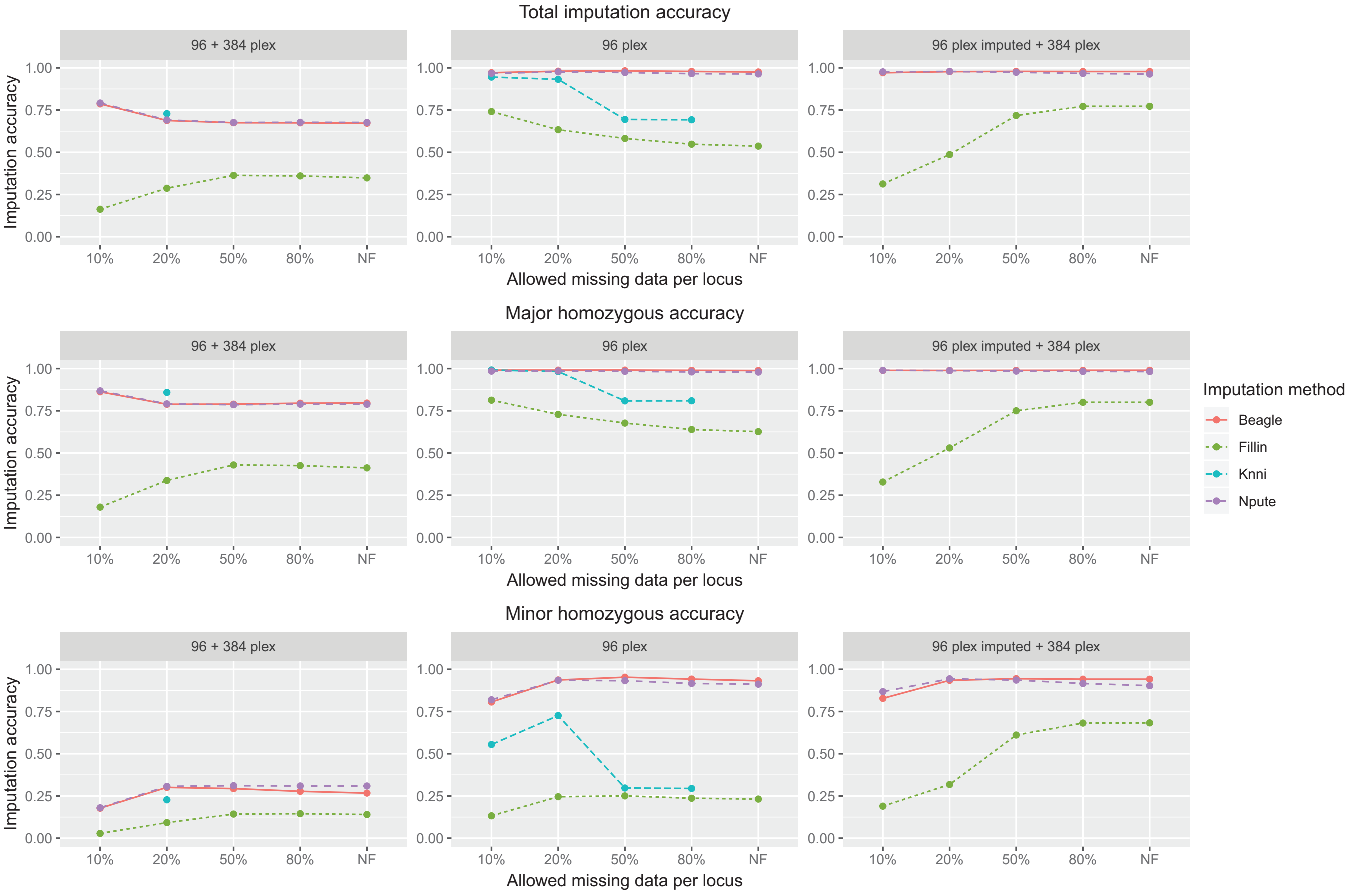

FIGURE 4 Imputation accuracy using single nucleotide polymorphism (SNP) calling strategy II, that is, SNP identification and genotype calling using all 1060 samples, removing heterozygous markers. Each row represents imputation accuracy for different genotype classes: total imputation accuracy, major homozygous accuracy, and minor homozygous accuracy. Each column represents an imputation strategy: 96- plus 384-plex, 96-plex, and 96-plex-imputed plus 384-plex. The *x*-axis represents the different filters of allowed missing data per locus (10, 20, 50, and 80% and not filtered). Line colors represent the four imputation methods: Beagle (solid red), FILLIN (dotted green), KNNI (dashed blue), and NPUTE (dashed purple)

method because it accepts all genotype classes, while NPUTE and FILLIN require exclusively homozygous genotypes. Although the KNNI method accepts heterozygous genotypes, its computational limitations precluded further evaluation. The accuracy for heterozygous genotypes was considerably lower than for the two homozygous genotypes (Figures 5 and 6). Again, the allowed missing data per locus did not affect the imputation accuracy.

We observed extensive differences between the two different imputation strategies (Figures 3 through 6). As expected, the 96-plex imputation scenario always showed the best accuracies, while the 96- plus 384-plex performed the worst. Interestingly, the imputation scenario 96-plex-imputed plus 384-plex showed imputation accuracies similar to the 96-plex scenario. This opens up the possibility of combining high- and low-depth GBS data without compromising the imputation accuracy.

3.3 | Computation time

We tracked the amount of time required to complete the imputation process for each method and imputation scenario in each SNP-calling strategy with and without removing heterozygotes (Tables 1 and 2). The number of markers for the different filters of allowed missing data varied considerably, which reflected in the computation times. The FILLIN method required, by far, the least computational times (Table 1). Even in the most complex scenario, that is, a higher number of markers and 96-plex-imputed plus 384-plex, the time required to complete the imputation process was never more than a few minutes. The second fastest method was Beagle, which however required noticeably more time in more complex scenarios (Tables 1 and 2). The NPUTE method was the slowest algorithm overall except for some scenarios where KNNI was slower (Table 1).

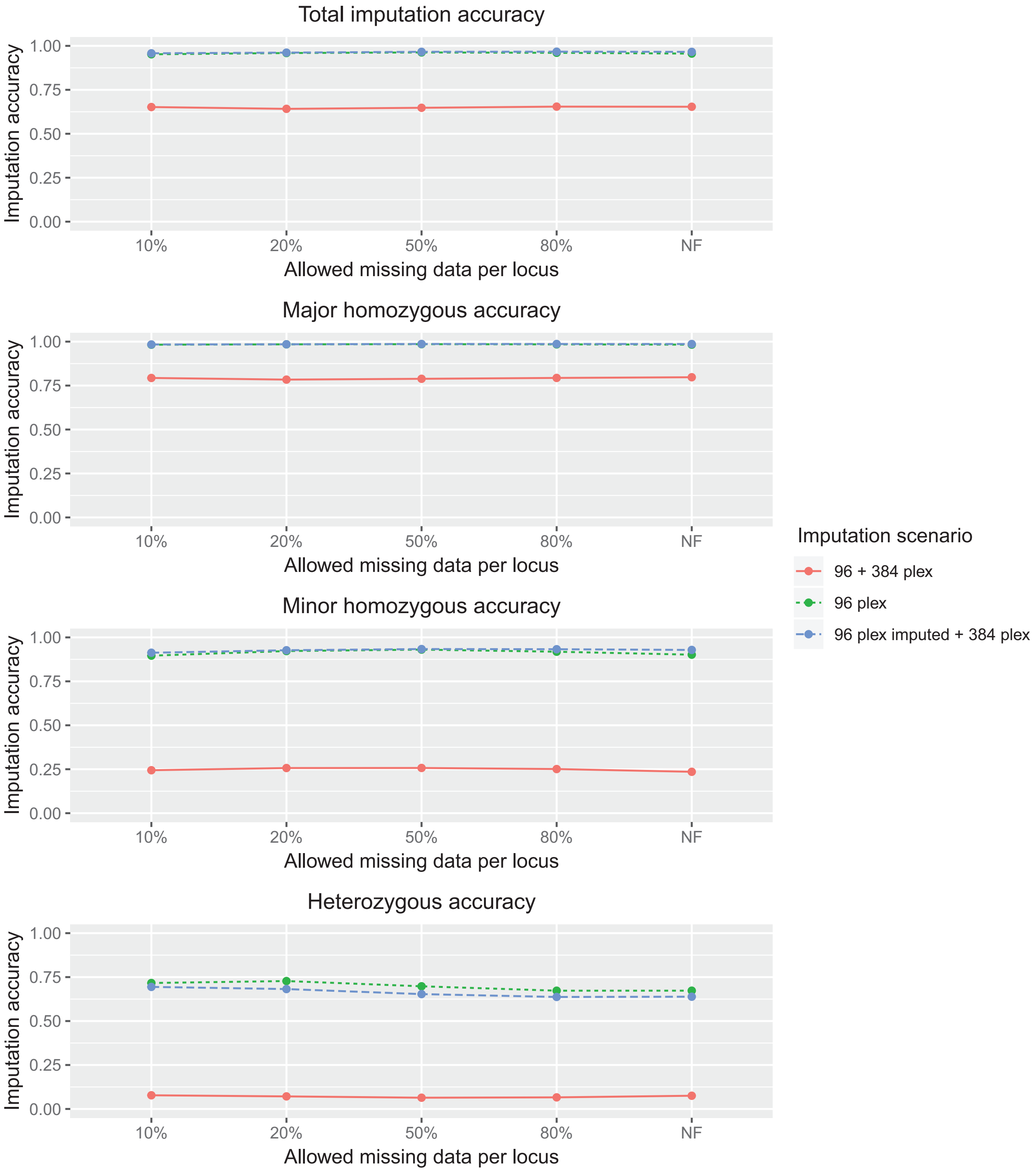

FIGURE 5 Imputation accuracy using single nucleotide polymorphism (SNP) calling strategy I, that is, SNP identification using only the 680 lines genotyped with 96 samples per sequencing lane, followed by genotype calling with all 1060 lines, not removing heterozygous markers for the Beagle imputation method. Each row represents imputation accuracy for different genotype classes: total imputation accuracy, major homozygous accuracy, minor homozygous accuracy, and heterozygous accuracy. The *x*-axis represents the different filters of allowed missing data per locus (10, 20, 50, and 80% and not filtered). Line colors represent the three imputation scenarios: 96 plus 384 plex (solid red), 96 plex (dotted green), and 96-plex-imputed plus 384-plex (dashed blue)

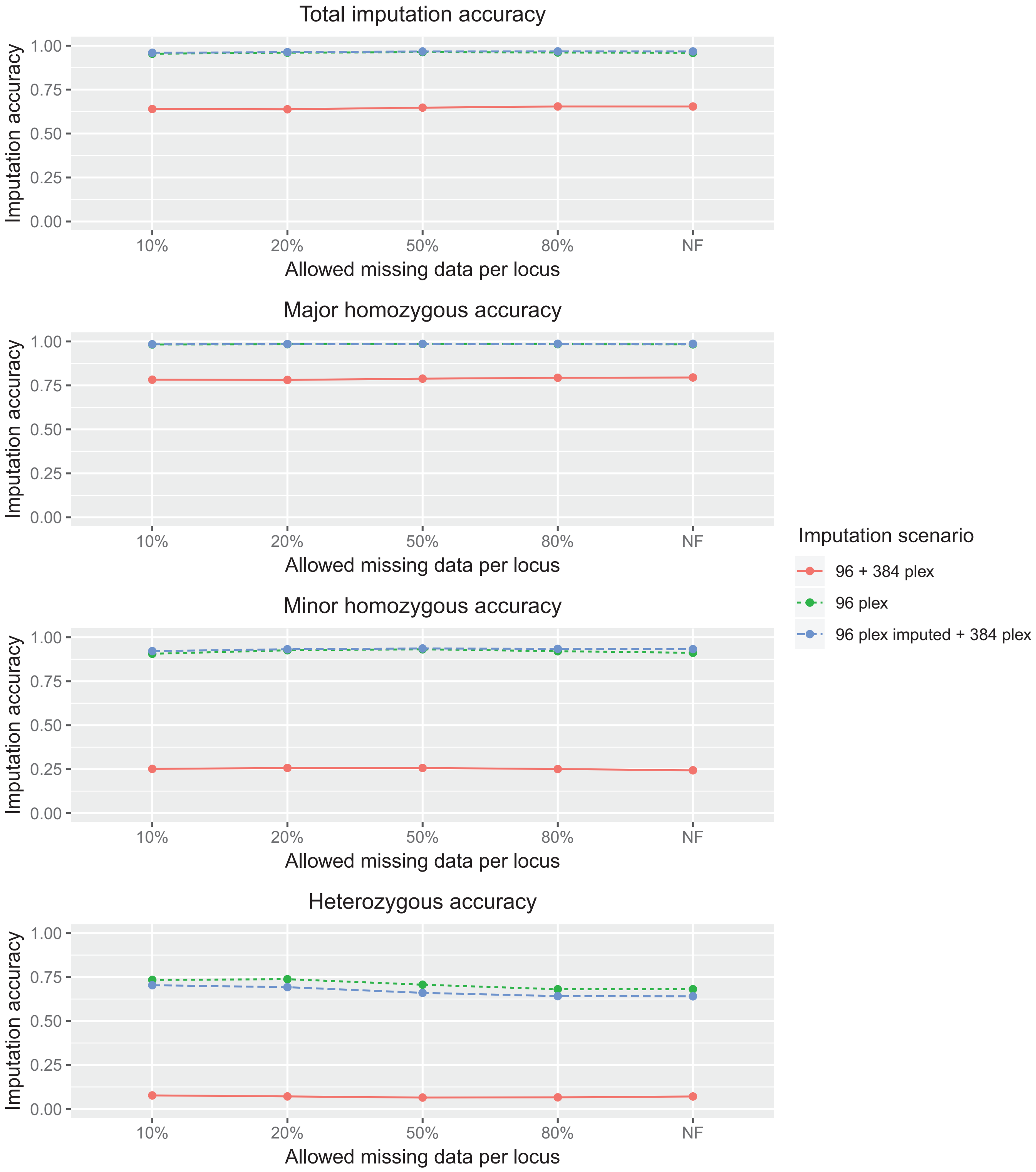

FIGURE 6 Imputation accuracy using single nucleotide polymorphism (SNP) calling strategy II, that is, SNP identification and genotype calling using all 1060 samples, not removing heterozygous markers for the Beagle imputation method. Each row represents imputation accuracy for different genotype classes: total imputation accuracy, major homozygous accuracy, minor homozygous accuracy, and heterozygous accuracy. The *x*-axis represents the different filters of allowed missing data per locus (10, 20, 50, and 80% and not filtered). Line colors represent the three imputation scenarios: 96- plus 384-plex (solid red), 96-plex (dotted green), and 96-plex-imputed plus 384-plex (dashed blue)

TABLE 1 Running times for the different imputation methods for each allowed missing data per locus (10, 20, 50, and 80% and not filtered [NF]) and respective number of markers in each imputation and single nucleotide polymorphism (SNP) calling strategy removing the heterozygous markers

SNP-calling strategy[a]	Imputation strategy	Allowed missing data per locus	No. of markers	Imputation method	Running time
		%			HH:MM:SS
I	96-plex	10	12,957	Beagle	00:02:37
				KNNI	00:07:29
				NPUTE	00:42:39
				FILLIN	00:00:12
		20	42,053	Beagle	00:18:18
				KNNI	00:56:47
				NPUTE	04:24:57
				FILLIN	00:00:35
		50	173,328	Beagle	01:23:00
				KNNI	21:19:08
				NPUTE	20:56:10
				FILLIN	00:01:51
		80	368,351	Beagle	02:58:00
				KNNI	107:06:20
				NPUTE	38:34:20
				FILLIN	00:03:49
		NF	474,367	Beagle	03:25:00
				KNNI	–
				NPUTE	41:25:49
				FILLIN	00:04:30
	96-plex imputed plus 384-plex	10	12,957	Beagle	04:17:24
				KNNI	–
				NPUTE	02:48:06
				FILLIN	00:00:13
		20	42,053	Beagle	29:28:46
				KNNI	–
				NPUTE	15:58:28
				FILLIN	00:00:54
		50	173,328	Beagle	43:28:34
				KNNI	–
				NPUTE	84:54:03
				FILLIN	00:02:53
		80	368,351	Beagle	61:54:41
				KNNI	–
				NPUTE	163:19:23
				FILLIN	00:07:03
		NF	474,367	Beagle	54:25:08
				KNNI	–
				NPUTE	189:12:28
				FILLIN	00:07:07

(Continues)

TABLE 1 (Continued)

SNP-calling strategy[a]	Imputation strategy	Allowed missing data per locus	No. of markers	Imputation method	Running time
	96- plus 384-plex	10	12,957	Beagle	01:11:34
				KNNI	–
				NPUTE	01:30:01
				FILLIN	00:00:15
		20	42,053	Beagle	02:57:11
				KNNI	07:32:04
				NPUTE	09:28:58
				FILLIN	00:00:45
		50	173,328	Beagle	04:17:45
				KNNI	07:32:04
				NPUTE	43:46:57
				FILLIN	00:02:49
		80	368,351	Beagle	06:16:15
				KNNI	–
				NPUTE	77:52:46
				FILLIN	00:05:28
		NF	474,367	Beagle	06:21:13
				KNNI	–
				NPUTE	83:50:49
				FILLIN	00:06:33
II	96-plex	10	17,508	Beagle	00:03:41
				KNNI	00:15:52
				NPUTE	00:57:22
				FILLIN	00:00:18
		20	50,793	Beagle	00:27:06
				KNNI	06:53:51
				NPUTE	05:31:46
				FILLIN	00:00:36
		50	187,440	Beagle	01:40:45
				KNNI	10:10:12
				NPUTE	24:51:52
				FILLIN	00:02:05
		80	380,955	Beagle	03:26:19
				KNNI	117:56:04
				NPUTE	41:17:26
				FILLIN	00:04:05
		NF	443,940	Beagle	03:33:55
				KNNI	–
				NPUTE	42:56:24
				FILLIN	00:04:30

(Continues)

TABLE 1 (Continued)

SNP-calling strategy[a]	Imputation strategy	Allowed missing data per locus	No. of markers	Imputation method	Running time
	96-plex-imputed plus 384-plex	10	17,508	Beagle	03:14:19
				KNNI	–
				NPUTE	03:00:25
				FILLIN	00:00:20
		20	50,793	Beagle	18:47:42
				KNNI	–
				NPUTE	16:42:40
				FILLIN	00:00:50
		50	187,440	Beagle	42:42:46
				KNNI	–
				NPUTE	78:20:06
				FILLIN	00:03:07
		80	380,955	Beagle	69:25:25
				KNNI	–
				NPUTE	155:48:01
				FILLIN	00:07:41
		NF	443,940	Beagle	57:18:19
				KNNI	–
				NPUTE	179:39:53
				FILLIN	00:09:06
	96- plus 384-plex	10	17,508	Beagle	01:41:32
				KNNI	–
				NPUTE	02:52:14
				FILLIN	00:00:27
		20	50,793	Beagle	03:02:00
				KNNI	13:10:43
				NPUTE	15:59:20
				FILLIN	00:01:13
		50	187,440	Beagle	04:26:05
				KNNI	–
				NPUTE	48:18:38
				FILLIN	00:04:14
		80	380,955	Beagle	06:28:54
				KNNI	–
				NPUTE	74:41:06
				FILLIN	00:08:41
		NF	443,940	Beagle	08:09:11
				KNNI	–
				NPUTE	81:20:55
				FILLIN	00:09:37

[a] SNP-calling strategy I: SNP identification using only the 680 lines genotyped with 96 samples per sequencing lane, followed by genotype calling with all 1060 lines; SNP-calling strategy II: SNP identification and genotype calling using all 1060 samples.

TABLE 2 Running times for Beagle for each allowed missing data per locus (10, 20, 50, and 80% and not filtered [NF]) and respective number of markers, in each imputation and single nucleotide polymorphism (SNP) calling strategy, considering the heterozygous markers

SNP-calling strategy[a]	Imputation strategy	Allowed missing data per locus	No. of markers	Running time
		%		HH:MM:SS
I	96-plex	10	12,957	00:27:02
		20	42,053	01:08:07
		50	173,328	04:01:21
		80	368,351	04:51:47
		NF	474,367	05:20:26
	96-plex-imputed plus 384-plex	10	12,957	02:28:25
		20	42,053	05:29:07
		50	173,328	11:43:59
		80	368,351	17:24:06
		NF	474,367	21:10:02
II	96- plus 384-plex	10	12,957	01:28:59
		20	42,053	02:56:30
		50	173,328	07:33:40
		80	368,351	08:29:33
		NF	474,367	11:16:18
	96-plex	10	17,508	00:21:33
		20	50,793	01:00:56
		50	187,440	03:12:39
		80	380,955	06:25:55
		NF	443,940	07:13:19
	96-plex-imputed plus 384-plex	10	17,508	03:08:15
		20	50,793	06:55:54
		50	187,440	13:15:22
		80	380,955	20:52:30
		NF	443,940	21:48:30
	96- plus 384-plex	10	17,508	01:47:01
		20	50,793	03:17:01
		50	187,440	06:44:40
		80	380,955	12:24:20
		NF	443,940	12:41:05

[a] SNP-calling strategy I: SNP identification using only the 680 lines genotyped with 96 samples per sequencing lane, followed by genotype calling with all 1060 lines; SNP-calling strategy II: SNP identification and genotype calling using all 1060 samples.

In general, running times for the imputation strategy 96-plex were much lower than for situations that included the samples sequenced at lower depth. The second imputation strategy in amount of time required was the 96-plex plus 384-plex. Finally, the 96-plex-imputed plus 384-plex imputation strategy largely exceeded the others in amount of time required to complete the imputation process (Tables 1 and 2).

4 | DISCUSSION

Results from our study indicate that combining SNP-calling and imputation strategies can enhance cost-effective genotyping, resulting in higher imputation accuracies. These approaches thus allow a more widespread adoption of genomic selection and genome-wide association studies in plant breeding programs. The different

SNP-calling strategies aimed to better explore the high- and low-multiplexing levels of our dataset and yielded different number of markers. Even after we removed SNPs with MAF <5% and inbreeding coefficient <0.8, the SNP-calling strategy I, that is, using only the high coverage dataset to discover SNPs, produced 30,000 more markers than the alternative scheme. Across the MAF plots, we observed that MAFs closer to zero were less frequent using only lines genotyped with 96 samples per sequencing lane (Figure 2). The higher coverage dataset likely enabled greater power of detection and less false positives. With regard to the inbreeding coefficient, both SNP-calling strategies showed similar patterns, with slightly higher values with SNP-calling strategy II (Figure 2). It is more difficult to call heterozygous SNPs with the lower-depth dataset, such that the homozygous genotypes tend to be called more frequently (Swarts et al., 2014). Overall, the SNP-calling strategies did not greatly affect the imputation accuracy but had influence on the number of markers found (Figures 3 through 6).

In SNP genotype imputation, it is important to evaluate not only the total imputation accuracy but also the per-class accuracy. Data balancedness refers to the ratio that each genotype class (AA, AB, BB) appears. Classification problems are more difficult when the data is unbalanced, that is, the three classes appear at different frequencies in the dataset. Data balancedness is directly related to MAF because very low MAFs arise when a class is underrepresented (Hickey et al., 2012; Nazzicari et al., 2016). Insofar as the classes of missing genotypes appear at different frequencies, the total imputation accuracy can be dominated by the most frequent class. The imputation accuracy tends to be higher for the more frequent genotype class, and the overall imputation accuracy will predominantly represent the imputation accuracy of that class. Indeed, we found significantly higher error rates in the less frequent class for all imputation methods.

Beagle and NPUTE methods produced the best imputation results with accuracies close to 100% in the imputation strategies 96-plex and 96-plex-imputed plus 384-plex in most scenarios of missing data. The KNNI method did not work in most evaluated scenarios. With large amounts of missing data, the complexity of the imputation problem increases and complicates the identification of *k* neighbors that are close enough to the data point to be imputed (Nazzicari et al., 2016). Probably as a consequence of the curse of dimensionality (Marimont & Shapiro, 1979), scenarios with large amounts of missing data could not be imputed with KNNI. The FILLIN method performed poorly in all tested scenarios, which may be explained by the fact that this algorithm is optimized for homogeneous inbred populations, while our dataset consists of a collection of lines from different heterotic groups. Similar results using Beagle, KNNI, and FILLIN methods were observed in a study with GBS data from rice (*Oryza sativa* L.) and alfalfa (*Medicago sativa* L.) (Nazzicari et al., 2016).

The allowed missing data per locus did not reflect on the imputation accuracy for Beagle and NPUTE methods (Figures 3 and 4). Nonetheless, in the 96-plex imputation strategy, with larger quantities of missing genotypes, KNNI showed a decrease in imputation accuracy. In the 96-plex imputation strategy, FILLIN also showed decreasing imputation accuracy with increasing missing rates for total and major homozygous imputation accuracies. However, overall in more stringent scenarios, with only 10 to 20% allowed missing data, imputation accuracy for all classes in the scenarios 96-plex-imputed plus 384-plex and 96-plex plus 384-plex as well as the minor homozygous accuracy in the scenario 96-plex were reduced for the FILLIN method.

Despite working with inbred maize lines, some of them were in the initial stages of the breeding program (F_3–F_4) and were not yet completely endogamic. Including heterozygous genotypes complicated the imputation problem, because this dataset showed relatively few heterozygotes, which are more susceptible to genotyping errors. As a consequence, the heterozygous accuracy was considerably lower than for both homozygote classes (Figures 3 and 4).

The complexity of the problem directly affected the running time required to complete the imputation process. The KNNI and NPUTE methods were the most demanding, with computation times growing both with the number of markers and with the number of missing genotypes to be imputed. The 96-plex-imputed plus 384-plex imputation strategy exceeded substantially the others in amount of time required. We believe that despite the smaller amount of missing data to impute, the initial step of identifying the haplotypes is likely more time consuming because there are more informative loci. Considering both imputation accuracy and computational time, the best imputation method was Beagle (Tables 1 and 2). In addition, this method allows for heterozygote genotypes, which is an interesting feature for panels that include individuals with few generations of inbreeding.

Several works have explored imputation strategies combining high- and low-density genotyping (Gonen et al., 2018; Gorjanc et al., 2017a; Hickey et al., 2012, 2015; Huang, Hickey, Cleveland, & Maltecca, 2012; Mulder, Calus, Druet, & Schrooten, 2012). These studies, however, do not focus on combining SNP-calling and imputation strategies using real GBS data. In this paper, we investigated the impact on imputation accuracies of combining different SNP-calling and imputation strategies using a real dataset of lines from a maize breeding program genotyped with GBS. We believe that our study is a first stage of what can be done regarding SNP-calling and imputation strategies with GBS data. Further research is necessary, for example, establishing the number of high-coverage individuals necessary to ensure high imputation accuracy of low-coverage individuals. We suggest that some key individuals could be genotyped using lower multiplexing levels, while others might be included in larger pools. This set of key individuals should be well-thought-out to represent the entire diversity of heterotic groups in the breeding program. Besides that, as sequencing costs continue to decrease, it should become feasible to increase sequencing depth for a larger portion of the samples or to increase the number of individuals in the genotyping panel at the same cost. Both approaches can increase overall genotyping and imputation accuracies as a result of the added information, allowing breeders to fine-tune the strategy to their particular needs. Finally, the accuracy of imputation may vary substantially depending on the genetic complexity of the species at hand. For instance, imputation in rice, which is nearly fully homozygous and has a reference genome available, outperformed imputation in alfalfa, a species with higher heterozygosity and which required the genome of a closely related species as a reference (Nazzicari et al., 2016).

Our results indicate that designing the SNP-calling and imputation strategies in order to better explore the different depths of coverage considerably improves the imputation accuracy and reduces costs since high-multiplexing levels are considerably cheaper. Bringing together SNP-calling strategies using only high-coverage data to discover variants followed by genotype calling for all sequenced samples with the imputation strategy 96-plex-imputed plus 384-plex produced the larger number of SNPs and higher imputation accuracies. These combined strategies encompass a wide range of applications in breeding programs representing an opportunity to reduce costs and time by optimizing the genotyping process.

ACKNOWLEDGMENTS

This research was supported by FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, program PREMIO 2045/2014, grant 23038.007195/2012-39), and Embrapa (Brazilian Agricultural Research Corporation). AAO received a fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ORCID

Amanda Avelar de Oliveira https://orcid.org/0000-0002-5595-6697
Maria Marta Pastina https://orcid.org/0000-0002-9112-3457
Gabriel Rodrigues Alves Margarido https://orcid.org/0000-0002-2327-0201

REFERENCES

Beissinger, T. M., Hirsch, C. N., Sekhon, R. S., Foerster, J. M., Johnson, J. M., Muttoni, G., … de Leon, Natalia (2013). Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, *193*, 1073–1081. https://doi.org/10.1534/genetics.112.147710

Bouwman, A. C., Hickey, J. M., Calus, M. P., & Veerkamp, R. F. (2014). Imputation of non-genotyped individuals based on genotyped relatives: Assessing the imputation accuracy of a real case scenario in dairy cattle. *Genetics Selection Evolution*, *46*, 6. https://doi.org/10.1186/1297-9686-46-6

Browning, B. L., & Browning, S. R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, *98*, 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Browning, B. L., & Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics*, *85*, 847–861. https://doi.org/10.1016/j.ajhg.2009.11.004

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, *81*, 1084–1097. https://doi.org/10.1086/521987

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C. S., Pedersen, M. G., … Janss, L. (2018). Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (*Lolium perenne* L.). *The American Journal of Human Genetics*, *9*, 369. https://doi.org/10.3389/fpls.2018.00369

Chang, L. Y., Toghiani, S., Ling, A., Aggrey, S. E., & Rekaya, R. (2018). High density marker panels, SNPs prioritizing and accuracy of genomic selection. *BMC Genetics*, *19*, 4. https://doi.org/10.1186/s12863-017-0595-2

Cleveland, M. A., Hickey, J. M., & Kinghorn, B. P. (2011). Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals. *BMC Proceedings*, *5*, S6. https://doi.org/10.1186/1753-6561-5-S3-S6

Deschamps, S., Llaca, V., & May, G. D. (2012). Genotyping-by-sequencing in plants. *Biology (Basel)*, *1*, 460–483. https://doi.org/10.3390/biology1030460

Desta, Z. A., & Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Science*, *19*, 592–601. https://doi.org/10.1016/j.tplants.2014.05.006

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., da Costa Silva, L., Parentoni, S. N., … Pastina, M. M. (2018). Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity*, *121*, 24–37. https://doi.org/10.1038/s41437-018-0053-6

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*, e19379. https://doi.org/10.1371/journal.pone.0019379

Faville, M. J., Ganesh, S., Cao, M., Jahufer, M. Z. Z., Bilton, T. P., Easton, H. S., … Barrett, B. A. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theoretical and Applied Genetics*, *131*, 703–720. https://doi.org/10.1007/s00122-017-3030-1

Gerard, G. S., Kobiljski, B., Lohwasser, U., Börner, A., & Simón, M. R. (2018). Genetic architecture of adult plant resistance to leaf rust in a wheat association mapping panel. *Plant Pathology*, *67*, 584–594. https://doi.org/10.1111/ppa.12761

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, *9*, e90346. https://doi.org/10.1371/journal.pone.0090346

Gonen, S., Wimmer, V., Gaynor, R. C., Byrne, Ed, Gorjanc, G., & Hickey, J. M. (2018). A heuristic method for fast and accurate phasing and imputation of single-nucleotide polymorphism data in biparental plant populations. *Theoretical and Applied Genetics*, *131*, 2345–2357. https://doi.org/10.1007/s00122-018-3156-9

Gorjanc, G., Battagin, M., Dumasy, J. F., Antolin, R., Gaynor, R. C., & Hickey, J. M. (2017a). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Science*, *57*, 216–228. https://doi.org/10.2135/cropsci2016.06.0526

Gorjanc, G., Dumasy, J. F., Gonen, S., Gaynor, R. C., Antolin, R., & Hickey J. M. (2017b). Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Science*, *57*, 1404. https://doi.org/10.2135/cropsci2016.08.0675

Haile, J. K., N'diaye, A., Clarke, F., Clarke, J., Knox, R., Rutkoski, J., … Pozniak, C. J. (2018). Genomic selection for grain yield and quality traits in durum wheat. *Molecular Breeding*, *38*, 75. https://doi.org/10.1007/s11032-018-0818-x

Han, S., Miedaner, T., Utz, H. F., Schipprack, W., Schrag, T. A., & Melchinger, A. E. (2018). Genomic prediction and GWAS of Gibberella ear rot resistance traits in dent and flint lines of a public maize breeding program. *Euphytica*, *214*, 6. https://doi.org/10.1007/s10681-017-2090-2

Hickey, J. M., Crossa, J., Babu, R., & De Los Campos, G. (2012). Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Science*, *52*, 654–663. https://doi.org/10.2135/cropsci2011.07.0358

Hickey, J. M., Gorjanc, G., Varshney, R. K., & Nettelblad, C. (2015). Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a Hidden Markov Model. *Crop Science*, *55*, 1934–1946. https://doi.org/10.2135/cropsci2014.09.0648

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*, e1000529. https://doi.org/10.1371/journal.pgen.1000529

Huang, Y., Hickey, J. M., Cleveland, M. A., & Maltecca, C. (2012). Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genetics Selection Evolution*, *44*, 25. https://doi.org/10.1186/1297-9686-44-25

Jacobson, A., Lian, L., Zhong, S., & Bernardo, R. (2015). Marker imputation before genomewide selection in biparental maize populations. *The Plant Genome*, *8*, 1–9. https://doi.org/10.3835/plantgenome2014.10.0078

Kayondo, S. I., Pino Del Carpio, D., Lozano, R., Ozimati, A., Wolfe, M., Baguma, Y., … Jannink, J. L. (2018). Genome-wide association mapping and genomic prediction for CBSD resistance in Manihot esculenta. *Scientific Reports*, *8*, 1549. https://doi.org/10.1038/s41598-018-19696-1

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923

Law, M., Childs, K. L., Campbell, M. S., Stein, J. C., Olson, A. J., Holt, C., … Yandell, M. (2015). Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. *Plant Physiology*, *167*, 25–39. https://doi.org/10.1104/pp.114.245027

Liu, E. Y., Li, M., Wang, W., & Li, Y. (2013). MaCH-Admix: Genotype imputation for admixed populations. *Genetic Epidemiology*, *37*, 25–37. https://doi.org/10.1002/gepi.21690

Marimont, R. B., & Shapiro, M. B. (1979). Nearest neighbour searches and the curse of dimensionality. *IMA Journal of Applied Mathematics*, *24*, 59–70. https://doi.org/10.1093/imamat/24.1.59

Mulder, H. A., Calus, M. P. L., Druet, T., & Schrooten, C. (2012). Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *Journal of Dairy Science*, *95*, 876–889. https://doi.org/10.3168/jds.2011-4490

Nazzicari, N., Biscarini, F., Cozzi, P., Brummer, E. C., & Annicchiarico, P. (2016). Marker imputation efficiency for genotyping-by-sequencing data in rice (*Oryza sativa*) and alfalfa (*Medicago sativa*). *Molecular Breeding*, *36*, 69. https://doi.org/10.1007/s11032-016-0490-y

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, *5*, 92–102. https://doi.org/10.3835/plantgenome2012.05.0005

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roberts, A., McMillan, L., Wang, W., Parker, J., Rusyn, I., & Threadgill, D. (2007). Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows *Bioinformatics*, *23*, i401–i407. https://doi.org/10.1093/bioinformatics/btm220

Saghai-Maroof, M. A., Soliman, K. M., Jorgensen, R. A., & Allard, R. W. (1984). Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proceedings of the National Academy of Sciences*, *81*, 8014–8018

Schwender, H., & Fritsch, A. (2015). Scrime: Analysis of high-dimensional categorical data such as SNP data. Retrieved from https://cran.r-project.org/web/packages/scrime/scrime.pdf

Swarts, K., Li, H., Romero Navarro, J. A., An, D., Romay, M. C., Hearne, 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*, 1–12. https://doi.org/10.3835/plantgenome2014.05.0023

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*, 520–525.

SUPPORTING INFORMATION

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

How to cite this article: de Oliveira AA, Guimarães LJM, Guimarães CT, et al. Single nucleotide polymorphism calling and imputation strategies for cost-effective genotyping in a tropical maize breeding program. *Crop Science*. 2020;60:3066–3082.
https://doi.org/10.1002/csc2.20255