Optimizing imputation of marker data from genotyping-by-sequencing (GBS) for genomic selection in non-model species: Rubber tree (Hevea brasiliensis) as a case study

Norman Munyengwa, Hermine Ngalle Bille, André Clément-Demange, Livia Moura de Souza, Pierre Mournet, Aurélien Masson, Mouman Soumahoro, Daouda Kouassi, David Cros

Optimizing imputation of marker data from genotyping-by-sequencing (GBS) for genomic selection in non-model species: rubber tree (*Hevea brasiliensis*) as a case study.

Authors: Norman Munyengwa [1,5], Vincent Le Guen [2,3*], Hermine Ngalle Bille [5], Livia M. Souza [8], André Clément-Demange [2,3], Pierre Mournet [2,3], Aurélien Masson [6], Mouman Soumahoro [7], Daouda Kouassi [6], David Cros [2,3,4]

(1) Crop Science Department, Faculty of Agriculture, University of Zimbabwe, P. O. Box MP167, Mt Pleasant, Harare, Zimbabwe

(2) CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398, Montpellier, France

(3) University of Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

(4) CETIC (African Center of Excellence in Information and Communication Technologies), University of Yaoundé 1, Yaoundé, Cameroon

(5) Department of Plant Biology, Faculty of Science, University of Yaounde 1, Yaounde, Cameroon

(6) SOCFIN/SOGB (Société des caoutchoucs de Grand Bereby), Abidjan, Côte d'Ivoire

(7) SAPH (Société Africaine de Plantations d'Hévéas), Abidjan, Côte d'Ivoire

(8) Molecular Biology and Genetic Engineering Center (CBMEG), University of Campinas (UNICAMP), Campinas, Brazil

* corresponding author: vincent.le_guen@cirad.fr

ABSTRACT

Genotyping-by-sequencing (GBS) provides the marker density required for genomic predictions (GP). However, GBS gives a high proportion of missing SNP data which, for species without a chromosome-level genome assembly, must be imputed without knowing the SNP physical positions. Here, we compared GP accuracy with seven map-independent and two map-dependent imputation approaches, and when using all SNPs against the subset of genetically mapped SNPs. We used two rubber tree (*Hevea brasiliensis*) datasets with three traits. The results showed that the best imputation approaches were LinkImputeR, Beagle and FImpute. Using the genetically mapped SNPs increased GP accuracy by 4.3%. Using LinkImputeR on all the markers allowed avoiding genetic mapping, with a slight decrease in GP accuracy. LinkImputeR gave the highest level of correctly imputed genotypes and its performances were further improved by its ability to define a subset of SNPs imputed optimally. These results will contribute to the efficient implementation of genomic selection with GBS. For *Hevea*, GBS is promising for rubber yield improvement, with GP accuracies reaching 0.52.

Keywords Genomic predictions, *Hevea brasiliensis*, genotyping-by-sequencing, clonal breeding, single nucleotide polymorphisms.

1 Introduction

Genomic selection (GS) [1] has emerged as one of the most promising selection strategies to enhance genetic gain in plant and animal breeding, and its advantages over both phenotypic and quantitative trait loci (QTL) - based marker assisted selection (MAS) have been demonstrated [2–4]. Unlike QTL-based MAS, GS utilizes dense genome-wide markers simultaneously without testing the significance of their effects, to predict the genetic values of individuals in the selection population [5]. One key assumption in GS is that every QTL controlling the phenotypes of interest is in linkage disequilibrium (LD) with at least one SNP. GS requires a training population with phenotypes for the traits of interest and genotyped for a genome-wide panel of markers. The phenotypic and marker data from the training population are then used to develop a genomic prediction equation. The selection population (for example, seedlings) is also genotyped but not phenotyped, and the prediction model calculates the genomic estimated breeding values (GEBVs) or genomic estimated genetic values (GEGVs) of the selection population. Selection among a large number of individuals to identify new elite cultivars or superior genotypes to advance to the next selection stages could be done early in the breeding pipeline based solely on GEBVs or GEGVs, thus increasing selection intensity and reducing the generation interval.

The genotyping-by-sequencing (GBS) approach [6] has gained popularity in GS as a technology for high throughput and low-cost genotyping [7]. Prior sequencing of the genome is not mandatory for GBS, which allows genotyping of individuals with complex genomes without prior SNP discovery, thus making the approach most useful for non-model species [8]. However, GBS also comes with a higher proportion of missing data and sequencing errors as compared to SNP arrays [9]. This is problematic for GS, as genomic prediction models cannot handle

incomplete molecular datasets, making imputation of missing genotypes a requirement before GBS data can be used for genomic predictions.

A number of imputation algorithms have been developed to infer missing marker genotypes but the majority of them require ordered markers to perform imputation [10]. For species lacking a reference genome and without the construction of a reliable genetic linkage map, the majority of markers from GBS are unordered. To deal with this situation, imputation approaches that do not use marker positions were developed, but existing comparisons among them and with approaches using marker positions are not comprehensive (i.e. they focus on just a few imputation methods) [10–13]. Another solution for non-model species without a reference genome would be to make a genetic map of SNPs obtained from GBS and use the genetic positions for imputation. A number of software such as JoinMap [14], Lep-MAP3 [15], OneMap [16] and HighMap [17] have been developed for the construction of genetic linkage maps, especially for biparental populations. Using SNP array genotyping in wheat, map-dependent imputation methods led to higher GS accuracy than map-independent methods [18]. However, to the best of our knowledge, this was not investigated in the case of GBS data, and the interest of using genetic positions when physical positions are not available has not been evaluated.

Therefore, in the context of genomic predictions with SNPs obtained by GBS, in particular for non-model species without a reference genome, a comparison of SNP imputation approaches using (i) map-independent algorithms and (ii) map-dependent algorithms and SNP genetic positions is necessary. For this purpose, we compared the effect of nine imputation approaches on GS accuracy using rubber data (*Hevea brasiliensis*). Despite the great economic importance of this species, which is the only economically viable source of natural rubber, covering around 14 million hectares worldwide and producing around 14 Mt of natural rubber

per year [19], so far only two articles on rubber GS were published [20,21] and a reference genome with 18 chromosomes was published very recently [22]. In this study, we used two rubber datasets, i.e. dataset 1 comprising of 304 full-sib clones of cross PB260 × RRIM600 genotyped with 3,420 SNP markers and phenotyped for rubber yield and sucrose content, and dataset 2 consisting of 251 full-sib clones of cross PR255 × PB217 genotyped with 9,088 SNP markers and phenotyped for trunk circumference. In detail, we used seven map-independent imputation methods which comprised of LinkImputeR 1.2.2 [23], expectation-maximization imputation (EMI) [24], k–nearest neighbors imputation (kNNI) [25], random forest regression imputation (RFI) [26], probabilistic principal component analysis imputation (PPCAI) [27], singular value decomposition imputation (SVDI) [25] and mean imputation (MNI). In addition, two map-dependent imputation methods were evaluated, Beagle 5.1 [28] and FImpute 2.2 [29], using genetic linkage maps constructed with JoinMap 5.0 [14]. We also studied the effect of marker density on GS accuracy and determined the minimum marker density required to maximize GS accuracy. To better understand the results, the percentage of genotypes correctly imputed by each method was computed.

2 Materials and methods

2.1 Plant material and field phenotyping

Dataset 1 comprised of a panel of 304 F1 rubber clones that originated from a cross between clones PB260 and RRIM600, both from Malaysia. A detailed description of the two parents is available in Cros et al. [21]. The clones were phenotyped for rubber yield and sucrose content in two different small-scale clonal trials (SSCTs) in the South-Western parts of Côte d'Ivoire. Sucrose content is a trait of interest for rubber breeding as sucrose is a precursor of

rubber synthesis and an indicator of the clone ability to withstand latex production intensification. A total of 179 clones were evaluated at HR46 (Site 1) which is located at Société des Caoutchoucs de Grand-Béréby (SOGB) estate, and the remaining 125 clones were phenotyped at Sapest13 (Site 2) which is found at Société Africaine de Plantations d'Hévéas (SAPH) estate. Two clones were phenotyped at both site 1 and site 2, thus giving a total of 302 different clones. The two common clones were only used to train the GS model and were excluded from the validation sets. Site 1 trial was planted in July 2012 whilst site 2 trial was planted in July 2013. The trials followed an almost complete block design across the two sites, with individual trees randomized within each block. Individual ramets were planted at a spacing of 2.5 × 2.5 meters to give a plant density of 1,600 per hectare. Data collection on sucrose content was recorded for each ramet by sampling seven latex drops from the tapping panel and measuring the optical absorbance (λ = 627 nm) with a spectrometer after an anthrone reaction [30]. On average the sucrose content was measured twice for each ramet.

Dataset 2 consisted of 251 F1 clonal hybrids from a cross between PR255 and PB217. As described in Souza et al. [20], the field trial was laid out in a randomized block design over four replications and with each plot consisting of four grafted trees from the same individual. The individuals were phenotyped for trunk circumference in a single site trial in Itiquira (Mato Grosso, Brazil) from March 2006 to March 2007. Trunk circumference is a major trait for rubber breeding as growth speed determines the earliness of tapping initiation and the onset of rubber production, at around 6 years old. Trunk circumference data were collected before tapping initiation on 6.5 years old trees grown under low-water conditions by measuring stem diameter at 50 cm from the ground using a tape measure.

2.2 *Estimation of clone genetic values*

The raw phenotypic data collected on the ramets were used to estimate the clone genetic values (i.e. phenotypes adjusted for effects related to the trial experimental design), which were used to train and to validate the GS models. The clone genetic values were obtained using univariate linear mixed models and best linear unbiased predictor (BLUP) analysis implemented with ASReml-R version 3.0 statistical package [31].

In dataset 1, the clone genetic values for rubber yield and the clonal mean heritability (H^2) were taken from Cros et al. [21]. To get clone genetic values using raw sucrose data from the two sites, the model was adjusted for block effects. The clonal mean heritability (H^2) for sucrose content in site 1 and site 2 was 0.66 and 0.85, respectively. The H^2 was calculated as per equation below:

$$H^2 = \sigma_G^2 / \left(\sigma_G^2 + \frac{\sigma_E^2}{h_r}\right)$$

Where σ_G^2 is the genetic variance of clones, σ_E^2 is the residual error variance and h_r is the trial harmonic mean number of ramets per clone, with σ_E^2 and σ_G^2 obtained from the linear mixed model [21].

In dataset2, to get clone genetic values for stem circumference data, the model was adjusted for block, year and age effect. The clonal mean heritability (H^2) for trunk circumference was 0.9 and was calculated using the same approach used for sucrose content. For a more detailed description of this dataset, refer to Souza et al. [20].

2.3 Molecular data

2.3.1 Genotyping-by-sequencing

For dataset 1, high-quality genomic DNA was extracted from young and healthy leaflets on Macherey-Nagel NucleoMag magnetic beads, using the Beckman robot [32]. After DNA extraction, the DNA was digested by two enzymes (*Pst*1 and *Mse*1), in league with the barcode and adapters, and multiplexed to 96 individuals per library. The libraries were sent to Genewiz (USA) for next-generation sequencing. In brief, the GBS protocol followed the following steps: normalization of genomic DNA, digestion with Pst1 and Mse1, ligation of barcoded adapter sequences, direct pooling of PCR products (pooling of 96 genotypes to get one GBS library), DNA purification on column, PCR amplification with adapter specific primers, double purification of DNA and sequencing on Hiseq 4000 platform [6,32,33], generating two opposite reads of 150 nucleotides for each analyzed fragment. The SNPs were called from the sequence reads using VcfHunter, a pipeline available at https://github.com/SouthGreenPlatform/VcfHunter/ [34]. The reference genome used for SNP calling was the one published by Tang et al. [35], which was the best sequence available at the time of the study. It contains 7,453 scaffolds with a median length of 1.28 Mbp.

For dataset 2, genomic DNA was extracted using the approaches of Conson et al. [36] and Souza et al. [37], and GBS library preparation and sequencing were done according to Elshire et al. [6]. Genomic DNA samples were digested by *EcoT22I* to reduce genome complexity and SNP calling was performed using TASSEL 5 GBS version 2 pipeline [38] and the reference genome of Tang et al. [35]. Refer to Souza et al. [20] for a more detailed description of genomic DNA sequencing and SNP calling on dataset 2.

2.3.2 Marker filtering and quality checking

For dataset 1, the following criteria were used to filter SNPs: indels and SNPs that were not biallelic were discarded using VCFtools [39]. All SNPs with a MAF (minor allele frequency) of less than 15% were also removed, as this was not compatible with the possible segregation patterns expected with biallelic markers in a biparental cross. In addition, SNP data points with a read depth of less than eight were set as missing and SNPs with more than 50% missing data were removed. The remaining markers were thinned to keep only one SNP per window of 500 bp using VCFtools. All individuals with more than 50% missing data were removed and all SNPs with a mean read depth greater than 400 were removed, as it was assumed that they were found in duplicated regions of the genome. All SNPs with a heterozygosity percentage greater than 80% and less than 20% were discarded as this was not compatible with the possible values expected with biallelic markers in a biparental cross. Furthermore, all SNPs with a segregation pattern that significantly differed from the segregation expected from the parental genotypes were discarded. This was assessed with the Monte Carlo exact multinomial test (p-value < 0.05), using the function *multinomial.test* in the EMT R package [40]. This resulted in a final marker dataset of 3,420 SNPs, with the percentage of missing data per SNP averaging at 11% and ranging from 0 to 50% (see Figure 1) and percentage of missing data per individual ranging from 0 to 30%. The depth per SNP in dataset 1 was on average 58.8 and ranged from 8 to 401.3.

For dataset 2, SNPs that were not biallelic were removed. Only SNPs with a MAF greater than or equal to 5% were kept. All SNP data points with a read depth of less than 9 were set to missing and only SNPs with less than 60% missing data were kept. If a genotype was present in less than 10% of the non-missing data of a given SNP, it was set as missing (i.e. not compatible with the possible values expected with biallelic markers in a biparental cross). All SNPs which showed inconsistencies between parents and offsprings, i.e. with offsprings showing genotypes

that were not possible given the parental genotypes, were discarded. All SNPs with unknown parental genotypes were removed. This resulted in an SNP dataset with 9,088 markers. The 9,088 SNPs had an average percentage of missing data of 15.2% which ranged from 0 to 60%. The depth per SNP was on average 27.7 and ranged from 7.7 to 162.6.

2.4 Construction of genetic linkage maps

For dataset 1, the 302 F1 individuals of the cross PB260 × RRIM600 and 3,750 markers (3,420 SNPs and 330 SSRs) were used to construct the genetic linkage map. For dataset 2, 253 F1 individuals from the PR255 × PB217 cross and 9,253 markers (9,088 SNPs and 165 SSRs used in Rosa et al. [41]) were used to construct the genetic map. The two genetic linkage maps for the two populations were constructed using the JoinMap 5.0 software [14] using parameters set for cross-pollinated population types and the regression mapping procedure. Assignment to linkage groups was done using a high logarithm of the odds (LOD) threshold value of 5.0. We used linkages with a recombination rate of < 0.4, a map LOD value of 0.05, and a goodness-of-fit jump threshold set at five for inclusion into the linkage map and for calculating the linear order of markers within a linkage group. The Kosambi mapping function [42] was used to estimate map distance and to convert the recombination fractions between markers to map distances in centiMorgans (cM). We integrated SSRs into the genetic maps in order to identify linkage groups according to those already described in previous studies such as Lespinasse et al. [43] and Pootakham et al. [44]. The SSR data were not used in the rest of the study, i.e. imputation and genomic predictions, as the practical GS application is expected to rely only on SNPs, for cost efficiency and practical reasons. With dataset 2, for computational reasons, a first map was made using maximum likelihood. This allowed subsetting the SNPs by removing the ones that mapped very closely, and the final map was then made on this subset of SNPs by

regression mapping. The final linkage maps for the two datasets were plotted using the R package LinkageMapView [45].

2.5 Marker imputation methods

Imputation algorithms Beagle, LinkImputeR, RFI, kNNI, EMI, PPCAI, SVDI, FImpute and MNI were used to impute the sporadic missing genotypes. Beagle and FImpute were applied only to the subset of SNPs located on the genetic map (1,769 SNPs on dataset 1 and 3,111 SNPs on dataset 2, see results), while the other methods were applied on both markers located on the genetic map and all the markers (3,420 SNPs on dataset 1, and 9,088 SNPs on dataset 2). In the rest of the article, the *M* matrix refers to the matrix of genotypes, with *m* rows corresponding to individuals, *n* columns of SNP markers and data points presented in (0, 1, 2, NA) format, representing the three possible genotypes coded in number of copies of the reference allele and the missing value (NA), respectively.

2.5.1 Beagle 5.1

We chose Beagle version 5.1 [28] as one of the two map-dependent imputation methods since it is a widely used imputation approach. In addition, Beagle has been used in *Hevea* GS with SSRs [21].

The Beagle 5.1 algorithm is based on the Li and Stephens hidden Markov model (HMM). The algorithm makes use of the physical position of markers in order to perform imputation. Initially, tightly linked markers are collapsed to form a single aggregate marker. The map position of an aggregate marker *l* is the average map position of the first and last markers. Only target markers are initially included in the HMM and after computing the state probabilities at these markers, the algorithm proceeds to estimate HMM state probabilities at each imputed marker by using

interpolation on genetic distance. The algorithm finishes by using a long sliding window to limit the amount of data stored in memory when imputing whole chromosomes, and markers that overlap between two adjacent sliding windows are imputed twice (once in each window). Imputation accuracy increases with distance from the window boundary, and therefore, imputed values from a sliding window in which the position is in close proximity to the window boundary are discarded.

Beagle 5.1 takes SNP physical positions expressed in base pairs and assumes that 1 cM corresponds to 1 Mbp. Therefore, in order to make this software use the positions in cM on the genetic map, we converted them into base pairs, multiplying them by 1,000,000. We used an *Ne* value of 100,000 and, for the other parameters, the default settings.

2.5.2 LinkImputeR 1.2.2

LinkImputeR version 1.2.2 [23] was chosen among the map-independent imputation methods because it was specially designed for the imputation of unordered biallelic markers. LinkImputeR performs genotype calling and imputation by making use of both read count and available genome sequence information. The algorithm starts by using a likelihood calculation to infer genotypes using read count information. The genotype likelihood (L_g) of seeing the observed read counts if that is the true genotype is calculated for each genotype, g$\in$ {0,1,2}, as follows:

$$L_0 = f(r_R, r_R + r_A, 1 - e)$$

$$L_1 = f(r_R; r_R + r_A, 0.5)$$

$$L_2 = f(r_A; r_R + r_A, 1 - e)$$

Where r_R refers to the number of reference reads, r_A the number of alternative reads, e the error rate and $f(k, tp)$ is the probability mass function of the binomial distribution. The probability of each genotype, p_g^t, is calculated from the genotype likelihoods as follows:

$$p_g^t = \frac{L_g}{L_0 + L_1 + L_2}$$

LinkImputeR then uses a modified LD-kNNI [46] algorithm which produces the most likely genotype (imputed genotype) and a probability for each genotype (imputed probabilities) to impute marker data points that fall below the chosen read depth threshold (i.e. missing marker data). To impute a missing genotype at SNP *a* in sample *b*, the modified LD-kNNI algorithm first considers the *l* SNPs most in LD with SNP *a* to calculate the distance from sample *b* to every other sample for SNP *a*. LD-kNNI then picks the *k* nearest neighbors to *b* that have an inferred genotype at SNP *a* before calculating the score of each of the possible genotypes. The score of each possible genotype is then used to calculate the imputation probability. In order to obtain the highest accuracy, LinkImputeR optimizes the *k* and *l* values.

LinkImputeR allows the user to run the algorithm in accuracy mode to test different read depth thresholds and additional data quality filters such as MAF, percentage of missing data per SNP and per sample, thus allowing for selection of the best parameters for imputation. After testing different MAF and read depth levels, for dataset 1 we used a MAF value of 0.17 and read depth of 10. In addition to the MAF and read depth thresholds, we used a *max depth* = 500, *min depth* = 40 and *numbermasked* = 10,000. For dataset 2, a MAF value of 0.26 and a read depth of 40 was used, with *max depth* = 500, *min depth* = 40 and *numbermasked* = 25,000.

2.5.3 Random forest imputation (RFI)

We included RFI among the map-independent imputation methods as it gave the highest prediction accuracy in a GS study with GBS data in perennial ryegrass [47]. RFI was also the best imputation approach in winter wheat with GBS data where it was compared to EMI [48]. In addition, RFI gave the highest imputation accuracy in a study comparing four map-independent imputation methods with markers from GBS in wheat, maize and barley [10].

The RFI algorithm was implemented in R using the package MissForest [26] and the function *missForest*, using 15 iterations and 300 regression trees. The RFI procedure as described by Rutkoski et al. [10] was implemented as follows:

1. For the marker matrix *M*, SNP markers were first sorted in ascending order according to the percentage of missing data and MNI was used to impute missing values.
2. At each marker *j* with missing values, non-missing values (*Y*) were used to grow 300 random forest regression trees ($\boldsymbol{\Theta}1$, …, 300). Each of the 300 RF regression trees were grown using a bootstrap sample of *Y*, and a random sample of $\sqrt{n-1}$ marker predictors were used, with *n*-1 the number of markers excluding marker *j*. Each of the trees ($\boldsymbol{\Theta}$) contained terminal node values and instructions (split variables at each node and the value of the split variable that is used for partitioning) for recursive partitioning of observations into the terminal nodes.
3. The missing values at each marker *j* were imputed according to:

$$\hat{Y} = \frac{1}{300}\sum_{1}^{300} h(x, \Theta)$$

where *x* is an input vector.

4. Marker *j* was then updated in the marker matrix *M* using the $\hat{Y}$ values as the estimate of missing values.

5. Steps two to four were repeated for each successive marker until all the markers were imputed.
6. Using the imputed matrix, steps 2 to 5 were repeated until convergence occurred or for a maximum of 15 iterations. The convergence was declared as soon as the difference between the previous and newly imputed marker matrix (ΔN) went up for the first time, with ΔN computed as follows:

$$\Delta N = \frac{\sum_{j \epsilon n}(M_1 - M_0)^2}{\sum_{j \epsilon n}(M_1)^2}$$

where M_1 is the newly imputed marker matrix and M_0 is the previously imputed marker matrix. When the convergence criterion was met, M_0 was used as the final estimate of M.

2.5.4 K-Nearest Neighbour Imputation (kNNI)

We included kNNI in this study since it was used to impute GBS data for a GS study in rubber tree [20]. In addition, it gave the highest imputation accuracy with GBS data from alfalfa [49].

For the kNNI algorithm [25], missing marker genotypes were imputed by replacing them with a weighted average of the data points at the k nearest neighbors. Imputation with kNNI was done in three stages. The algorithm starts by replacing missing markers using MNI and then computes the Euclidean distance between all possible pairs of SNP marker vectors. Each marker genotype is included in the marker matrix both in its original state and in the flipped form to avoid markers in negative LD to be considered distant. Secondly, for each marker j, SNP markers were ranked based on the Euclidean distance to marker j. Lastly, for each individual i, an estimate of marker data point x_{ij} was done using the weighted average of the k markers that

were closest to the non-missing values at the i^{th} individual. The weight of each SNP marker was computed as $1/d^2$, where *d* refers to the Euclidean distance between the marker to be weighted and marker *j*. Imputation with kNNI was performed using the function *raw.data* in the R package snpReady [50]. We used a *sweep.sample* threshold of 0.6, a call rate of 0.4 and a MAF threshold of 0.15 to avoid losing any data.

2.5.5 Expectation Maximization Imputation (EMI)

We included EMI in this study since it was designed for use with GBS markers and gave the highest GS accuracy for fusarium head blight incidence and kernel quality index among five map-independent imputation methods in wheat [51].

The multivariate normal - expectation maximization (EM) algorithm [24] is a general approach for computing maximum likelihood estimates of unknown parameters when there is missing data. Imputation of missing data points using the EM algorithm involves two steps which are the expectation step and the maximization step. The EM imputation algorithm is based on the multivariate normal distribution and it imputes missing markers based on realized relationship averaged over all the markers. We implemented the EMI using the R package rrBLUP [52]. A detailed description of the EM algorithm is found in Poland et al. [53].

2.5.6 Probabilistic PCA imputation (PPCAI)

PPCAI was chosen for this study because it gave the best imputation accuracy on a comparison involving three map-independent imputation methods with GBS data from maize and wheat [54].

PPCAI [27] is a PCA based imputation approach. Missing data-points were first set to marker averages, and singular value decomposition (SVD) was then used to construct orthogonal principal components. The algorithm finishes by using the principal components, which correspond to the largest eigenvalues, to infer the missing genotypes in the SNP matrix. PPCAI was implemented using the R package pcaMethods [27]. We used the function *KEstimate* to identify the optimal number of principal components, which was chosen as the one that gave the highest prediction accuracy.

*2.5.7 **Singular Value Decomposition Imputation (SVDI)***

SVDI was chosen amongst map-independent imputation methods as it gave the best prediction accuracy for fusarium damaged kernels in wheat with GBS data [51]. For the SVDI [25] algorithm, singular value decomposition of marker matrix *M* was used to obtain the *k* most significant eigengenes (different numbers of principal components), i.e. with the highest eigengene value. Missing value *j* was then estimated in SNP *i* by first regressing SNP *i* against the *k* eigengenes. The regression coefficients were then used to reconstruct missing value *j* using a linear combination of *k* eigengenes. SVDI was implemented using the R package pcaMethods [27].

2.5.8 FImpute 2.2

We included FImpute version 2.2 [29] in the comparison of imputation methods as it outperformed Beagle imputation with SNP array data in beef cattle [55]. In addition, FImpute gave a similar accuracy as Beagle 4.0 on GBS data in dairy cows [56].

FImpute makes use of pedigree information to impute missing genotypes and if pedigree information is not available, the algorithm uses population information to construct haplotypes

using an overlapping sliding window approach. Like Beagle, FImpute requires SNP physical positions expressed in base pairs and assumes that 1 cM corresponds to 1 Mbp. Consequently, we multiplied the genetic positions by 1,000,000 to convert them into physical positions. We used the default settings of FImpute (*shrink factor* = 0.15 and *overlap* = 0.65 of sliding windows) for this study and for both dataset 1 and dataset 2. Imputation with FImpute was performed on the set of SNPs that were mapped. We implemented FImpute with and without pedigree information. We used FImpute with pedigree information as Cros et al. [57] showed that including pedigree information improved GS accuracy in oil palm GS with GBS data.

2.5.9 Mean imputation (MNI)

For MNI, each missing marker data-point at a given marker was replaced with the mean of the non-missing values at that marker.

2.6 Impact of marker density on prediction accuracy

The effect of marker density on GS accuracy was assessed using dataset 1 by making genomic predictions using marker subsets of varying sizes from the mapped and non-mapped markers. From the non-mapped marker data set, marker subsets of 25, 250, 500, 1000, 1,769 and 3,420 SNPs were created, and from the mapped marker dataset, marker subsets of 25, 250, 500, 1,000 and 1,769 SNPs were created. Marker subsets were made by random sampling, with 45 replicates by level of SNP number.

2.7 Model for genomic predictions

A purely additive GS model and the random regression best linear unbiased prediction (rr-BLUP) method [1] were used to obtain the genomic estimated genetic values (GEGVs) of the validation rubber clones. The additive model was chosen as Cros et al. [21] did not find any difference in accuracy when using this approach and other approaches modeling non-additive effects in dataset 1. Marker effects were estimated using the following linear mixed model:

$$y = X\beta + Zm + \varepsilon$$

Where *y* is the vector of clone genetic values (i.e. adjusted phenotypic values), β is the vector of fixed effects (the mean phenotype) with incidence matrix *X*, *m* is the vector of the random marker effects with incidence matrix *Z*, i.e. the matrix of genotypes, and ε is the vector of residual effects. The *Z* matrix was the imputed version of the *M* matrix and contained the number of reference alleles that were observed in each marker coded in -1, 0 and 1.

The structure of the means and variances for the rr-BLUP model was as follows: $m \sim \mathrm{N}(0,\ G)$, $G = I\sigma_m^2,\ E(y) = X\beta$, $\varepsilon \sim \mathrm{N}(0,\ R)$, $R = I\sigma_\varepsilon^2$, and $Var\ (y) = V = ZGZ' + R$ where σ_m^2 is the variance of the SNP effects (common to each SNP) and σ_ε^2 is the residual error variance [58]. The rr-BLUP mixed model for prediction of *m* is equivalent to:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\frac{\sigma_\varepsilon^2}{\sigma_m^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

An estimate of the effect of each SNP marker was obtained by solving the mixed model equations presented above. The predicted GEGV of the individual rubber clone *j* was given by $\hat{g}_j = \sum_i Z_{ji}\hat{m}_i$. The variances were calculated by restricted maximum likelihood (REML). The rr-BLUP analysis was performed using the *mixed.solve* function in the R package rrBLUP [52].

2.8 Validation approaches for genomic predictions

2.8.1 Across site validation within dataset 1

Across-site genomic predictions were performed with dataset 1, leading to two different validation scenarios (Site 1 towards Site 2 and Site 2 towards Site 1). In the first scenario, 179 individual clones from Site 1 were used to train the GS model to predict the GEGVs of the 123 clones in Site 2, that were used as the validation population. In the second scenario, the 125 individual clones in Site 2 were used to train the GS model and the 177 clones in Site 1 were used for the validation. The two common clones in the two sites were only used to train the GS model and excluded from the validation sets.

2.8.2 Cross-validation within dataset 2

Five-fold cross-validation was performed with dataset 2. The population was randomly divided into five partitions (folds), and four of these partitions were used as the training population to predict the genetic values of the remaining partition. All five partitions were used in turn for validation.

2.8.3 Prediction accuracy

Prediction accuracy was obtained for each validation set as the Pearson correlation between the GEGV ($\hat{g}$) and the genetic value (y) of the clones composing the set. The GS accuracy was calculated by dividing the prediction accuracy by the square root of clonal mean heritability (H^2). For dataset 2, accuracies of the five cross-validations were averaged to give aggregated accuracies.

Pairwise comparisons of prediction accuracies among imputation methods were made for each trait using the Hotelling-Williams t-test [59]. This test compares two dependent Pearson

correlation coefficients, i.e. with one variable in common (which, in our case, is the genetic value of clones used for validation). The test was implemented using the function *r.test* in the R package psych [60].

2.9 Percentage of correctly imputed genotypes

Among the genotypes (i.e. datapoints) with high read depth, some genotypes were set as missing (validation genotypes), imputed with the different methods and used to compute the percentage of correct imputations. This was done in dataset 1 and dataset 2 considering the mapped SNPs, in order to compare the nine imputation methods. There were 17,503 and 6,291 validation genotypes in dataset 1 and in dataset 2, respectively, defined as genotypes with read depth ≥250 and ≤550 in dataset 1 and with read depth ≥105 and ≤125 in dataset 2. In each dataset, three replicates were defined by setting as missing one third of the validation genotypes, with the validation genotypes set as missing in a given replicate chosen randomly. For LinkImputeR, as this software makes use of the read depth information, the actual read depth of the validation genotypes was replaced by read depth randomly sampled among the read depths observed in the genotypes that were originally missing. The actual allelic read depth of each validation genotype was then scaled to match with the read depth obtained in the previous step.

The effect of the imputation method on the percentage of correctly imputed genotypes was investigated by analyses of variance (ANOVA), performed separately in each dataset, with imputation method and replicate as factors. The mean levels of factors were compared using Tukey's honest significant difference test.

3 Results

3.1 Genetic linkage maps

From the 3,420 SNPs and 330 SSRs used to construct the genetic linkage map of PB260 × RRIM600 individuals, 1,769 SNPs and 308 SSRs were positioned. We obtained a genetic linkage map with 2,077 markers that were spread over 19 linkage groups (LG), which almost corresponds to the haploid chromosome number of the hevea tree [43], except that LG6 was split into two LGs, LG 6a and LG 6b. The linkage map encompassed 2600.9 cM, with the size of linkage groups ranging from 111 cM (LG 12) to 181.8 cM (LG 10). The number of markers mapped in each linkage group ranged from 71 (59 SNPs and 12 SSRs) in LG 04 to 171 (139 SNPs and 32 SSRs) in LG10, and the number of SNPs in each LG ranged from 59 (LG 04) to 143 in LG 05 (Table 1). The average inter-marker (SSRs and SNPs combined) distance across the linkage groups was 1.33 cM and the average distance between SNP markers was 1.47 cM. Large gaps of more than 10 cM were observed in nine linkage groups.

From the 9,088 SNP markers used to construct the genetic map of PR255 × PB217 individuals, we obtained a genetic map which comprised of 3,111 SNPs that were spread across 19 linkage groups (Table 2). The genetic map spanned a cumulative length of 2,215.1 cM and the size of linkage groups ranged from 61.6 cM (LG10 + LG17) to 171.2 cM (LG10). The number of unique SNPs located on each linkage group ranged from 55 (LG18) to 216 (LG07), with an average of 194 SNP markers per linkage group. The average inter-marker distance in the map from PR255 × PB217 population was 0.71 cM which is shorter than in the map of PB260 × RRIM600, where the average SNP inter-marker distance was 1.47 cM. We observed only three gaps larger than 10 cM and these were located at the terminal portions of LG01, LG07 and LG08.

The SNP genetic maps are presented in Figure 2 and Figure 3. The segregation patterns of SNP markers located on the two genetic maps are shown in Supplementary Table S 1.

3.2 Effect of imputation method on percentage of correctly imputed genotypes

Large differences among imputation methods were found in terms of percentage of correctly imputed genotypes (Figure 4). LinkImputeR gave the best results, with 92% and 97% of correct imputations in datasets 1 and 2, respectively. It was significantly better than all the other methods, and was followed in terms of imputation accuracy by FImpute using pedigree information (88% on average over the two datasets) and Beagle (86%). MNI was the worst method in both datasets, with a mean percentage of correctly imputed genotypes of 42%.

3.3 Effect of imputation method on GS accuracy

Mapped SNPs: When considering the subsets of mapped SNPs (i.e. 1,769 SNPs on dataset 1 and 3,111 on dataset 2), LinkImputeR, Beagle and FImpute without pedigree information were the best imputation methods in terms of GS accuracy, on average over the three traits, although there were variations according to trait (Table 3). Thus, for genomic predictions made for site 1 of dataset 1, considering rubber yield, the GS accuracy varied from 0.53 for LinkImputeR, Beagle and SVDI to 0.48 for MNI, corresponding to a 10.8% increase between the worst and the best methods (Table 3). For rubber yield in site 2, LinkImputeR was the best imputation method with a GS accuracy of 0.5, closely followed by Beagle and FImpute without pedigree information (0.49). The method giving the smallest GS accuracy was again MNI (0.45), corresponding to an increase between the worst and the best methods similar to what was found in site 1 (+11.1%). For sucrose content, kNNI (0.16) was the best imputation method for site 1,

closely followed by Beagle, LinkImputeR and FImpute without pedigree information (0.15 for the three of them). The lowest GS accuracy was recorded with MNI and PPCAI, both of them with a GS accuracy of 0.13. For site 2, FImpute without pedigree information gave the highest GS accuracy (0.25) followed by Beagle (0.24), FImpute with pedigree information (0.23), LinkImputeR and SVDI (0.22). The worst method was MNI (0.17), making GS accuracies obtained with FImpute without pedigree information 47.1% higher. When we performed cross-validations for GS prediction of circumference within dataset 2, LinkImputeR was the imputation method which led to the highest GS accuracy (0.21), closely followed by Beagle, kNNI, FImpute without pedigree information, FImpute with pedigree information and MNI (GS accuracy of 0.2 for the five of them).

Over the three cases (i.e. combinations of sites and traits) investigated here, LinkImputeR, Beagle and FImpute without pedigree information therefore achieved the best results, as they obtained the best average rank and highest mean GS accuracy over traits and sites (0.32, against 0.29 to 0.31 for the other methods). However, although the differences in GS accuracy between the best and worst methods could be high, many differences were not significant (see Supplementary Table S 2 and Supplementary Table S 3).

All SNPs: When considering all the SNPs, with or without known genetic positions (i.e. 3,420 SNPs on dataset 1 and 9,088 on dataset 2), LinkImputeR was on average the best imputation method, although there were again variations according to trait (Table 3). When making GS predictions for rubber production, LinkImputeR gave the highest GS accuracy in the two sites of dataset 1, reaching 0.53 in site 1 (10.4% higher than the worst method, MNI) and 0.5 in site 2 (13.6% higher than the worst method, MNI). For sucrose content, kNNI (0.16) was the best imputation method in site 1 followed by EMI (0.15), PPCAI (0.15), SVDI (0.15),

LinkImputeR (0.14), RFI (0.14) and lastly MNI (0.14). In site 2, PPCAI and SVDI were equally the best imputation methods (0.19) followed by RFI (0.18), EMI (0.18), LinkImputeR (0.17), MNI (0.16) and lastly kNNI (0.15). For circumference in dataset 2, the highest GS accuracy was obtained using LinkImputeR (0.22) followed by kNNI and MNI (GS accuracy of 0.2 for both of them). As with markers from the genetic map, the lowest GS accuracies were obtained after using markers imputed with RFI, EMI, PPCAI and SVDI. Over these cases, LinkImputeR was the most efficient method for GS predictions, as it had the best average rank and mean GS accuracy (0.31, against 0.28 to 0.30 for the other methods). However, many differences were not significant (see Supplementary Table S 4 and Supplementary Table S 5).

Improved GS accuracies were obtained when using only the markers with known position on the genetic maps compared to when using all the markers (Table 3). When considering the map-independent imputation methods, which were used on both SNP sets, there was an increase in GS accuracy, despite the associated reduction in marker density. This occurred in 62.9% of the cases studied, i.e. combinations of traits, site and imputation method, and on average GS accuracy increased by 4.3% (ranging from -13.3% to +29.4%, according to the cases). For LinkImputeR, which was the best method when using the SNPs with known position (although performing as well as Beagle and FImpute without pedigree information), and the best method when using all the SNPs, using the mapped SNPs increased GS accuracy on average by 6.4%.

3.4 Impact of marker density on GS accuracy

Figure 5 and Figure 6 summarize the effect of marker density on GS accuracy in dataset 1 for the two traits under study (rubber yield and sucrose content). When marker density became higher, there was for the two traits a strong increase in GS accuracy, before a plateau was

reached at around 1,000 markers with markers from the genetic map and around 1,600 markers with unordered markers. Only marginal increases in GS accuracy were observed after the plateau was reached. Rubber yield showed a more pronounced increase in GS accuracy with marker density as compared to sucrose content.

Also, from the results of the map-independent imputation methods, we can see in Figure 5 and Figure 6 that using only the markers with known position on the genetic maps increased GS accuracies compared to when using all the markers. This was shown by the lower number of markers (1,000) needed to reach the plateau when markers from the genetic map were used as compared to the 1,600 markers needed to reach the plateau with all the markers.

4 Discussion

The results obtained here highlight the importance of the choice of the imputation method with GBS data. LinkImputeR gave the highest percentage of correctly imputed genotypes (>90%), followed by FImpute with pedigree information and Beagle. Considering three traits of interest for rubber breeding and SNP data obtained by GBS, we showed that for genomic predictions, using the subset of SNPs with known genetic position and imputing the sporadic missing SNP data with LinkImputeR, Beagle or FImpute without pedigree information was a good strategy, as it maximized GS accuracy compared to using all the SNPs. Alternatively, using LinkImputeR on all the available markers allowed avoiding the prior construction of a genetic map, with only a slight decrease in GS accuracy. As expected, the performances of the imputation methods in terms of GS accuracy correlated with their performances measured in percentage of correctly imputed genotypes. However, the differences among methods in GS accuracy were of much smaller magnitude than the differences in percentage of correctly

imputed genotypes. These results are particularly interesting for all species for which a reference genome of good quality (i.e. with a chromosome-level assembly) is not available yet.

4.1 Effect of imputation method

Here we found that, for the three traits studied, map-independent imputation methods resulted in higher GS accuracies when imputation was done on markers with known positions on the genetic map as compared to imputation on all the markers. This suggests that the linkage mapping process leads to filtering SNPs on relevant criteria for genomic predictions, such as the consistency between parental genotypes and segregation patterns in progenies (which is related to a low rate of genotyping errors and reduced distortion of segregation) and the elimination of redundant markers (i.e. markers located at similar positions in the genetic map, and that likely provide similar genomic information due to LD). The fact that GS accuracy can be increased by using a subset of markers has already been reported. This was, for example, the case in rubber, by filtering SSR markers on observed heterozygosity [21], in oil palm, by filtering SNPs from GBS on the percentage of missing data [57], and in pigs, by filtering whole-genome sequence variants on LD [61]. However, to our knowledge, this is the first article reporting the fact that using a subset of SNPs positioned on a genetic map could be efficient in the context of genomic predictions, i.e. could increase GS accuracy or at least allow reaching the same GS accuracy with a reduced marker density.

The study also found that map-independent imputation methods, if applied on markers from the genetic map, can outperform or at least perform similarly with map-dependent imputation methods such as Beagle and FImpute in terms of GS accuracy. This indicates that the relevant filtering of SNPs is at least as important as the choice of the imputation algorithm.

Overall, LinkImputeR was the best imputation method because of its consistent performance over the traits and SNP datasets (i.e. the subsets of SNPs located on the genetic maps and all the markers), and its high percentage of correctly imputed genotypes. LinkImputeR has a unique feature compared to the other imputation approaches presented here, i.e. the ability to define a subset of SNPs imputed optimally: the user specifies a range of values that will be used as filters for different parameters, in particular minimum minor allele frequency, maximum percentage of missing genotype per SNP and sample, and minimum read depth per genotype, and LinkImputeR measures the imputation accuracy for every combination of the different filters. This allows the users to easily identify the optimal quality filters to apply to their SNP dataset to achieve the best imputation. To better understand the performances of LinkImputeR, we computed the percentage of correctly imputed genotypes when no filtering was made by the software, i.e. forcing it to keep all the SNPs. The mean percentage of correctly imputed genotypes over the two datasets remained high (0.86), but was slightly below FImpute with pedigree information (0.88) and identical to Beagle. LinkImputeR therefore performed best as a result of its high imputation accuracy and its ability to define a subset of SNPs imputed optimally. This also suggests that FImpute and Beagle could perform as well as, or even outperform, LinkImputeR, if their use was preceded by a preliminary analysis to identify quality filters to apply to the SNP dataset. This should be further investigated. In addition, a possible reason for the consistent high performance of LinkImputeR, in particular compared to the other map-independent imputation methods, is that it makes use of LD between markers, which is a key information regarding the correspondence between the genotypes at two different SNPs. For future studies, it would be interesting to evaluate the performances of LinkImputeR in

populations with a higher degree of complexity, and therefore a contrasting pattern of LD compared to the biparental populations used here.

The differences among methods in terms of GS accuracy were relatively small, with a coefficient of variation of 3%, on average over combinations of traits, sites and SNP datasets, while the percentage of correctly imputed genotypes varied more widely among methods, with an average coefficient of variation of 22%. This suggests that most of the imputation errors can be considered as random, which makes that, on average, they have a marginal effect on the genomic relationships between individuals, and therefore have a small effect on GS accuracy. However, for genetic approaches where the reliability of individual genotypes is of greater importance than in GS, e.g. in QTL-based studies or legitimacy tests, it will be crucial to carefully choose the imputation method.

The results we obtained here are particularly interesting for species without a reference genome with a chromosome-level assembly. For rubber, the recent publication of a sequence with such an assembly [22] opens the way for further investigation on the optimization of the genomic predictions, and in particular regarding the imputation of the sporadic SNP data.

For FImpute, it was better to use pedigree information in terms of imputation accuracy but better not to use it in terms of GS accuracy. Thus, with both datasets, FImpute imputed significantly better when using pedigree information. By contrast, for GS accuracy, FImpute without pedigree information consistently gave higher GS accuracies for the two traits (yield and sucrose) and across the two sites as compared to FImpute with pedigree information, except for yield in site 1 and trunk circumference where the two imputation methods gave the same GS accuracy. This discrepancy remains unclear, in particular as the legitimacy of the clones in dataset 1 was controlled using SSRs (not shown), and because the SNPs with a segregation

pattern that significantly differed from the segregation expected from the parental genotypes were discarded.

4.2 Effect of marker density on GS accuracy

The effect of marker density on GS accuracy was assessed for rubber yield and sucrose content across the two sites. The two traits showed a strong response to increases in marker density up to 1,000 markers and 1,600 markers for the subset of mapped SNPs and all the SNPs, respectively, showing that these marker densities were high enough to achieve the maximum possible GS accuracies for the families and traits under study. Rubber yield benefited more from an increase in marker density than sucrose content, which suggests that rubber yield is a more complex trait than sucrose content (i.e. with a genetic architecture that follows more closely the infinitesimal model). This makes sense considering that rubber yield depends on sucrose content, as sucrose is a precursor of rubber synthesis, but also on many other parameters.

4.3 Effect of marker type on GS accuracy

The results obtained in this study confirm that GBS is an efficient genotyping method for the application of GS in rubber breeding, as indicated by the results obtained by Souza et al. [20]. Indeed, for rubber yield, the mean between site GS accuracy that we obtained for rubber yield with dataset 1 using markers imputed with LinkImputeR was 0.52, which is, for the cross considered here, sufficient to increase the rate of genetic progress by the inclusion of a stage of genomic selection before the conventional small scale clonal trials in Cros et al. [21]. In addition, this value is similar to the GS accuracy of Cros et al. [21], where a mean between site GS accuracy of 0.53 was obtained with the same dataset genotyped with 332 SSRs (instead of GBS),

and using a larger number of clones (330 against 302 here, due to technical difficulties with the GBS analysis). We can even assume that, if data had been available here for the same number of clones as in Cros et al. [21], higher GS accuracies could have been achieved, as GS accuracy increases with the size of the training population [62]. It shows that SNPs from GBS, despite the fact that they have a higher percentage of missing data and higher error rate than SSRs, can be used for GS predictions. This opens the way to the practical application of GS, as Cros et al. [21] showed that, to increase the rate of genetic gain, GS required to be applied on a large number of selection candidates (>1,000), which would not be cost-effective with SSRs but is with GBS. Further studies remain needed to be able to reach satisfactory GS accuracies for the other traits of interest for rubber breeding, such as trunk circumference and sucrose content.

4.4 Effect of trait on GS accuracy

Our results show that GS accuracy varied strongly depending on trait, and hence, before the practical application of GS, it is useful to evaluate its accuracy for the target trait(s). Although rubber yield and trunk circumference had the same clonal mean heritability, the GS accuracies differed. We hypothesize that could be due to differences in the genetic architecture among the two traits, that would affect the performance of rr-BLUP, as shown for example in Resende et al. [58]. Thus, rubber production could be closer to the infinitesimal model, with a large number of genes with small effects involved, while trunk circumference could involve some genes with large effects, as suggested by the fact that more QTLs are found for this trait than for rubber production [41].

4.5 High-density rubber genetic linkage maps

We obtained, for dataset 1 and 2, respectively, genetic linkage maps that contained 1,769 and 3,111 SNP markers located at non-redundant positions and spread across 19 LGs. The reason for an extra LG could be due to a large interval without a marker, making the SNP alleles on the two chromosome segments appear as statistically unrelated. This could also be because of a recombination hotspot for which it is difficult to prove the genetic proximity between physically close markers. The two genetic maps spanned a total length of 2,600.9 cM and 2,215.1 cM, respectively, which is comparable with the lengths of previously published maps in rubber (2,144 cM in Lespinasse et al. [43], 2,041 cM in Pootakham et al. [44] and 2,441 cM in Le Guen et al. [63]).

The two GBS-based linkage maps, with a marker density of one SNP in every 1.47 cM and 0.71 cM in dataset 1 and 2, respectively, showed a much denser genome coverage as compared to previously published microsatellite-based linkage maps of *Hevea* (one marker in every 3.5 cM in Conson et al. [36], in every 8 cM in Le Guen et al. [63], in every 10 cM in Souza et al. [37]). Their marker density is comparable to what was achieved with GBS in the composite map obtained by Pootakham et al. [44], where SNP density was one SNP in every 0.89 cM.

Acknowledgments

We thank the Institut Francais du Caoutchouc (IFC), and the companies Michelin, SIPH and SOCFIN for financial support and for providing the data used for this research. We also thank the companies SOGB (Société Grand Bereby) and SAPH (Société Africaine de Plantations d'Hévéa) for their logistical assistance in the field experiments in Cote d'Ivoire. The authors are grateful to the Michelin group (Brazil) for allowing access to phenotypic and molecular data

from their breeding trials in Itiquira (Mato Grosso, Brazil). We acknowledge the GENES program of the Intra-Africa Academic Mobility Scheme of the European Union and the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) for financial support.

Author contribution statement

NM carried out data analysis, with the help of DC. The paper was written by NM and DC, with the help of HBN. LS carried out preliminary phenotypic and genomic data analyses on dataset 2. AM, MS and DK helped in designing and performing field experiments and phenotyping. ACD and VLG designed field experiments, supervised phenotypic data collection and, with help of PM, the generation of molecular data on dataset 1.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability statement

The data from dataset 1 can be found at the following link:

https://www.ncbi.nlm.nih.gov/sra/PRJNA645262

Dataset 2 is publicly available and can be found at the following links:

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA540286 and

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA541308.

Funding

This work was partly funded by the Institut Francais du Caoutchouc (IFC) and the companies Michelin, SIPH and SOCFIN in the framework of the IFC project "*Hevea varietal creation in West-Africa*". NM received a master scholarship from the GENES program of the Intra-Africa Academic Mobility Scheme of the European Union. LS received a postdoctoral fellowship from the FAPESP (2012/05473-1 and 2012/50491-8). KeyGene N.V. owns patents and patent applications protecting its Sequence Based Genotyping technologies.

References

[1] T.H. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps., Genetics. 157 (2001) 1819–1829.

[2] B.J. Hayes, N.O.I. Cogan, L.W. Pembleton, M.E. Goddard, J. Wang, G.C. Spangenberg, J.W. Forster, Prospects for genomic selection in forage plant species, Plant Breed. 132 (2013) 133–143. https://doi.org/10.1111/pbr.12037.

[3] E.L. Heffner, M.E. Sorrells, J.-L. Jannink, Genomic Selection for Crop Improvement, CROP Sci. 49 (2009) 13.

[4] K.P. Voss-Fels, M. Cooper, B.J. Hayes, Accelerating crop genetic gains with genomic selection, Theor. Appl. Genet. (2018). https://doi.org/10.1007/s00122-018-3270-8.

[5] X. Wang, Y. Xu, Z. Hu, C. Xu, Genomic selection methods for crop improvement: Current status and prospects, Crop J. 6 (2018) 330–340. https://doi.org/10.1016/j.cj.2018.03.001.

[6] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species, PLOS ONE. 6 (2011) e19379. https://doi.org/10.1371/journal.pone.0019379.

[7] D.P. Wickland, G. Battu, K.A. Hudson, B.W. Diers, M.E. Hudson, A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy, BMC Bioinformatics. 18 (2017) 586. https://doi.org/10.1186/s12859-017-2000-6.

[8] S. Kagale, C. Koh, W.E. Clarke, V. Bollina, I.A.P. Parkin, A.G. Sharpe, Analysis of Genotyping-by-Sequencing (GBS) Data, in: D. Edwards (Ed.), Plant Bioinforma. Methods Protoc., Springer New York, New York, NY, 2016: pp. 269–284. https://doi.org/10.1007/978-1-4939-3167-5_15.

[9] B. Darrier, J. Russell, S.G. Milner, P.E. Hedley, P.D. Shaw, M. Macaulay, L.D. Ramsay, C. Halpin, M. Mascher, D.L. Fleury, P. Langridge, N. Stein, R. Waugh, A Comparison of Mainstream Genotyping Platforms for the Evaluation and Use of Barley Genetic Resources, Front. Plant Sci. 10 (2019). https://doi.org/10.3389/fpls.2019.00544.

[10] J.E. Rutkoski, J. Poland, J.-L. Jannink, M.E. Sorrells, Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy, G3 GenesGenomesGenetics. 3 (2013) 427–439. https://doi.org/10.1534/g3.112.005363.

[11] V. Edriss, Y. Gao, X. Zhang, M.B. Jumbo, D. Makumbi, M.S. Olsen, J. Crossa, K.C. Packard, J.-L. Jannink, Genomic Prediction in a Large African Maize Population, Crop Sci. 57 (2017) 2361. https://doi.org/10.2135/cropsci2016.08.0715.

[12] C. Miao, J. Fang, D. Li, P. Liang, X. Zhang, J. Yang, J.C. Schnable, H. Tang, Genotype-Corrector: improved genotype calls for genetic mapping in F2 and RIL populations, Sci. Rep. 8 (2018) 10088. https://doi.org/10.1038/s41598-018-28294-0.

[13] Y. Wang, G. Lin, C. Li, P. Stothard, Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle, Springer Sci. Rev. 4 (2016) 79–98. https://doi.org/10.1007/s40362-017-0041-x.

[14] P. Stam, Construction of integrated genetic linkage maps by means of a new computer package: Join Map, Plant J. 3 (1993) 739–744. https://doi.org/10.1111/j.1365-313X.1993.00739.x.

[15] P. Rastas, Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data, Bioinformatics. 33 (2017) 3726–3732. https://doi.org/10.1093/bioinformatics/btx494.

[16] G.R.A. Margarido, A.P. Souza, A. a. F. Garcia, OneMap: software for genetic mapping in outcrossing species, Hereditas. 144 (2007) 78–79. https://doi.org/10.1111/j.2007.0018-0661.02000.x.

[17] D. Liu, C. Ma, W. Hong, L. Huang, M. Liu, H. Liu, H. Zeng, D. Deng, H. Xin, J. Song, C. Xu, X. Sun, X. Hou, X. Wang, H. Zheng, Construction and analysis of high-density linkage map using high-throughput sequencing data, PloS One. 9 (2014) e98855. https://doi.org/10.1371/journal.pone.0098855.

[18] S. He, Y. Zhao, M.F. Mette, R. Bothe, E. Ebmeyer, T.F. Sharbel, J.C. Reif, Y. Jiang, Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.), BMC Genomics. 16 (2015). https://doi.org/10.1186/s12864-015-1366-y.
[19] FAOSTAT, Food and agriculture data. Accessed 11 February 2020. http://www.fao.org/faostat/en/#data/QC, (2020).
[20] L.M. Souza, F.R. Francisco, P.S. Gonçalves, E.J. Scaloppi Junior, V. Le Guen, R. Fritsche-Neto, A.P. Souza, Genomic Selection in Rubber Tree Breeding: A Comparison of Models and Methods for Managing G×E Interactions, Front. Plant Sci. 10 (2019). https://doi.org/10.3389/fpls.2019.01353.
[21] D. Cros, L. Mbo-Nkoulou, J.M. Bell, J. Oum, A. Masson, M. Soumahoro, D.M. Tran, Z. Achour, V. Le Guen, A. Clement-Demange, Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production, Ind. Crops Prod. 138 (2019) 111464. https://doi.org/10.1016/j.indcrop.2019.111464.
[22] J. Liu, C. Shi, C.-C. Shi, W. Li, Q.-J. Zhang, Y. Zhang, K. Li, H.-F. Lu, C. Shi, S.-T. Zhu, Z.-Y. Xiao, H. Nan, Y. Yue, X.-G. Zhu, Y. Wu, X.-N. Hong, G.-Y. Fan, Y. Tong, D. Zhang, C.-L. Mao, Y.-L. Liu, S.-J. Hao, W.-Q. Liu, M.-Q. Lv, H.-B. Zhang, Y. Liu, G.-R. Hu-Tang, J.-P. Wang, J.-H. Wang, Y.-H. Sun, S.-B. Ni, W.-B. Chen, X.-C. Zhang, Y.-N. Jiao, E.E. Eichler, G.-H. Li, X. Liu, L.-Z. Gao, The Chromosome-Based Rubber Tree Genome Provides New Insights into Spurge Genome Evolution and Rubber Biosynthesis, Mol. Plant. 13 (2020) 336–350. https://doi.org/10.1016/j.molp.2019.10.017.
[23] D. Money, Z. Migicovsky, K. Gardner, S. Myles, LinkImputeR: user-guided genotype calling and imputation for non-model organisms, BMC Genomics. 18 (2017) 1–12. https://doi.org/10.1186/s12864-017-3873-5.
[24] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data Via the EM Algorithm, J. R. Stat. Soc. Ser. B Methodol. 39 (1977) 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
[25] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics. 17 (2001) 520–525. https://doi.org/10.1093/bioinformatics/17.6.520.
[26] D.J. Stekhoven, P. Bühlmann, MissForest - nonparametric missing value imputation for mixed-type data, Bioinformatics. 28 (2012) 112–118. https://doi.org/10.1093/bioinformatics/btr597.
[27] W. Stacklies, H. Redestig, M. Scholz, D. Walther, J. Selbig, pcaMethods—a bioconductor package providing PCA methods for incomplete data, Bioinformatics. 23 (2007) 1164–1167. https://doi.org/10.1093/bioinformatics/btm069.
[28] B.L. Browning, Y. Zhou, S.R. Browning, A One-Penny Imputed Genome from Next-Generation Reference Panels, Am. J. Hum. Genet. 103 (2018) 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.
[29] M. Sargolzaei, J.P. Chesnais, F.S. Schenkel, A new approach for efficient genotype imputation using information from relatives, BMC Genomics. 15 (2014) 478. https://doi.org/10.1186/1471-2164-15-478.
[30] G. Ashwell, Colorimetric Analysis of sugar in: Method in Enzymology.[Colowick, SP and Koplan, NO eds.].. Vol. III. 73-105, Acad. Press Inc Publ. N. Y. (1957).
[31] D.G. Butler, B.R. Cullis, A.R. Gilmour, B.J. Gogel, Mixed Models for S Language Environments: ASReml-R Reference Manual (Version 3). Queensland Department of Primary Industries and Fisheries, pp., (2009) 398.
[32] F. Cormier, F. Lawac, E. Maledon, M.-C. Gravillon, E. Nudol, P. Mournet, H. Vignes, H. Chaïr, G. Arnau, A reference high-density genetic map of greater yam (Dioscorea alata L.), Theor. Appl. Genet. 132 (2019) 1733–1744. https://doi.org/10.1007/s00122-019-03311-6.
[33] D. Bhatia, R. Wing, K. Singh, Genotyping by sequencing, its implications and benefits, Crop Improv. 40 (2013) 101–111.

[34] O. Garsmeur, G. Droc, R. Antonise, J. Grimwood, B. Potier, K. Aitken, J. Jenkins, G. Martin, C. Charron, C. Hervouet, L. Costet, N. Yahiaoui, A. Healey, D. Sims, Y. Cherukuri, A. Sreedasyam, A. Kilian, A. Chan, M.-A. Van Sluys, K. Swaminathan, C. Town, H. Bergès, B. Simmons, J.C. Glaszmann, E. van der Vossen, R. Henry, J. Schmutz, A. D'Hont, A mosaic monoploid reference sequence for the highly complex genome of sugarcane, Nat. Commun. 9 (2018) 2638. https://doi.org/10.1038/s41467-018-05051-5.

[35] C. Tang, M. Yang, Y. Fang, Y. Luo, S. Gao, X. Xiao, Z. An, B. Zhou, B. Zhang, X. Tan, H.-Y. Yeang, Y. Qin, J. Yang, Q. Lin, H. Mei, P. Montoro, X. Long, J. Qi, Y. Hua, Z. He, M. Sun, W. Li, X. Zeng, H. Cheng, Y. Liu, J. Yang, W. Tian, N. Zhuang, R. Zeng, D. Li, P. He, Z. Li, Z. Zou, S. Li, C. Li, J. Wang, D. Wei, C.-Q. Lai, W. Luo, J. Yu, S. Hu, H. Huang, The rubber tree genome reveals new insights into rubber production and species adaptation, Nat. Plants. 2 (2016) 16073. https://doi.org/10.1038/nplants.2016.73.

[36] A.R.O. Conson, C.H. Taniguti, R.R. Amadeu, I.A.A. Andreotti, L.M. de Souza, L.H.B. dos Santos, J.R.B.F. Rosa, C.C. Mantello, C.C. da Silva, E. José Scaloppi Junior, R.V. Ribeiro, V. Le Guen, A.A.F. Garcia, P. de S. Gonçalves, A.P. de Souza, High-Resolution Genetic Map and QTL Analysis of Growth-Related Traits of Hevea brasiliensis Cultivated Under Suboptimal Temperature and Humidity Conditions, Front. Plant Sci. 9 (2018) 1255. https://doi.org/10.3389/fpls.2018.01255.

[37] L.M. Souza, R. Gazaffi, C.C. Mantello, C.C. Silva, D. Garcia, V. Le Guen, S.E.A. Cardoso, A.A.F. Garcia, A.P. Souza, QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (Hevea brasiliensis) Evaluated in a Sub-Tropical Climate, PLoS ONE. 8 (2013) e61238. https://doi.org/10.1371/journal.pone.0061238.

[38] J.C. Glaubitz, T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, E.S. Buckler, TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline, PLOS ONE. 9 (2014) e90346. https://doi.org/10.1371/journal.pone.0090346.

[39] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, R. Durbin, The variant call format and VCFtools, Bioinformatics. 27 (2011) 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

[40] B. Lawal, Categorical data analysis with SAS and SPSS applications, Lawrence Erlbaum Associates, Mahwah, N.J, 2003.

[41] J.R.B.F. Rosa, C.C. Mantello, D. Garcia, L.M. de Souza, C.C. da Silva, R. Gazaffi, C.C. da Silva, G. Toledo-Silva, P. Cubry, A.A.F. Garcia, A.P. de Souza, V. Le Guen, QTL detection for growth and latex production in a full-sib rubber tree population cultivated under suboptimal climate conditions, BMC Plant Biol. 18 (2018) 223. https://doi.org/10.1186/s12870-018-1450-y.

[42] D.D. Kosambi, The estimation of map distance from recombination values. Ann. Eugen. 12, (1944) 172–175. https://doi.org/10.1111/j.1469-1809.1943.tb02321.x.

[43] D. Lespinasse, M. Rodier-Goud, L. Grivet, A. Leconte, H. Legnate, M. Seguin, A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP, microsatellite, and isozyme markers, Theor. Appl. Genet. 100 (2000) 127–138. https://doi.org/10.1007/s001220050018.

[44] W. Pootakham, P. Ruang-Areerate, N. Jomchai, C. Sonthirod, D. Sangsrakru, T. Yoocha, K. Theerawattanasuk, K. Nirapathpongporn, P. Romruensukharom, S. Tragoonrung, S. Tangphatsornruang, Construction of a high-density integrated genetic linkage map of rubber tree (Hevea brasiliensis) using genotyping-by-sequencing (GBS), Front. Plant Sci. 6 (2015). https://doi.org/10.3389/fpls.2015.00367.

[45] L.A. Ouellette, R.W. Reid, S.G. Blanchard, C.R. Brouwer, LinkageMapView—rendering high-resolution linkage and QTL maps, Bioinformatics. 34 (2018) 306–307. https://doi.org/10.1093/bioinformatics/btx576.

[46] D. Money, K. Gardner, Z. Migicovsky, H. Schwaninger, G.-Y. Zhong, S. Myles, LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms, G3 GenesGenomesGenetics. 5 (2015) 2383–2390. https://doi.org/10.1534/g3.115.021667.
[47] F. Cericola, I. Lenk, D. Fè, S. Byrne, C.S. Jensen, M.G. Pedersen, T. Asp, J. Jensen, L. Janss, 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.), Front. Plant Sci. 9 (2018). https://doi.org/10.3389/fpls.2018.00369.
[48] I.S. Elbasyoni, A.J. Lorenz, M. Guttieri, K. Frels, P.S. Baenziger, J. Poland, E. Akhunov, A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat, Plant Sci. Int. J. Exp. Plant Biol. 270 (2018) 123–130. https://doi.org/10.1016/j.plantsci.2018.02.019.
[49] N. Nazzicari, F. Biscarini, P. Cozzi, E.C. Brummer, P. Annicchiarico, Marker imputation efficiency for genotyping-by-sequencing data in rice (Oryza sativa) and alfalfa (Medicago sativa), Mol. Breed. 36 (2016) 69. https://doi.org/10.1007/s11032-016-0490-y.
[50] I.S.C. Granato, G. Galli, E.G. de Oliveira Couto, M.B. e Souza, L.F. Mendonça, R. Fritsche-Neto, snpReady: a tool to assist breeders in genomic analysis, Mol. Breed. 38 (2018) 102. https://doi.org/10.1007/s11032-018-0844-8.
[51] M.P. Arruda, P.J. Brown, A.E. Lipka, A.M. Krill, C. Thurber, F.L. Kolb, Genomic Selection for Predicting Fusarium Head Blight Resistance in a Wheat Breeding Program, Plant Genome. 8 (2015). https://doi.org/10.3835/plantgenome2015.01.0003.
[52] J.B. Endelman, Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP, Plant Genome J. 4 (2011) 250. https://doi.org/10.3835/plantgenome2011.08.0024.
[53] J. Poland, J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, J.-L. Jannink, Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing, Plant Genome J. 5 (2012) 103. https://doi.org/10.3835/plantgenome2012.06.0006.
[54] Y.-B. Fu, Genetic Diversity Analysis of Highly Incomplete SNP Genotype Data with Imputations: An Empirical Assessment, G3 Genes Genomes Genet. 4 (2014) 891–900. https://doi.org/10.1534/g3.114.010942.
[55] T. Chud, R. Ventura, F. Schenkel, R. Carvalheiro, M. Buzanskas, I. Urbinati, L. Regitano, C. Marcondes, D. Munari, Imputation Accuracy using FImpute and BEAGLE Software in Brazilian Synthetic Cattle Breed, in: 2014.
[56] J.-S. Brouard, B. Boyle, E.M. Ibeagha-Awemu, N. Bissonnette, Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation, BMC Genet. 18 (2017) 32. https://doi.org/10.1186/s12863-017-0501-y.
[57] D. Cros, S. Bocs, V. Riou, E. Ortega-Abboud, S. Tisné, X. Argout, V. Pomiès, L. Nodichao, Z. Lubis, B. Cochard, T. Durand-Gasselin, Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses, BMC Genomics. 18 (2017) 839. https://doi.org/10.1186/s12864-017-4179-3.
[58] M.F.R. Resende, P. Muñoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando, J.M. Davis, E.J. Jokela, T.A. Martin, G.F. Peter, M. Kirst, Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.), Genetics. 190 (2012) 1503–1510. https://doi.org/10.1534/genetics.111.137026.
[59] J.H. Steiger, Tests for comparing elements of a correlation matrix, Psychol. Bull. 87 (1980) 245–251. https://doi.org/10.1037/0033-2909.87.2.245.
[60] W. Revelle, psych: Procedures for psychological, psychometric, and personality research, R package 1.8. 4, (2019).

[61] H. Song, S. Ye, Y. Jiang, Z. Zhang, Q. Zhang, X. Ding, Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs, Genet. Sel. Evol. GSE. 51 (2019) 58. https://doi.org/10.1186/s12711-019-0500-8.
[62] A.J. Lorenz, S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, J.-L. Jannink, Genomic Selection in Plant Breeding, in: Adv. Agron., Elsevier, 2011: pp. 77–123. https://doi.org/10.1016/B978-0-12-385531-2.00002-5.
[63] V. Le Guen, D. Garcia, F. Doaré, C.R.R. Mattos, V. Condina, C. Couturier, A. Chambon, C. Weber, S. Espéout, M. Seguin, A rubber tree's durable resistance to Microcyclus ulei is conferred by a qualitative gene and a major quantitative resistance factor, Tree Genet. Genomes. 7 (2011) 877–889. https://doi.org/10.1007/s11295-011-0381-7.

Tables

Table 1 Distribution of SSR and SNP markers on the genetic linkage map obtained from cross PB260 × RRIM600

Linkage group	Number of SSRs	Number of SNPs	Length (cM)	Average SNP interval (cM)	Maximum SNP interval (cM)
LG01	13	95	130.5	1.37	9.91
LG02	21	105	131.7	1.25	6.72
LG03	15	99	152.7	1.54	8.65
LG04	12	59	137.7	2.33	7.92
LG05	20	143	147.9	1.03	18.16
LG06a	12	51	114.1	2.24	9.14
LG06b	2	20	42.1	2.11	16.65
LG07	18	78	165.9	2.13	17.29
LG08	22	84	148.1	1.76	7.87
LG09	23	112	155.4	1.39	7.17
LG10	32	139	181.8	1.31	5.26
LG11	13	120	156.5	1.30	5.59
LG12	10	86	111.0	1.29	5.3
LG13	15	108	150.3	1.39	10.04
LG14	19	124	134.3	1.08	11.84
LG15	18	112	171.9	1.53	10.56
LG16	14	70	135.7	1.94	13.56
LG17	16	68	101.4	1.49	10.23

LG18	13	96	132.1	1.38	14.55
Average	16	93	136.9	1.47	-
Total	**308**	**1,769**	**2,600.9**	-	-

Table 2: Distribution of SNP markers on the genetic linkage map obtained from cross PR255 × PB217

Linkage group	Number of SNPs	Length (cM)	Average SNP interval (cM)	Maximum interval (cM)
LG01	191	127.1	0.67	10.51
LG02	175	114.9	0.66	6.39
LG03	186	142.3	0.77	6.83
LG04	179	128.7	0.72	7.08
LG05	159	128.8	0.81	6.63
LG06	160	107.6	0.67	5.3
LG07	216	145.9	0.68	10.43
LG08	180	138.9	0.77	10.85
LG09	206	144.6	0.7	8.01
LG10	178	171.2	0.96	9.05
LG10+LG17	70	61.6	0.88	7.69
LG11	111	86.6	0.78	6.31
LG12	92	83.9	0.91	5.78
LG13	186	140.1	0.75	7.53
LG14	192	123.5	0.64	4.16

LG15	195	90.3	0.46	5.23
LG16	199	85.6	0.43	3.81
LG17	181	114.1	0.63	5.36
LG18	155	79.4	0.51	6.18
Average	164	116.6	0.71	-
Total	**3,111**	**2,215.1**	-	-

Table 3. Performance of imputation methods in terms of GS accuracy after imputation. For dataset 1, when predictions were done in site 1, site 2 individuals were used to train the model and vice versa. Bold values represent the highest GS accuracy in respective comparisons (i.e. among the table row). LIR: LinkImputeR

Trait / site	***LIR***	***RFI***	***EMI***	***PPCAI***	***SVDI***	***kNNI***	***MNI***	***Beagle***	***FImpute***	***FImpute (pedigree)***
Dataset 1										
Mapped markers (1,769):										
Yield / site 1	**0.53**	0.52	0.52	0.52	**0.53**	0.5	0.48	**0.53**	0.52	0.52
Yield / site 2	**0.5**	0.48	0.48	0.47	0.48	0.46	0.45	0.49	0.49	0.47
Yield (mean)	***0.52***	*0.5*	*0.5*	*0.5*	*0.5*	*0.48*	*0.47*	*0.51*	*0.5*	*0.5*
Sucrose / site 1	0.15	0.14	0.14	0.13	0.14	**0.16**	0.13	0.15	0.15	0.14
Sucrose / site 2	0.22	0.21	0.21	0.2	0.22	0.19	0.17	0.24	**0.25**	0.23
Sucrose (mean)	*0.18*	*0.18*	*0.17*	*0.17*	*0.18*	*0.19*	*0.15*	*0.19*	***0.2***	*0.18*
All markers (3,420):										

Yield / site 1	**0.53**	0.5	0.51	0.5	0.5	0.5	0.48	-	-	-
Yield / site 2	**0.5**	0.45	0.46	0.45	0.45	0.45	0.44	-	-	-
Yield (mean)	***0.5***	*0.48*	*0.48*	*0.48*	*0.48*	*0.47*	*0.46*	-	-	-
Sucrose / site 1	0.14	0.14	0.15	0.15	0.15	**0.16**	0.14	-	-	-
Sucrose / site 2	0.17	0.18	0.18	**0.19**	**0.19**	0.15	0.16	-	-	-
Sucrose (mean)	*0.16*	*0.16*	***0.17***	***0.17***	***0.17***	*0.16*	*0.15*	-	-	-
Trait	*LIR*	*RFI*	*EMI*	*PPCAI*	*SVDI*	*kNNI*	*MNI*	*Beagle*	*FImpute*	*FImpute (pedigree)*
Dataset 2										
Mapped markers (3,111):										
Circumference	**0.21**	0.19	0.19	0.19	0.19	0.2	0.2	0.2	0.2	0.2
All markers (9,088):										
Circumference	**0.22**	0.18	0.18	0.18	0.18	0.2	0.2	-	-	-

Figures

Figure 1. Distribution of mean depth per SNP (top), and percentage of missing data per SNP (bottom) in dataset 1 (left) and 2 (right)

Figure 2. SNP map of cross PB260 × RRIM600, containing 1,769 SNPs. SNP density increases from blue to red colors.

Figure 3. SNP map of cross PR255 × PB217, containing 3,111 SNPs. SNP density increases from blue to red colors.

Figure 4. Percentage of correctly imputed genotypes according to imputation method in dataset 1 (left) and dataset 2 (right). Figures are means over three replicates. Values with the same letter are not significantly different within a dataset at P = 1%.

Figure 5. Effect of imputation approach and marker density on GS accuracy in site 1 (left) and site 2 (right) (dataset 1) for rubber yield. When not all markers were used, values are means over 45 replicates. FIPed/FIPnoP: FImpute using / not using pedigree information.

Figure 6. Effect of imputation approach and marker density on GS accuracy in site 1 (left) and site 2 (right) (dataset 1) for sucrose content. When not all markers were used, values are means over 45 replicates. FIPed/FIPnoP: FImpute using / not using pedigree information.

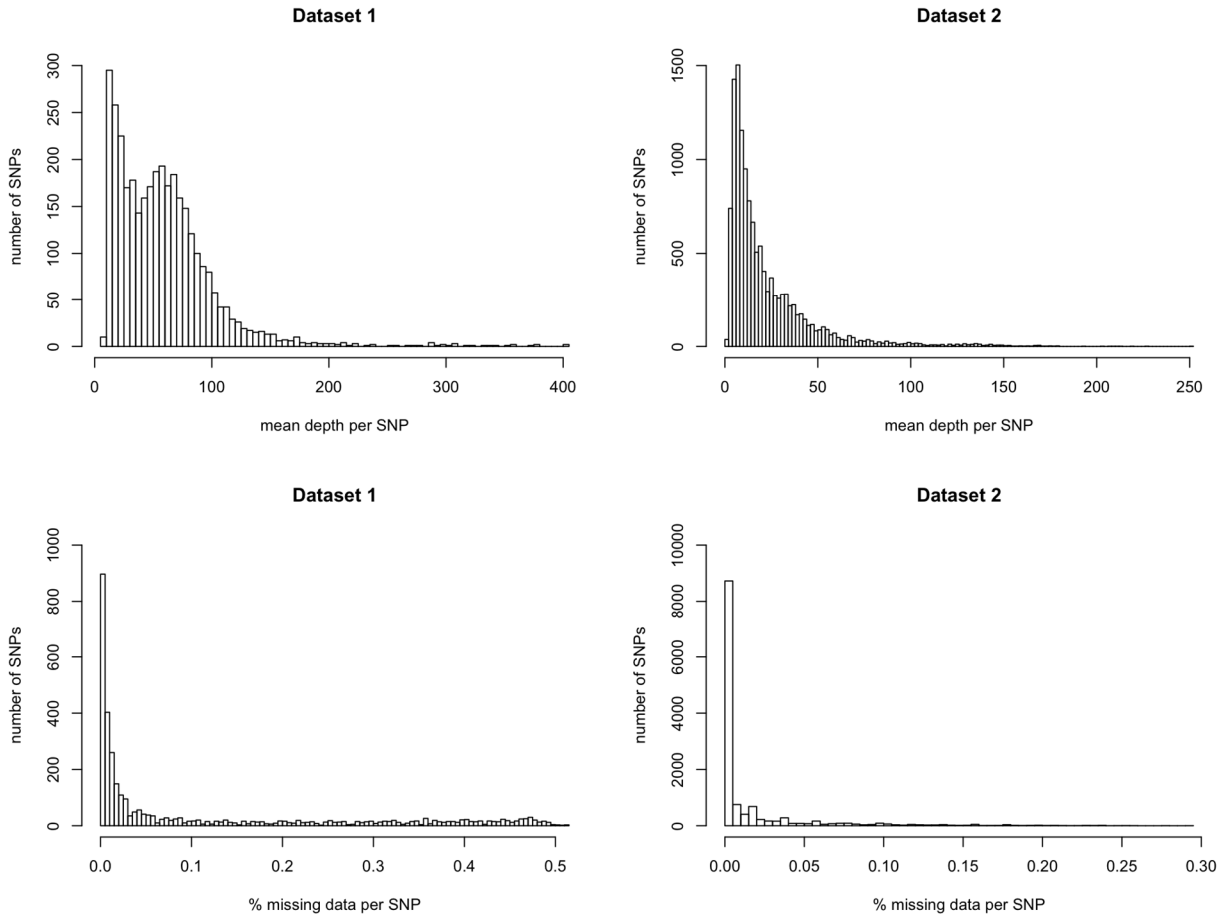
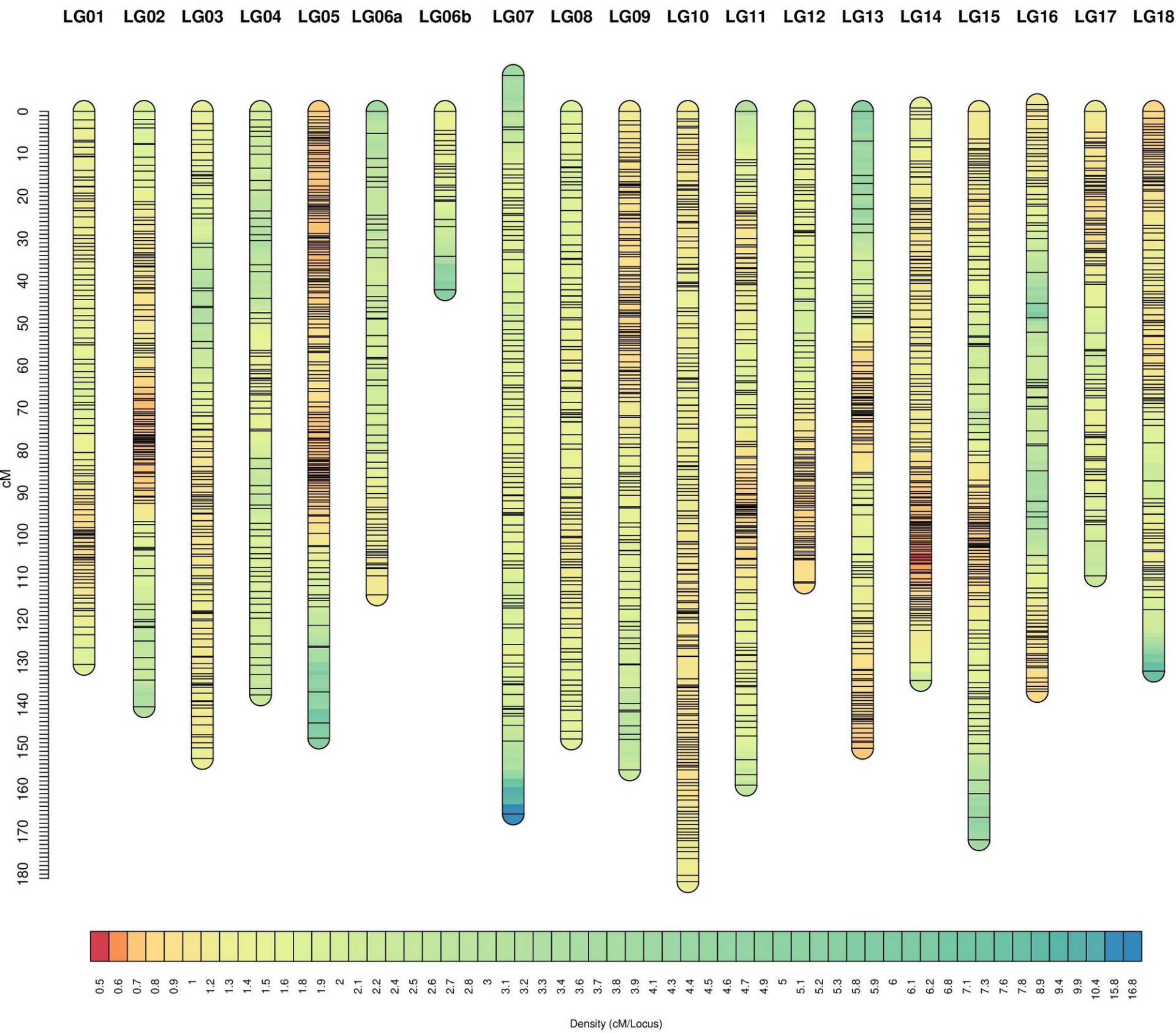
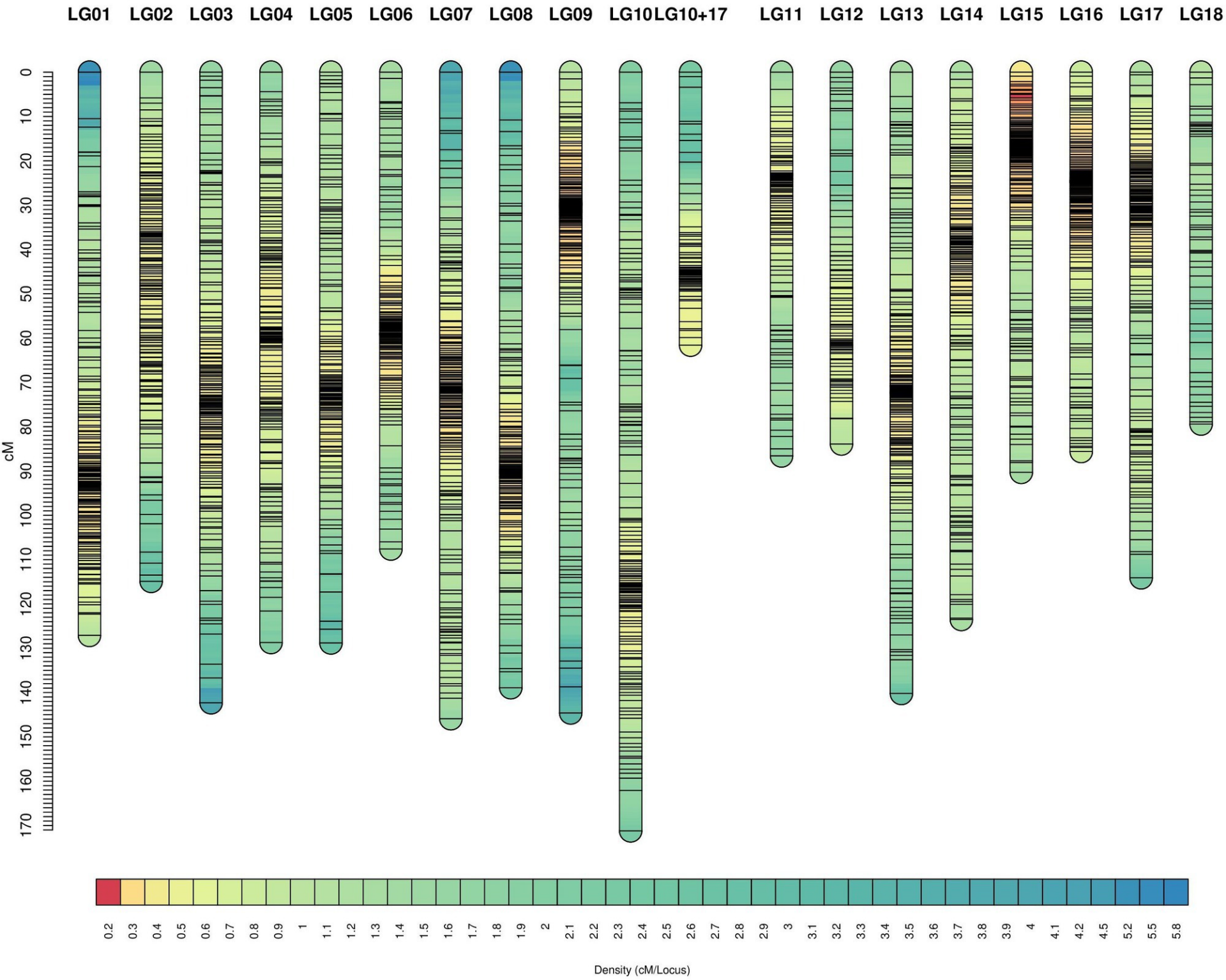


Figure 1. Distribution of mean depth per SNP (top), and percentage of missing data per SNP (bottom) in dataset 1 (left) and 2 (right)


LG01
LG02
LG03
LG04
LG05
LG06a
LG06b
LG07
LG08
LG09
LG10
LG11
LG12
LG13
LG14
LG15
LG16
LG17
LG18
cM
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
0.5
0.6
0.7
0.8
0.9
1
1.2
1.3
1.4
1.5
1.6
1.8
1.9
2
2.1
2.2
2.4
2.5
2.6
2.7
2.8
3
3.1
3.2
3.3
3.4
3.6
3.7
3.8
3.9
4.1
4.3
4.4
4.5
4.6
4.7
4.9
5
5.1
5.2
5.3
5.8
5.9
6
6.1
6.2
6.8
7.1
7.3
7.6
7.8
8.9
9.4
9.9
10.4
15.8
16.8
Density (cM/Locus)


LG01
LG02
LG03
LG04
LG05
LG06
LG07
LG08
LG09
LG10
LG10+17
LG11
LG12
LG13
LG14
LG15
LG16
LG17
LG18
cM
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
4.1
4.2
4.5
5.2
5.5
5.8
Density (cM/Locus)

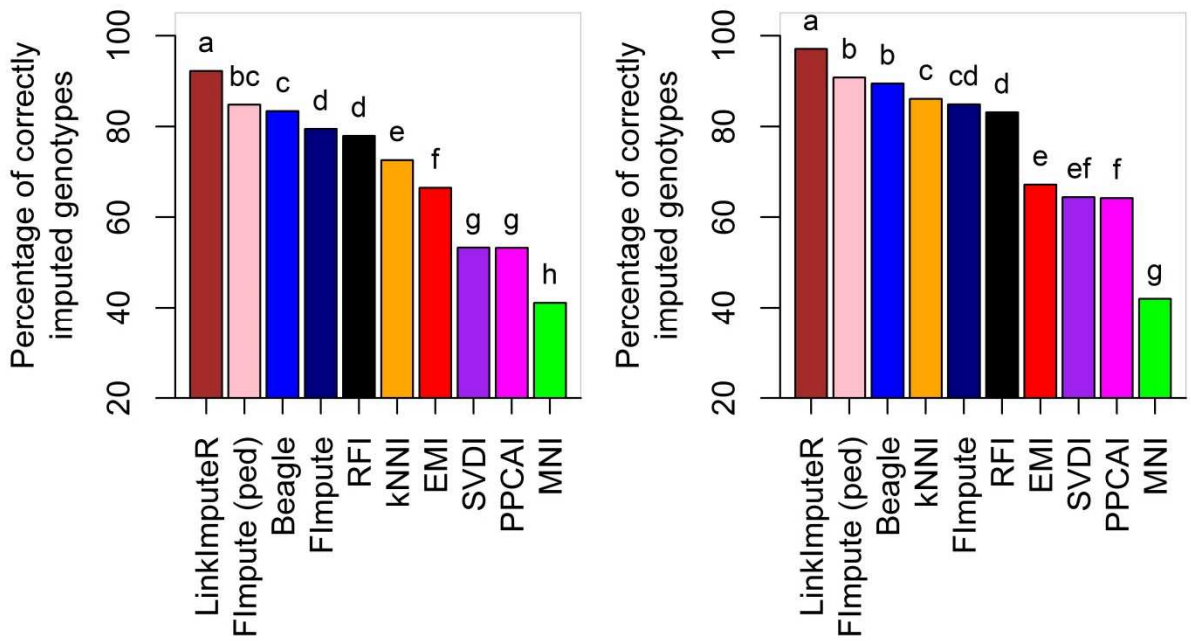


Figure 4 Percentage of correctly imputed genotypes according to imputation method in dataset 1 (left) and dataset 2 (right). Figures are means over three replicates. Values with the same letter are not significantly different within a dataset at P = 1%.

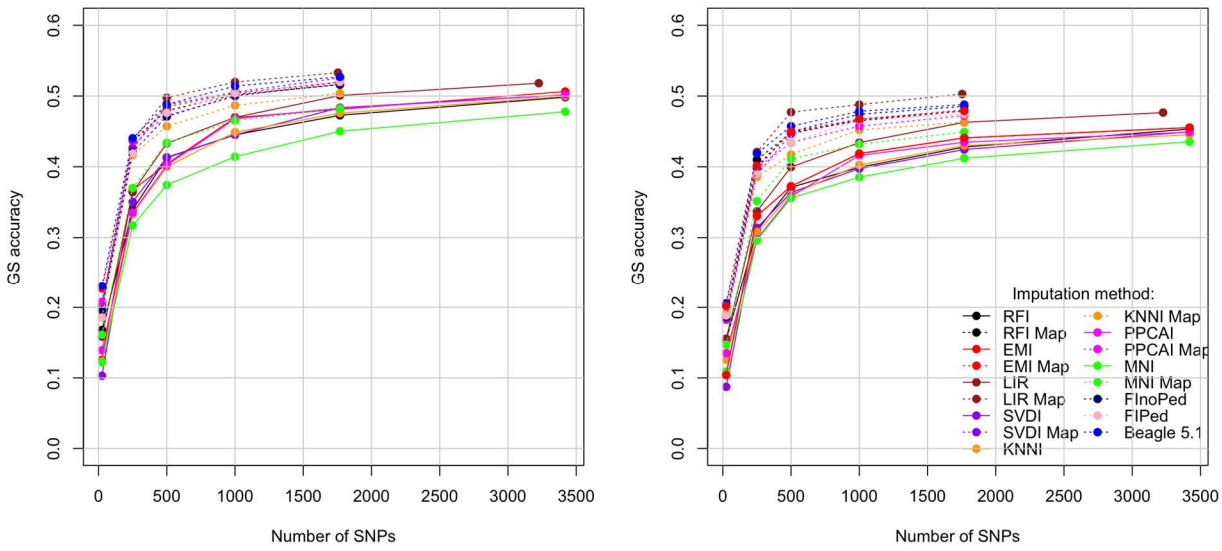


Figure 5. Effect of imputation approach and marker density on GS accuracy in site 1 (left) and site 2 (right) (dataset 1) for rubber yield. When not all markers were used, values are means over 45 replicates. FIPed/FIPnoP: FImpute using / not using pedigree information.

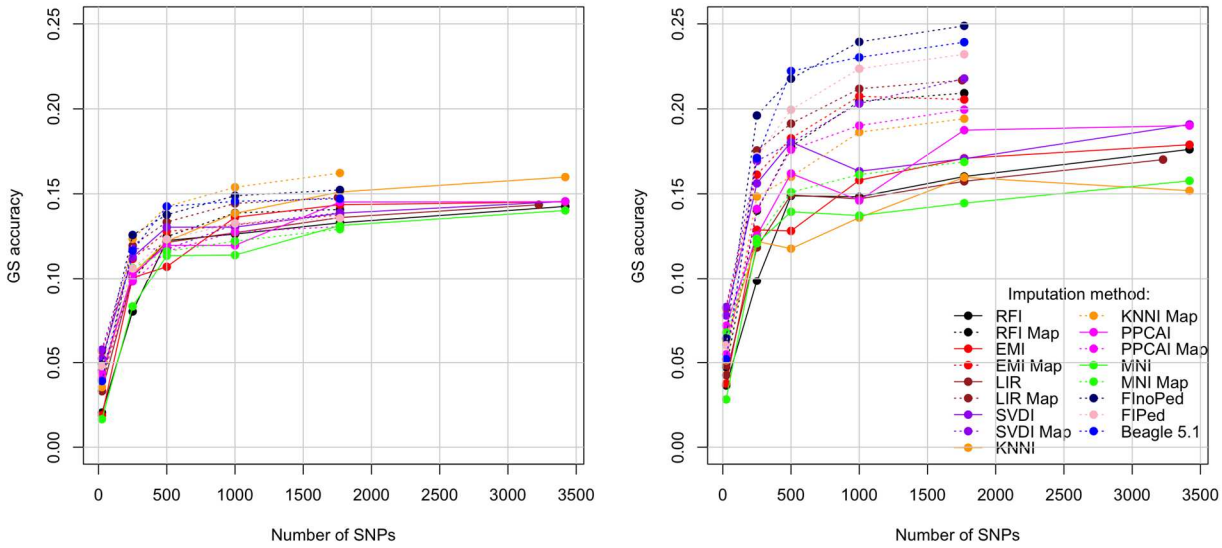


Figure 6. Effect of imputation approach and marker density on GS accuracy in site 1 (left) and site 2 (right) (dataset 1) for sucrose content. When not all markers were used, values are means over 45 replicates. FIPed/FIPnoP: FImpute using / not using pedigree information.