Published August 15, 2019

Expected Genotype Quality and Diploidized Marker Data from Genotyping-by-Sequencing of *Urochloa* spp. Tetraploids

Filipe Inácio Matias, Karem Guimarães Xavier Meireles, Sheila Tiemi Nagamatsu, Sanzio Carvalho Lima Barrios, Cacilda Borges do Valle, Marcelo Falsarella Carazzolle, Roberto Fritsche-Neto, and Jeffrey B. Endelman*

F.I. Matias, R. Fritsche-Neto, Genetics Dep., Luiz de Queiroz College of Agriculture, Univ. of São Paulo, Av. Pádua Dias, 11, C. P. 9, 13.418-900, Piracicaba, São Paulo, Brazil; S.C.L. Barrios, K.G.X. Meireles, C.B. do Valle, Embrapa Beef Cattle, Av. Rádio Maia, 830, Zona Rural, 79.106-550, Campo Grande, Mato Grosso do Sul, Brazil; S.T. Nagamatsu, M.F. Carazzolle, Genetics and Evolution Dep., Univ. of Campinas, Cidade Univ. Zeferino Vaz, 13.083-970, Campinas, São Paulo, Brazil; J.B. Endelman, Dep. Horticulture, Univ. of Wisconsin–Madison, Madison, WI, US, 53706.

ABSTRACT Although genotyping-by-sequencing (GBS) is a well-established marker technology in diploids, the development of best practices for tetraploid species is a topic of current research. We determined the theoretical relationship between read depth and the phred-scaled probability of genotype misclassification conditioned on the true genotype, which we call expected genotype quality (EGQ). If the GBS method has 0.5% allelic error, then 17 reads are needed to classify simplex tetraploids as heterozygous with 95% accuracy (EGQ = 13) vs. 61 reads to determine allele dosage. We developed an R script to convert tetraploid GBS data in variant call format (VCF) into diploidized genotype calls and applied it to 267 interspecific hybrids of the tetraploid forage grass *Urochloa*. When reads were aligned to a mock reference genome created from GBS data of the *Urochloa brizantha* (Hochst. ex A. Rich.) R. D. Webster cultivar Marandu, 25,678 biallelic single nucleotide polymorphism (SNPs) were discovered, compared with ~3000 SNPs when aligning to the closest true reference genomes, *Setaria viridis* (L.) P. Beauv. and *S. italica* (L.) P. Beauv. Cross-validation revealed that missing genotypes were imputed by the random forest method with a median accuracy of 0.85 regardless of heterozygote frequency. Using the *Urochloa* spp. hybrids, we illustrated how filtering samples based only on genotype quality (GQ) creates genotype bias; a depth threshold based on EGQ is also needed regardless of whether genotypes are called using a diploidized or allele dosage model.

Abbreviations: EGQ, expected genotype quality; EMBRAPA, Brazilian Agricultural Research Corporation; GATK, Genome Analysis Toolkit; GBS, genotyping-by-sequencing; GQ, genotype quality; HMA, homozygous for the major allele; LD, linkage disequilibrium; MAD, minor allele depth; MAP, maximum a posteriori; PCR, polymerase chain reaction; SNP, single nucleotide polymorphism; VCF, variant call format.

CORE IDEAS

- Introduced concept of expected genotype quality (EGQ) and software to calculate it
- Provided read depth guidelines for GBS in tetraploids
- Developed software to generate diploidized genotype calls from VCF files
- Demonstrated value of aligning GBS reads to a mock reference genome for SNP discovery
- Recommend filtering based on GQ and read depth to prevent genotype bias

Urochloa is the most cultivated genus as pasture on tropical livestock farms because of its tolerance to acidic soils, good carrying capacity, insect resistance, and nutritional value (Jank et al., 2014; Pessoa-Filho et al., 2017). The most economically important species are *U. decumbens* (Stapf) R. D. Webster (syn. *Brachiaria decumbens* Stapf) and *U. brizantha* (Hochst. ex A. Rich.) R. D. Webster [syn. *B. brizantha* (Hochst. ex A. Rich.) Stapf], which are both tetraploid ($2n = 4x = 36$). Apomixis is the normal mode of reproduction in these species, and

Citation: Matias, F.I., K.G.X. Meireles, S.T. Nagamatsu, S.C.L. Barrios, C.B. do Valle, M.F. Carazzolle, R. Fritsche-Neto, and J.B. Endelman. 2019. Expected genotype quality and diploidized marker data from genotyping-by-sequencing of *Urochloa* spp. tetraploids. Plant Genome 12:190002. doi: 10.3835/plantgenome2019.01.0002

Received 4 Jan. 2019. Accepted 29 Apr. 2019.
*Corresponding author (endelman@wisc.edu).

for many years, genetic improvement in South America was based on screening new introductions from Africa (Miles, 2007; Jank et al., 2011). To facilitate breeding by sexual hybridization, Swenne et al. (1981) used colchicine-induced tetraploids of the diploid species *U. ruziziensis* (R. Germ. & C. M. Evrard) Crins ($2n = 2x = 18$) as female parents to cross with apomictic tetraploids. This interspecific hybridization scheme has become the foundation of the *Urochloa* spp. breeding programs at the International Center for Tropical Agriculture in Colombia and the Brazilian Agricultural Research Corporation (EMBRAPA) (Lutts et al., 1991; Miles et al., 2006; Monteiro et al., 2016). As in other crops, genome-wide markers can provide significant value for *Urochloa* spp. breeding programs. Several previous studies have used microsatellite markers to study population structure in *Urochloa* (Jungmann et al., 2010; Vigna et al., 2011; Silva et al., 2013), but the ubiquity and cost-effectiveness of SNPs are advantageous for discovering genetic variants and predicting complex traits.

Arrays and GBS of multiplexed, reduced-representation libraries have been used to generate large, biallelic SNP datasets in heterozygous tetraploids, including potato (*Solanum tuberosum* L.) (Felcher et al., 2012; Uitdewilligen et al., 2013), alfalfa (*Medicago sativa* L.) (Li et al., 2014), rose (*Rosa* L.) (Koning-Boucoiran et al., 2015), kiwi [*Actinidia deliciosa* (A. Chev.) C. F. Liang & A. R. Ferguson] (Melo et al., 2016), and *Urochloa* spp. (Worthington et al., 2016; Ferreira et al., 2019). Both arrays and GBS generate a signal for each allele that can be used to predict allele dosage, that is, the tetraploid genotype. For the SNP array, signal intensity is not necessarily proportional to allele dosage, and therefore, different classification algorithms have been explored (Voorrips et al., 2011; Serang et al., 2012; Schmitz Carley et al., 2017). For GBS data, the allele signal intensity is the read count, which can be analyzed using the aforementioned classifiers, but the focus of this manuscript is genotype calling based on a binomial model. The binomial model is central to well-established software packages such as the Genome Analysis Toolkit (GATK) (McKenna et al., 2010; DePristo et al., 2011) and FreeBayes (Garrison and Marth, 2012) as well as more recent tools developed specifically for polyploids (Blischak et al., 2018; Gerard et al., 2018; Clark et al., 2019).

It is generally recognized that higher read depth is needed to estimate allele dosage in polyploids, but the literature contains a number of different approaches and recommendations. Uitdewilligen et al. (2013) developed KASP assays for 270 GBS markers in potato and compared the genotype calls from the two methods; the results under different filtering criteria led the authors to conclude that "~60–80× can be used as a lower boundary for reliable assessment of allele copy number…" Bastien et al. (2018) used a threshold of 53 reads for determining allele dosage in potato because it was deemed "sufficient to distinguish between the five expected genotypic classes based on a chi-square distribution." Gerard et al. (2018) developed software to investigate the effects of allelic bias, overdispersion, and sequencing error on read-depth thresholds.

Our approach is similar to Gerard et al. (2018) (and was developed independently) in that a binomial model is used to estimate the probability of genotype misclassification. However, whereas Gerard et al. (2018) reported population-level statistics, our focus is the difference between simplex and duplex genotypes. We also account for the nonmonotone relationship between the probability of genotype misclassification and read depth (see Methods section). Our results are reported on the phred-scale (defined as $-10 \log_{10} q$, where q is the error probability), analogous to the GQ field of the VCF (Danecek et al., 2011). However, whereas GQ is conditioned on the predicted genotype, our metric is conditioned on the true genotype. Because the metric can be viewed as an expectation over all possible allele counts for a given total read depth, we call it EGQ.

Expected genotype quality was used to guide the analysis of GBS data for a panel of 267 tetraploid *U. ruziziensis* × *U. brizantha* hybrids. Because few markers had sufficient read depth to determine allele dosage with reasonable accuracy, genotype calls were made using a diploid approximation, in which the three heterozygotes were not distinguished. This approximation is common for GBS in heterozygous tetraploids, and typically a threshold of 11 reads is used to ensure the probability of misclassifying a heterozygote as homozygous is <5% (Li et al., 2014). However, this threshold is based on the assumption of no error in the GBS method, and our theoretical treatment elucidates how the threshold increases with error.

Even with a diploid approximation, the *Urochloa* dataset contained missing data. Imputation of missing genotypes in GBS datasets has been studied extensively in inbred lines and heterozygous diploids, with hidden Markov models being the preferred method when a genetic or physical map for the markers is available (Hickey et al., 2012; Swarts et al., 2014; Fragoso et al., 2015). When a map is not available, as was the case for the *Urochloa* spp. hybrids, the random forest algorithm (Breiman, 2001) can still be used and has performed well in other species (Rutkoski et al., 2013; Money et al., 2015). Our objectives were to evaluate different filtering criteria, references genomes, and imputation accuracy for the *Urochloa* dataset.

MATERIALS AND METHODS

Expected Genotype Quality

A binomial model was used to determine the statistical relationship between read depth and EGQ. Let $f(k,N,\rho)$ denote the probability mass function for the binomial distribution with k successes out of N trials and success probability ρ. The likelihood of observing k reads of the alternate allele given N total reads for tetraploid genotype $x \in \{0,1,2,3,4\}$ was modeled as $f\left(k,N,\rho_x = \frac{x}{4}[1-\varepsilon]+\left[1-\frac{x}{4}\right]\varepsilon\right)$, where the allelic error rate ε is the probability that a read is generated by one allele but counted toward the other (e.g., as a result of errors during polymerase chain reaction [PCR] or sequencing). Under a uniform prior, the maximum a posteriori (MAP) tetraploid genotype call for the observed result (k,N) is the value of x that maximizes f. For some values of k, the MAP solution does not equal the true value. Summing f over these values of k and expressing the result on the phred scale leads to the following expression for $\mathrm{EGQ}_{\mathrm{tet}}$:

$$\mathrm{EGQ}_{\mathrm{tet}}(x,N,\varepsilon) = -10\log_{10}\sum\nolimits_{k=0}^{N} f(k,N,\rho_x)\left[1-\delta(x,\mathrm{MAP}_{\mathrm{tet}})\right] \quad [1]$$

The symbol δ in Eq. [1] is the Kronecker delta function, which equals 1 when its two arguments are equal and 0 when they are unequal.

For diploidized genotype calls, the three possible genotypic states are denoted {A, H, B}, where the heterozygous state H = dosages 1, 2, or 3, and the homozygous states A = dosage 0 and B = dosage 4. The corresponding three-vector of posterior probabilities is proportional to $(p_A, p_H, p_B) \equiv (f_0, f_1 + f_2 + f_3, f_4)$, and the MAP solution (under a uniform prior) for the observed result (k,N) is the value of j that maximizes p_j. For some values of k, the MAP solution does not equal the diploidized genotype y corresponding to the true tetraploid state x. Summing f over these values of k and expressing the result on the phred scale leads to the following expression for $\mathrm{EGQ}_{\mathrm{dip}}$:

$$\mathrm{EGQ}_{\mathrm{dip}}(x,N,\varepsilon) = -10\log_{10}\sum\nolimits_{k=0}^{N} f(k,N,\rho_x)\left[1-\delta(y,\mathrm{MAP}_{\mathrm{dip}})\right] \quad [2]$$

Although Eq. [1] and [2] tend to increase with read depth, they are not monotone functions of N. Our results for EGQ correspond to the following monotone extension:

$$\phi(x,N,\varepsilon) = \min\nolimits_{M \geq N} \mathrm{EGQ}(x,M,\varepsilon) \quad [3]$$

which has the property $\phi(x,N,\varepsilon) \geq \phi(x,M,\varepsilon)$ for $N > M$. Using the R programming language (R Development Core Team, 2017), a function was created (Supplemental File S1) to calculate $\phi(x,N,\varepsilon)$.

Genotyping-by-Sequencing of *Urochloa* Species

Genomic DNA was extracted using the Qiagen DNeasy kit for 267 tetraploid *U. ruziziensis* × *U. brizantha* hybrids from EMBRAPA, as well as for the *U. brizantha* Marandu. The GBS libraries were prepared according to Elshire et al. (2011) using the *Ape*KI enzyme and sequenced on five lanes of the Illumina Hi-Seq 2500 platform with 1×100 bp reads. Reads were demultiplexed and trimmed using Cutadapt (Martin, 2011) and then aligned to five different Poaceae family genomes with bwa-mem (Li, 2013): *Setaria viridis* (v1.1. http://phytozome.jgi.doe.gov/), *Setaria italica* (Bennetzen et al., 2012), sorghum [*Sorghum bicolor* (L.) Moench] (v3.1. http://phytozome.jgi.doe.gov/), rice (*Oryza sativa* L.) (Ouyang et al., 2006), and corn (*Zea mays* L.) (Schnable et al., 2009). The alignment percentage for each reference was evaluated with Bowtie2 (Langmead and Salzberg, 2012). Reads were also aligned to a mock reference genome generated from the reads for Marandu with the software GBS–SNP–CROP (Melo et al., 2016). The GATK (McKenna et al., 2010; DePristo et al., 2011) HaplotypeCaller was used for SNP discovery with the ploidy flag set to 4, followed by removal of SNPs that did not meet the recommended thresholds (Broad Inst., 2016): Fisher strand bias (FS) $\leq$ 60.0, RMS mapping quality (MQ) $\geq$ 40.0, rank sum test for mapping quality (MQRankSum) $\geq$ −12.5, rank sum test for read position (ReadPosRankSum) $\geq$ −8.0.

Using the R programming language, a function was created (readVCF, Supplemental File S2) to process the VCF file and perform additional filtering. Only biallelic SNPs were retained. The VCF file includes variants relative to the reference genome regardless of whether they are polymorphic in the genotyped population. To identify polymorphic markers, the total number of reads for the minor allele, or minor allele depth (MAD), was calculated for each marker based on the allele depth field, and variants with MAD < 2 were removed. The Genome Analysis Toolkit calculates allele frequency based on the dosage of called genotypes, which was deemed unreliable because of low read depth. A suitable proxy for filtering that does not require allele dosage information is the frequency of genotypes homozygous for the major allele (HMA), which was capped at 0.99. For each sample, GATK provides the phred-scaled likelihood for each of the five tetraploid genotypes, which was converted into a posterior probability p_i for genotype $i \in \{0,1,2,3,4\}$ (assuming a uniform prior) by the following:

$$p_i = \frac{10^{-\mathrm{PL}_i/10}}{\sum_{i=0}^{4} 10^{-\mathrm{PL}_i/10}}$$

The tetraploid genotype call corresponds to the largest probability, and $\mathrm{GQ}_{\mathrm{tet}} = -10\log_{10}(1 - \max_i p_i)$.

Because of the low read depth per sample in the *Urochloa* dataset, diploidized genotype calls were made in which the three heterozygous genotypes were not differentiated. This corresponds to defining a new vector of posterior probabilities, $\tilde{\mathbf{p}} = (p_0, p_1 + p_2 + p_3, p_4)$, in which the probability of the heterozygous state is the sum of the probabilities for the simplex, duplex, and triplex genotypes. The diploidized genotype call corresponds to the largest probability, and $\mathrm{GQ}_{\mathrm{dip}} = -10\log_{10}(1 - \max_i \tilde{p}_i)$.

Missing genotypes in the diploidized marker dataset were imputed with the R package randomForest (Liaw and Wiener, 2002; Supplemental File S3), which is based on the algorithms in Breiman (2001). For each marker, a training set of 100 hybrids was randomly selected from the hybrids with genotypes, and all other hybrids with genotype data were masked and used for validation. Because each marker had no more than 50% missing data, this ensured at least 33 hybrids were available for validation. We used 300 classification trees for prediction, and all markers with $r^2 \geq 0.1$ were used as m potential predictors. We used the default setting of randomly sampling $\sqrt{m}$ predictors at each split. Classification accuracy is the proportion of hybrids in the validation set for which the predicted genotype is correct. Accuracy results were binned by heterozygosity (with a constant bin range of 0.1) and reported at the midpoint for each bin (e.g., 0.3 for bin 0.25–0.35). As a baseline for comparison, the missing genotypes for each marker were also imputed with the population mode (i.e., the most common genotype).

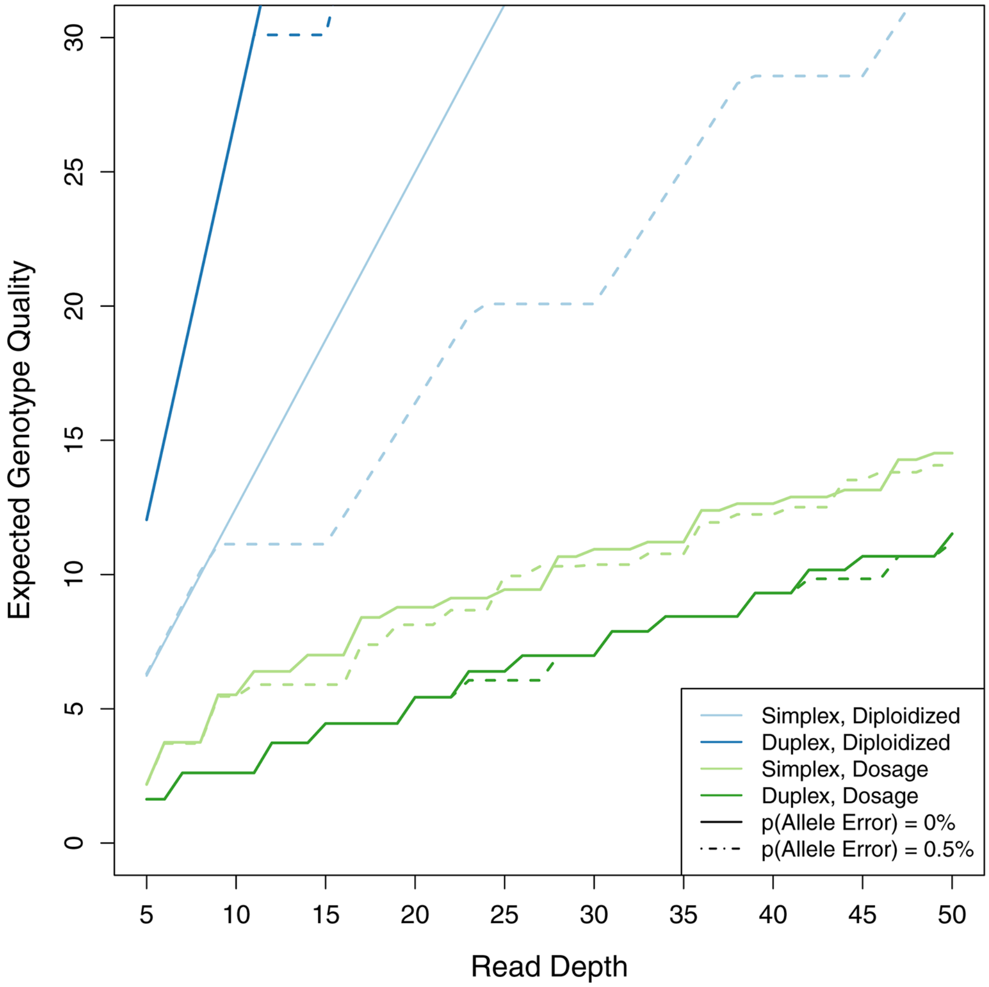

Fig. 1. Expected genotype quality (EGQ) as a function of read depth for two different allelic error rates.

RESULTS

Expected Genotype Quality

The EGQ is the phred-scaled probability of genotype misclassification conditioned on the true value. An EGQ of 13 corresponds to 95% genotype accuracy, and a score of 20 corresponds to 99% accuracy. Figure 1 shows EGQ for simplex and duplex genotypes as a function of total read depth. The allelic error rate, defined as the probability that a read is generated by one allele but counted toward the other (e.g., as a result of errors during PCR or sequencing), also affects EGQ. The blue lines in Fig. 1 correspond to diploidized genotypes, for which misclassifying simplex samples as homozygous (instead of heterozygous) is more likely than misclassifying duplex samples. The green lines correspond to genotype calls based on allele dosage, for which the relative EGQ of the two types of heterozygotes is reversed; misclassifying duplex samples is more likely than misclassifying simplex ones. The intuitive reason for this result is that a duplex genotype can appear as either simplex or triplex as a result of sampling variation, but comparable uncertainty for the simplex genotype exists only in the direction of higher dosage (i.e., with the duplex). In the absence of error (solid lines), 11 reads are needed to make diploidized genotype calls with EGQ 13 vs. 61 reads for determining allele dosage. Allelic errors have a greater effect on $\mathrm{EGQ}_{\mathrm{dip}}$ than $\mathrm{EGQ}_{\mathrm{tet}}$. With 0.5% error (dashed lines), the minimum depth needed to achieve EGQ 13 for diploidized genotypes increases to 17 reads, while the minimum depth for determining allele dosage remains 61.

Genotyping-by-Sequencing of *Urochloa* Species Hybrids

As no reference genome for the *Urochloa* spp. hybrids was available, the reference genomes of five other Poaceae species were evaluated for alignment. Figure 2 shows the number and percentage of aligned reads from the *Ape*KI-reduced representation of the *U. brizantha* cultivar Marandu. The percentage of reads aligned was low for all genomes, ranging from 1.92% for rice to 7.88% for *S. italica*. For both *Setaria* species and sorghum, over three-fourths of the aligned reads mapped to a unique location. For rice and corn, this proportion decreased to one half. The same five genomes were compared with respect to variant discovery in a panel of 267 tetraploid *U. ruziziensis* × *U. brizantha* hybrids. After removing variants with median depth less than eight, the two *Setaria* species generated the most biallelic SNPs (2809–3203) (Table 1).

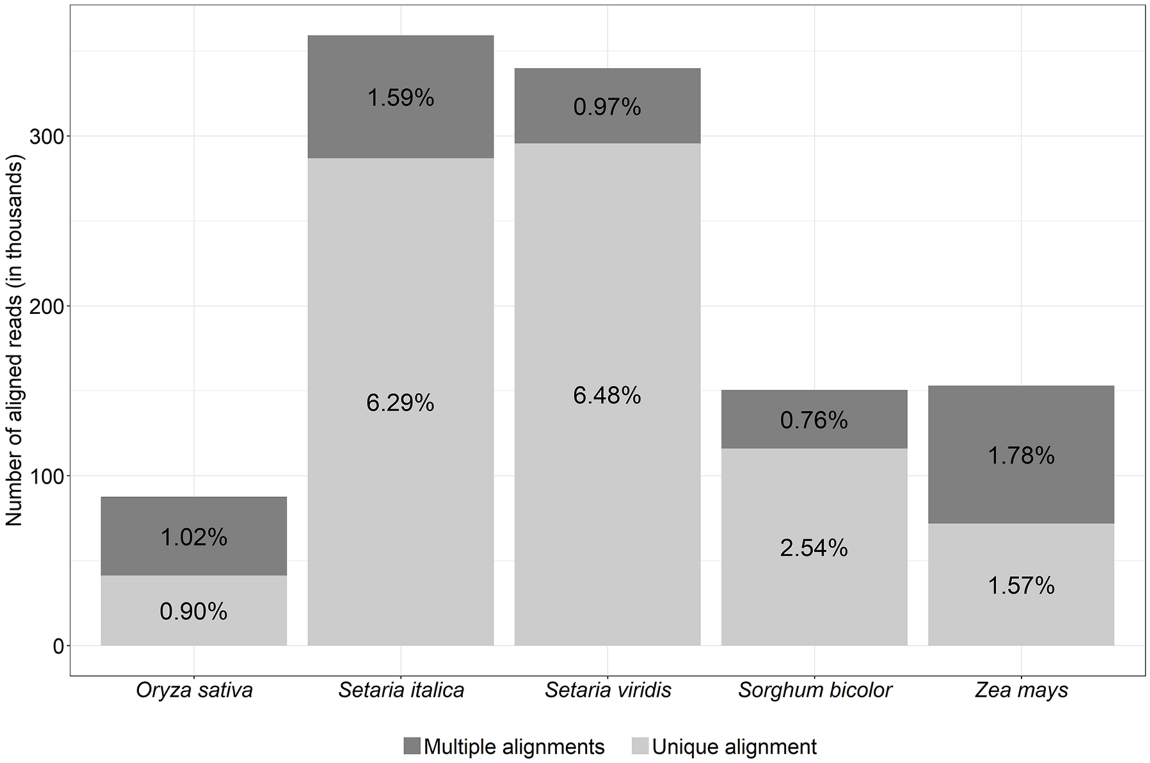

Fig. 2. Number and percentage of reads from the *U. brizantha* cultivar 'Marandu' that aligned to five Poaceae reference genomes.

To better utilize the GBS reads, a mock reference genome was built by clustering the trimmed reads from Marandu into 1,309,910 nonredundant consensus sequences, or centroids (Melo et al., 2016). A highly repetitive sequence was detected in the centroids, for which the first 50 bp are

GAGATCGGAAGAGCGGTTCAGCAGG-
AATGCCGAGACCGATCTCGTATGCC.

The entire 50 bp was present in 3.3% of the centroids, and when truncated to the first 40 or 30 bp, the frequency increased to 8.5 and 14.9%, respectively. The repetitive sequence was also detected in all 267 hybrids. A nucleotide BLAST search of the 50-bp sequence against the NCBI database returned highly significant matches to a diverse set of species, including *Larimichthys crocea* and *Cyprinus carpio* (100% identity across 49 bp), *Triticum aestivum* L., and *Solanum pennellii* Correll (98% identity across 50 bp).

When the GBS reads for the 267 hybrids were aligned to the centroids, the number of biallelic SNPs with median depth greater than eight increased to 25,678 (Table 1). A depth threshold of eight reads corresponds to $EGQ_{dip} \geq 10$ at 0.5% allelic error, whereas a depth threshold of 47 is needed for $EGQ_{tet} \geq 10$. As only 1955 SNPs had median depth greater than 47, tetraploid genotype calls were not pursued.

Table 1. Number of biallelic single nucleotide polymorphisms (SNPs) with <50% missing data based on a minimum sample depth of eight.

Reference	No. SNPs
Urochloa brizantha	25,678
Setaria viridis	3203
Setaria italica	2809
Sorghum bicolor	1331
Zea mays	763
Oryza sativa	571

Figure 3 is a histogram of the GQ_{dip} scores for all 153,589 diploidized genotypes (sample × marker combinations) with depth equal to eight in the filtered dataset. For homozygous genotypes, GQ_{dip} was peaked at 10, while for heterozygotes GQ_{dip} exceeded 30. The lower GQ for homozygotes is related to the low EGQ for simplex genotypes: heterozygous genotype calls have strong support because both alleles have been observed, whereas homozygous calls could be the result of misclassifying a simplex sample.

The cumulative distribution in Fig. 4 reveals the SNP dataset is dominated by rare alleles. The *x*-axis of Fig. 4 is the genotype frequency of hybrids homozygous for the major allele, and the *y*-axis is the proportion of SNPs for which the HMA frequency is less than or equal to the *x*-axis value. The SNP counts in Table 1 are based on an upper limit of 0.99 for HMA, and 51% of the SNPs discovered with the mock reference genome had HMA genotype frequencies between 0.95 and 0.99.

Genotype Imputation

The success of genotype imputation depends on the amount of linkage disequilibrium (LD) between markers, which is often quantified by the physical distance at which r^2 (the squared correlation) drops below some threshold. Since a physical reference genome was unavailable for this study, LD was quantified based on the maximum r^2 for

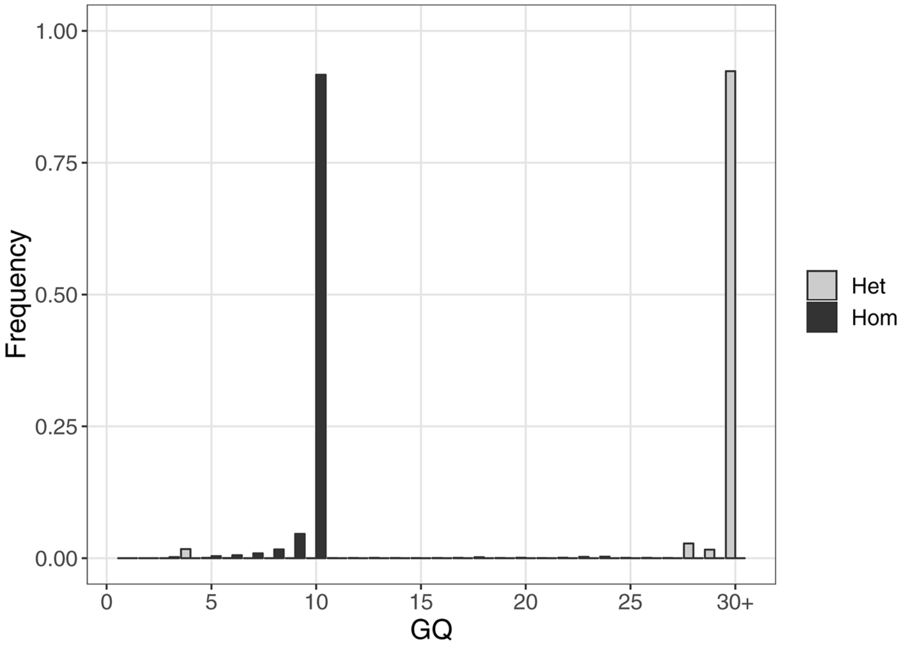

Fig. 3. Distribution of genotype quality (GQ_{dip}) scores for diploidized genotypes of *Urochloa* spp. hybrids with sample depth equal to eight. The figure is based on 25,678 SNPs, discovered using the *U. brizantha* mock reference genome for alignment. Heterozygous samples (Het) are shown in light gray, and homozygous samples (Hom) are shown in dark gray.

each SNP. Figure 5A shows the distribution of r^2_{max} for 3230 SNPs from the filtered dataset that are 25 to 75% heterozygous to capture a range of difficulty for imputation. The median value of r^2_{max} was 0.4 to 0.5 for heterozygote frequencies below 0.5 but gradually decreased as the proportion of heterozygotes increased toward 0.75.

Cross-validation accuracy was determined with a training set of 100 hybrids selected at random from all hybrids with genotype data for a particular marker. The accuracy shown in Fig. 5B is the proportion of predicted values equal to the masked value. The results are binned by heterozygote frequency, with the median accuracy shown by a solid line and the first and third quartiles by dashed lines. Imputation with the population mode is a simple baseline method that, by definition, has lower accuracy as the frequency of the modal genotype declines. By contrast, the random forest method was largely unaffected by heterozygote frequency, with a median accuracy of ~0.85.

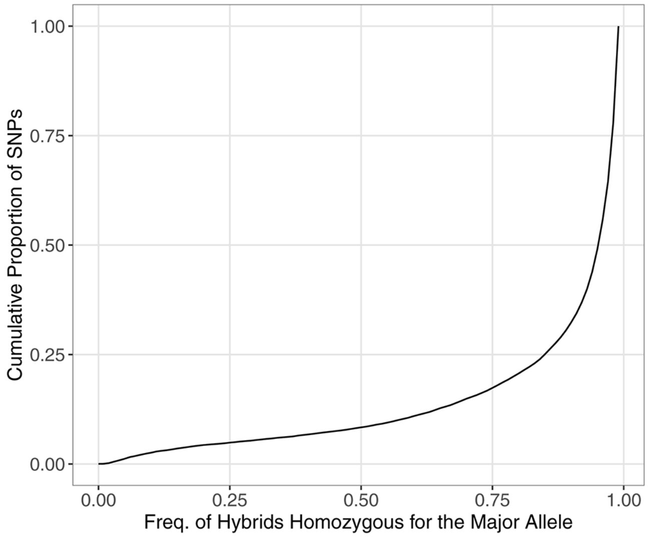

Fig. 4. Cumulative distribution for the frequency of *Urochloa* spp. hybrids homozygous for the major allele (HMA). The *y*-axis is the proportion of SNPs for which the HMA frequency is less than or equal to the *x*-axis value.

DISCUSSION

As mentioned in the introduction, there has been variation in the filtering criteria used in previous studies involving GBS of tetraploids. Uitdewilligen et al. (2013) recommended 60 to 80× for determining allele dosage, which corresponds to a EGQ_{tet} of 13 to 16. For diploidized genotype calling, the threshold of 11 reads from Li et al. (2014) is frequently used, which corresponds to EGQ = 13.7 in the absence of error but only EGQ = 11.1 when the allelic error is 0.5%. To achieve $EGQ_{dip} \geq 13$ with 0.5% allelic error, a threshold of 17 reads is needed.

The need for higher read depth per site to make accurate genotype calls in tetraploid species underscores the importance of selecting restriction enzymes to optimize the fragment size distribution. This study used *Ape*KI, which has a 5-bp recognition sequence, while Worthington et al. (2016) and Ferreira et al. (2019) used enzymes with 6-bp recognition sequences (*Hinc*II and *Nsi*I, respectively) for GBS of *Urochloa* spp. F_1 populations. Future research on GBS for *Urochloa* species should explore a two-enzyme system, such as the *Pst*I–*Msp*I combination introduced by Poland et al. (2012), as a way of generating more markers with higher read depth. Bastien et al. (2018) compared *Ape*KI against the *Pst*I–*Msp*I combination in tetraploid potato and obtained

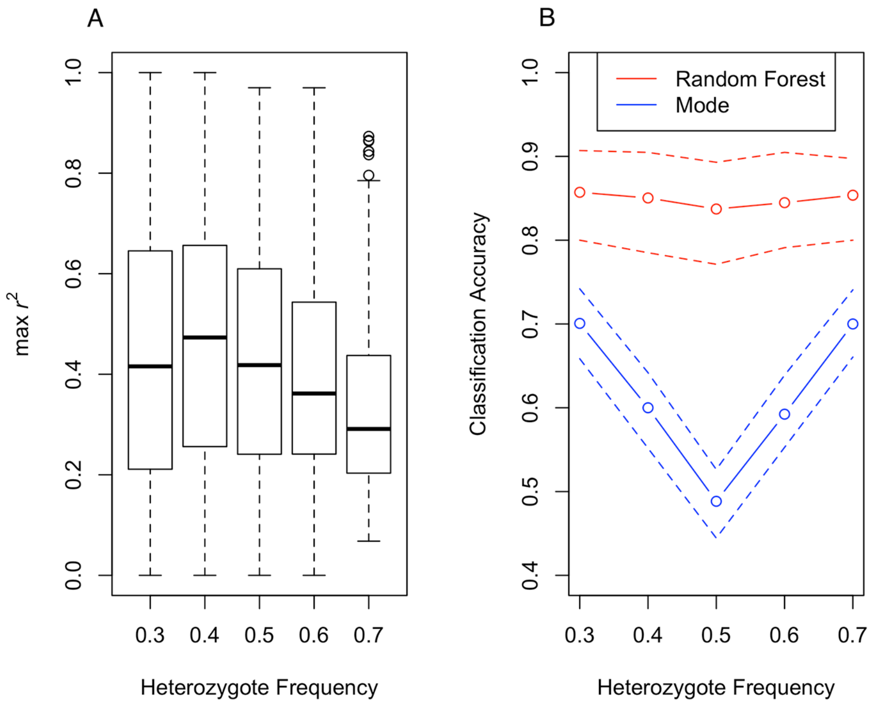

Fig. 5. (A) Distribution of the maximum LD (r^2) for 3230 SNPs with heterozygote frequency between 0.25 and 0.75 in the *Urochloa* spp. hybrids. The *x*-axis values are bin midpoints. (B) Imputation accuracy defined as the proportion of imputed values equal to the masked value.

tenfold more markers with the two-enzyme system when using a minimum sample depth of 53 reads.

The difference in EGQ for simplex (or triplex) vs. duplex genotypes has important implications for filtering GBS data. For diploidized genotype calls, setting a minimum GQ threshold creates bias against simplex (and triplex) samples relative to duplex samples; for tetraploid genotype calls, the opposite bias is present. Using a depth threshold based on the desired minimum EGQ does not introduce this bias, but filtering only on depth does not address reads with low base or mapping quality. Our conclusion is that a combination of the two approaches (depth and GQ) is needed. Supplemental File S1 can be used to calculate EGQ (Eq. [3]) for any depth and allelic error rate, and Supplemental File S2 can be used to generate matrices (marker × sample) of tetraploid or diploidized genotype calls and corresponding GQ scores from a VCF file.

The aforementioned considerations are appropriate for genotype calling based on the posterior mode. An alternative approach is to estimate allele dosage based on the posterior mean, which produces fractional genotype calls (Ashraf et al., 2014; Sverrisdoìttir et al., 2017; Clark et al., 2019). Such data are suitable when additive models are used in association analysis and genome-wide prediction, but a number of genetic analyses require integral estimates of dosage, including linkage analysis (Hackett et al., 2013; Zheng et al., 2016), dominance effects (Rosyara et al., 2016; Endelman et al., 2018), and haplotype inference (Su et al., 2008; Aguiar and Istrail, 2013).

This study used the traditional approach of setting hard thresholds for genotype calling followed by imputation of the missing data. We did not explore the interplay between threshold and imputation accuracy, but this is an interesting topic for future research. It seems appealing to select thresholds to achieve similar genotype accuracy in the samples called based on allele counts vs. those that are imputed. Ultimately, the traditional two-step approach (threshold then impute) is suboptimal because the read counts for the missing genotypes are not used during imputation. For ordered markers, this limitation can be overcome by using hidden Markov models with read counts as the emission states. This approach has been used in diploid mapping populations (Fragoso et al., 2015; Bilton et al., 2018) and can be extended to hidden Markov models that have been developed for SNP array markers in tetraploids (Hackett et al., 2013; Zheng et al., 2016). For unordered markers, alternative imputation methods need to be explored.

Author Contributions

JBE and FIM designed the study. SCLB and CBdoV crossed and developed the *Urochloa* population. KGXM performed the DNA extraction. STN and MFC built the mock reference genome. JBE, FIM, and STN analyzed the data and drafted the manuscript. RFN and MFC provided analytical expertise and edited the manuscript. JBE and RFN supervised the whole study. All authors read and approved the final version of the manuscript for publication.

Supplemental Information Available

Supplemental information is available with the online version of this manuscript as well as in Dryad at https://doi.org/10.5061/dryad.4j2c7h6.

Conflict of Interest

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENTS

Financial support was provided by the National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES), and the Brazilian Agricultural Research Corporation (EMBRAPA). Computing services were provided by the National Center for High Performance Processing in São Paulo (CENAPAD) and the Center for High Throughput Computing (CHTC) at UW-Madison. We thank Schuyler Smith for contributions to the bioinformatics pipeline and Sushan Ru for comments on the binomial model.

REFERENCES

Aguiar, D., and S. Istrail. 2013. Haplotype assembly in polyploid genomes and identical by descent shared tracts. Bioinformatics 29:i352–i360. doi:10.1093/bioinformatics/btt213

Ashraf, B.H., J. Jensen, T. Asp, and L.L. Janss. 2014. Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. Theor. Appl. Genet. 127:1331–1341. doi:10.1007/s00122-014-2300-4

Bastien, M., C. Boudhrioua, G. Fortin, and F. Belzile. 2018. Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. Genome 61:449–456. doi:10.1139/gen-2017-0236

Bennetzen, J.L., J. Schmutz, H. Wang, R. Percifield, J. Hawkins, A.C. Ponaroli, et al. 2012. Reference genome sequence of the model plant Setaria. Nat. Biotechnol. 30:555–561. doi:10.1038/nbt.2196

Bilton, T.P., M.R. Schofield, M.A. Black, D. Chagné, P.L. Wilcox, and K.G. Dodds. 2018. Accounting for errors in low coverage high-throughput sequencing data when constructing genetic maps using biparental outcrossed populations. Genetics 209:65–76. doi:10.1534/genetics.117.300627

Blischak, P.D., L.S. Kubatko, and A.D. Wolfe. 2018. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics 34:407–415. doi:10.1093/bioinformatics/btx587

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32. doi:10.1023/A:1010933404324

Broad Institute. 2016. Genome analysis toolkit: Understanding and adapting the generic hard-filtering recommendations. https://software.broadinstitute.org/gatk/documentation/article.php?id=6925 (accessed 3 Mar. 2018).

Clark, L.V., A.E. Lipka, and E.J. Sacks. 2019. polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. G3: Genes, Genomes, Genet. 9:663–673.

Danecek, P., 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, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158. doi:10.1093/bioinformatics/btr330

DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, A. McKenna, T.J. Fennell, A.M. Kernytsky, A.Y. Sivachenko, K. Cibulskis, S.B. Gabriel, D. Altshuler, and M.J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498. doi:10.1038/ng.806

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

Endelman, J.B., C.A. Schmitz Carley, P.C. Bethke, J.J. Coombs, M.E. Clough, W.L. da Silva, et al. 2018. Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. Genetics 209:77–87. doi:10.1534/genetics.118.300685

Felcher, K.J., J.J. Coombs, A.N. Massa, C.N. Hansey, J.P. Hamilton, R.E. Veilleux, C.B. Buell, and D.S. Douches. 2012. Integration of two diploid potato linkage maps with the potato genome sequence. PLoS One 7:e36347. doi:10.1371/journal.pone.0036347

Ferreira, R.C.U., L.A. de Castro Lara, L. Chari, S.C.L. Barrios, C.B. do Valle, J.R. Valerio, F.Z.V. Torres, A.A.F. Garcia, and A.P. de Souza. 2019. Genetic mapping with allele dosage information in tetraploid *Urochloa decumbens* (Stapf) R.D. Webster reveals insights into spittlebug (*Notozulia entreriana* Berg) resistance. Front. Plant Sci. 10:92. doi:10.3389/fpls.2019.00092

Fragoso, C.A., C. Heffelfinger, H. Zhao, and S.L. Dellaporta. 2015. Imputing genotypes in biallelic populations from low-coverage sequence data. Genetics 202:487–495. doi:10.1534/genetics.115.182071

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907v2.

Gerard, D., L.F.V. Ferrão, A.A.F. Garcia, and M. Stephens. 2018. Genotyping polyploids from messy sequencing data. Genetics 210:789–807. doi:10.1534/genetics.118.301468

Hackett, C.A., K. McLean, and G.J. Bryan. 2013. Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. PLoS One 8:e63939. doi:10.1371/journal.pone.0063939

Hickey, J.M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654–663. doi:10.2135/cropsci2011.07.0358

Jank, L., S.C. Barrios, C.B. do Valle, R.M. Simeão, and G.F. Alves. 2014. The value of improved pastures to Brazilian beef production. Crop Pasture Sci. 65:1132–1137. doi:10.1071/CP13319

Jank, L., C. Valle, and R. Resende. 2011. Breeding tropical forages. Crop Breed. Appl. Biotechnol. S1:27–34. doi:10.1590/S1984-70332011000500005

Jungmann, L., B.B.Z. Vigna, K.R. Boldrini, A.C.B. Sousa, C.B. do Valle, R.M.S. Resende, M.S. Pagliarini, M.I. Zucchi, and A.P. de Souza. 2010. Genetic diversity and population structure analysis of the tropical pasture grass *Brachiaria humidicola* based on microsatellites, cytogenetics, morphological traits, and geographical origin. Genome 53:698–709. doi:10.1139/G10-055

Koning-Boucoiran, C.F.S., G.D. Esselink, M. Vukosavljev, W.P.C. van't Westende, V.W. Gitonga, F.A. Krens, R.E. Voorrips, W.E. van de Weg, D. Schulz, T. Debener, C. Maliepaard, P. Arens, and M.J.M. Smulders. 2015. Using RNA-seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa* L.). Front. Plant Sci. 6:249. doi:10.3389/fpls.2015.00249

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359. doi:10.1038/nmeth.1923

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.

Li, X., Y. Wei, A. Acharya, Q. Jiang, J. Kang, and E.C. Brummer. 2014. A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. G3: Genes, Genomes, Genet. 4:1971–1979. doi:10.1534/g3.114.012245

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22. doi:10.1177/154405910408300516

Lutts, S., J. Ndikumana, and B.P. Louant. 1991. Fertility of *Brachiaria ruziziensis* in interspecific crosses with *Brachiaria decumbens* and *Brachiaria brizantha*: Meiotic behavior, pollen viability and seed set. Euphytica 57:267–274. doi:10.1007/BF00039673

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12. doi:10.14806/ej.17.1.200

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303. doi:10.1101/gr.107524.110

Melo, A.T.O., R. Bartaula, and I. Hale. 2016. GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC Bioinformatics 17:29. doi:10.1186/s12859-016-0879-y

Miles, J.W. 2007. Apomixis for cultivar development in tropical forage grasses. Crop Sci. 47:S238–S249. doi:10.2135/cropsci2007.04.0016IPBS

Miles, J.W., C. Cardona, and G. Sotelo. 2006. Recurrent selection in a synthetic Brachiariagrass population improves resistance to three spittlebug species. Crop Sci. 46:1088–1093. doi:10.2135/cropsci2005.06-0101

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, and G.Y. Zhong. 2015. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genet. 5:2383–2390. doi:10.1534/g3.115.021667

Monteiro, L.C., J.R. Verzignassi, S.C.L. Barrios, C.B. do Valle, G. de L. Benteo, and C.B. de Libório. 2016. Characterization and selection of interspecific hybrids of *Brachiaria decumbens* for seed production in Campo Grande-MS. Crop Breed. Appl. Biotechnol. 16:174–181.

Ouyang, S., W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R.L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, and C.R. Buell. 2006. The TIGR rice genome annotation resource: Improvements and new features. Nucleic Acids Res. 35:D883–D887. doi:10.1093/nar/gkl976

Pessoa-Filho, M., A.M. Martins, and M.E. Ferreira. 2017. Molecular dating of phylogenetic divergence between *Urochloa* species based on complete chloroplast genomes. BMC Genomics 18. doi:10.1186/s12864-017-3904-2

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253. doi:10.1371/journal.pone.0032253

R Development Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Rosyara, U.R., W.S. De Jong, D.S. Douches, and J.B. Endelman. 2016. Software for genome-wide association studies in autopolyploids and its application to potato. Plant Genome 9:1–10. doi:10.3835/plantgenome2015.08.0073

Rutkoski, J.E., J. Poland, J.L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes, Genomes, Genet. 3:427–439. doi:10.1534/g3.112.005363

Schmitz Carley, C.A., J.J. Coombs, D.S. Douches, P.C. Bethke, J.P. Palta, R.G. Novy, and J.B. Endelman. 2017. Automated tetraploid genotype calling by hierarchical clustering. Theor. Appl. Genet. 130:717–726. doi:10.1007/s00122-016-2845-5

Schnable, P.S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, et al. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112–1115. doi:10.1126/science.1178534

Serang, O., M. Mollinari, and A.A.F. Garcia. 2012. Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS One 7:e30906. doi:10.1371/journal.pone.0030906

Silva, P.I.T., A.M. Martins, E.G. Gouvea, M. Pessoa-Filho, and M.E. Ferreira. 2013. Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genomics 14. doi:10.1186/1471-2164-14-17

Su, S.Y., J. White, D.J. Balding, and L.J. Coin. 2008. Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. BMC Bioinformatics 9:513. doi:10.1186/1471-2105-9-513

Sverrisdoìttir, E., S. Byrne, H.E.R. Sundmark, H.Ø. Johnsen, H.G. Kirk, T. Asp, L. Janss, and K.L. Nielsen. 2017. Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. Theor. Appl. Genet. 130:2091–2108. doi:10.1007/s00122-017-2944-y

Swarts, K., H. Li, J.A.R. Navarro, D. An, M.C. Romay, S. Hearne, C. Acharya, J.C. Glaubitz, S. Mitchell, R.J. Elshire, E.S. Buckler, and P.J. Bradbury. 2014. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7. doi:10.3835/plantgenome2014.05.0023

Swenne, A., B.P. Louant, and M. Dujardin. 1981. Induction par la colchicine de formes autotétraploïdes chez *Brachiaria ruziziensis* Germain et Evrard (Graminée). Agron. Trop. 36:134–141.

Uitdewilligen, J.G.A.M.L., A.M.A. Wolters, B.B. D’hoop, T.J.A. Borm, R.G.F. Visser, and H.J. van Eck. 2013. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One 8:e62355. doi:10.1371/journal.pone.0062355

Vigna, B.B.Z., L. Jungmann, P.M. Francisco, M.I. Zucchi, C.B. do Valle, and A.P. de Souza. 2011. Genetic diversity and population structure of the *Brachiaria brizantha* germplasm. Trop. Plant Biol. 4:157–169. doi:10.1007/s12042-011-9078-1

Voorrips, R.E., G. Gort, and B. Vosman. 2011. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics 12. doi:10.1186/1471-2105-12-172

Worthington, M., C. Heffelfinger, D. Bernal, C. Quintero, Y.P. Zapata, J.G. Perez, J. De Vega, J. Miles, S. Dellaporta, and J. Tohme. 2016. A parthenogenesis gene candidate and evidence for segmental allopolyploidy in apomictic *Brachiaria decumbens*. Genetics 203:1117–1132. doi:10.1534/genetics.116.190314

Zheng, C., R.E. Voorrips, J. Jansen, C.A. Hackett, J. Ho, and M.C. Bink. 2016. Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. Genetics 203:119–131. doi:10.1534/genetics.115.185579 5579