pSBVB: A Versatile Simulation Tool To Evaluate Genomic Selection in Polyploid Species

María L. Zingaretti,*1 Amparo Monfort,*† and Miguel Pérez-Enciso*‡1

*Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Barcelona, Spain, †IRTA (Institut de Recerca i Tecnologia Agroalimentàries), Barcelona, Spain, and ‡ICREA, Passeig de Lluís Companys 23, 08010 Barcelona, Spain

ABSTRACT Genomic Selection (GS) is the procedure whereby molecular information is used to predict complex phenotypes and it is standard in many animal and plant breeding schemes. However, only a small number of studies have been reported in horticultural crops, and in polyploid species in particular. In this paper, we have developed a versatile forward simulation tool, called polyploid Sequence Based Virtual Breeding (pSBVB), to evaluate GS strategies in polyploids; pSBVB is an efficient gene dropping software that can simulate any number of complex phenotypes, allowing a very flexible modeling of phenotypes suited to polyploids. As input, it takes genotype data from the founder population, which can vary from single nucleotide polymorphisms (SNP) chips up to sequence, a list of causal variants for every trait and theirheritabilities, and the pedigree. Recombination rates between homeologous chromosomes can be specified, so that both allo- and autopolyploid species can be considered. The program outputs phenotype and genotype data for all individuals in the pedigree. Optionally, it can produce several genomic relationship matrices that consider exact or approximate genotype values. pSBVB can therefore be used to evaluate GS strategies in polyploid species (say varying SNP density, genetic architecture or population size, among other factors), or to optimize experimental designs for association studies. We illustrate pSBVB with SNP data from tetraploid potato and partial sequence data from octoploid strawberry, and we show that GS is a promising breeding strategy for polyploid species but that the actual advantage critically depends on the underlying genetic architecture. Source code, examples and a complete manual are freely available in GitHub https://github.com/lauzingaretti/pSBVB.

KEYWORDS
Genomic selection
Polyploids
Simulation
pSBVB
Genomic Prediction
GenPred
Shared Data Resources

Copyright © 2019 Zingaretti et al.
doi: https://doi.org/10.1534/g3.118.200942
Manuscript received October 10, 2018; accepted for publication December 19, 2018; published Early Online December 20, 2018.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1Corresponding authors: Centre for Research in Agricultural Genomics (CRAG) E-mail: laura.zingaretti@cragenomica.es; Centre for Research in Agricultural Genomics (CRAG), ICREA, Facultat de Veterinaria, 08193 Bellaterra, Barcelona, Spain. E-mail: miguel.perez@uab.es

Genomic selection (GS) (Meuwissen et al. 2001) is the breeding strategy consisting in predicting future performance using DNA information from the whole genome, typically SNPs (single nucleotide polymorphisms). It relies on genome wide linkage disequilibrium (LD) between markers and the causal mutations, without the need to identify them. Due to dramatic reduction in genotyping costs, GS is becoming standard in many animal and plant breeding schemes, replacing or complementing traditional methods based solely on pedigree information. So far, GS has been mainly applied to diploid species. Yet, polyploidy is a very common phenomenon in evolution and include numerous species of interest (e.g., strawberry, potato, wheat). Traditionally, polyploid species have been classified into autopolyploids, caused by one or more genome duplication events in a single species, and allopolyploids, the result of hybridization between closely related species (Stebbins 1947). The impact of GS on either auto- or allopolyploid species breeding, however, remains largely unexplored.

In principle, the application of GS in polyploid species can have a positive impact in the rates of genetic gain through improved accuracy of predicted breeding values and/or reduction of generation intervals (Slater et al. 2016; Bassi et al. 2016; Sverrisdóttir et al. 2017; Gezan et al. 2017; Enciso-Rodriguez et al. 2018). However, the complex genetic structure of polyploids has delayed the availability of genome-wide
genotyping SNP arrays that are needed for GS. Polyploid SNP detection can be challenging due to a high similarity between homologous and homeologous sequences, which generates complications to differentiate true SNPs from nuisance paralogous variants (Bassil et al. 2015; Cleve and Ozias-Akins 2015).

Further, accurate genotyping is also important but becomes more complex as ploidy level increases. Several tools to perform genotype estimation from SNP array platforms are already available (Voorrips et al. 2011; Voorrips and Gort 2018; Schmitz Carley et al. 2017; Blischak et al. 2017). However, the rising of Next Generation Sequencing technologies requires of new tools adapted for this type of data, which are also being developed (Bourke et al. 2018; Meirmans et al. 2018; Gerard et al. 2018).

Computer simulation is a fundamental tool to evaluate alternative breeding schemes, since it allows the exploration of a wide range of hypothesis at no cost and can help to interpret the outcome of selection in complex situations. In this regard, numerous simulation tools have been developed such as easyPOP (Balloux 2001), simuPOP (Peng and Kimmel 2005; Peng and Amos 2008), forqS (Kessner and Novembre 2014) Slim (Messner 2013), PedigreeSim (Voorrips and Maliepaard 2012) among others. However, simulation approaches may not be straightforward to interpret owing to unknowns on the genetic architecture, among other factors. These problems are exacerbated in polyploid species and, to the best of our knowledge, only simuPOP and PedigreeSim allow polyploids organisms. simuPOP is not developed to compare breeding schemes, whereas PedigreeSim does not directly generate phenotypes nor produces genomic relationship matrices.

Here we present a flexible simulation tool for complex phenotypes adapted to polyploids and we propose several approaches to compute the molecular relationship matrix in polyploids. The software is an extension of Sequence-Based Virtual Breeding (SBVB, Pérez-Enciso et al. 2017), called pSBVB. This tool employs complete or partial genome data as input and simulates new genomes by gene dropping. We illustrate the software with data from two economically important polyploid species: potato, an autopolyploid, and strawberry, an allopolyploid.

**METHODS**

**Polyploid sequence based virtual breeding (pSBVB)**

pSBVB is a modification of SBVB software (Pérez-Enciso et al. 2017) that allows simulating genotypes and phenotypes of an arbitrary genetic complexity in polyploids. Compared to SBVB designed for diploid organisms only, pSBVB enables simulating meiosis in autopolyploid or allopolyploid species (see below). It takes ploidy into account to generate the phenotypes and incorporates several options to compute the molecular relationship matrix that are pertinent to polyploids, as described below.

**Software algorithm**

As input, pSBVB needs genotypes in vcf format (https://samtools.github.io) or a text file with genotypes coded to 0 up to h (where h is the ploidy level). For diploids, the vcf genotype format is of the kind 0/0, 0/1, and 1/1 for the three possible genotypes in a biallelic SNP. The polyploid vcf format is an extension of the type 0/0/0/0, 0/0/0/1 and so on in the case of an unphased tetraploid genotypes. Phased genotypes are represented by vertical bars, (e.g., genotype 0|0|0|1) is different from 1|0|0|0). No missing values are allowed. Phased genotypes are needed in pSBVB to identify which chromosomes are passed to offspring. A number of accurate phasing algorithms for diploids are available such as beagle (Browning and Browning 2007) or minimap (Howie et al. 2012). For polyploids, several approaches are also developed (He et al. 2018; Shen et al. 2016), but their accuracy has not been completely validated and seems critically dependent on ploidy level. If phase is unknown, pSBVB randomly generates a phase configuration. Further, linkage disequilibrium can be obtained by generating an individual genome out of a random pedigree starting with the founders’ genotypes. To do that, pSBVB incorporates the option ‘EXPAND_BASEPOP’, which generates additional founders’ by randomly crossing the available ones and random breeding for a pre-specified number of generations (see SBVB manual, https://lauzingaretti.GitHub.io/pSBVB/). A list with QTNs (Quantitative Traits Nucleotides) positions, a list of SNP positions to be used for GS, a pedigree file and a parameter file are also necessary. The pedigree file is used to perform the gene dropping simulation, i.e., genotypes of the descendants along the pedigree are generated following Mendelian rules and a pre-specified pairwise rate between homologous and homeologous pairs; for autopolyploids, pairing is at random. While performing gene dropping, pSBVB stores only the recombination breakpoints, which results in an efficient algorithm to recover marker genotypes and phenotypes.

pSBVB is very flexible in terms of the genetic architectures; it can simulate any number of traits with their specific QTNs and allelic effects. QTNs effects can be specified in a file or sampled from gamma, normal, or uniform distributions. In contrast to SBVB, though, pSBVB does not allow for epistasis. The Figure 2 shows a general representation of the pSBVB software, as well as screen shots.

As output, pSBVB produces phenotype and marker data of individuals obtained from the pedigree-based gene-dropping procedure. In addition, pSBVB can also compute molecular relationship matrices G using predefined marker subsets (e.g., a genotyping array) or the whole sequence. For polyploids, G is computed by default from:

\[
G = \frac{(M - hp)(M - hp)^T}{hp(1-p)^T}
\]

where M is a n×m matrix with elements containing the number of copies of the alternative allele for f_{ih} individual (i = 1..n) and f_{ih}
SNP \( (j = 1 \ldots m) \), and \( p \) is a \( m \)-dimension vector with marker allele frequencies. Note that Equation 1 reduces to the standard formula in the case of diploidy \( (h = 2) \) (VanRaden 2008).

Assessing the genotype for polyploids can be inferred from fluorescence intensity in SNP arrays or from read count in sequence data (Bourke et al. 2018) but may not be as accurate as for diploid organisms, especially at high ploidy levels. If genotyping is not accurate, a simple alternative is assuming that only one full homozygous can be distinguished for the rest of genotypes, i.e., that a given marker allele behaves as fully dominant. To accommodate this, pSBVB allows computing a modified \( G \) where element \( m_{ij} \) is coded as 0 if all alleles are 0 and 1 otherwise. This is specified with the MIMIC_HAPLOID statement in the parameter file. The software also incorporates a `MIMIC_DIPLOID` option, which assumes only presence or absence of the alternative allele can be ascertained for genotype values higher than 2. In summary, the software is able to generate three \( G \) matrices:

- **Default option**: The true genotype, i.e., number of copies of the alternative allele, is known without error \( (G_Y) \). In this approach \( M \) (Equation 1) has elements varying between 0 to \( h \).
- **MIMIC_DIPLOID option**: Only 0, 1 and 2 or more copies of a given allele can be distinguished. In this case, all genotypes with values larger than 2 area assigned a value ‘2’, thus \( M \) (Equation 1) has elements ranging between 0 and 2 and ploidy is set to 2.
- **MIMIC_HAPLOID option**: It considers that only one full homozygous can be distinguished for the rest of genotypes, then \( M \) (Equation 1) has elements ranging between 0 and 1 and ploidy is set to 1.

### Modeling meiosis in polyploids

Autopolyploids species have polysomic inheritance where homologous and homeologous chromosomes are randomly paired during meiosis. In contrast, most of allopolyploids have disomic inheritance, resulting from preferential pairing between homologous chromosomes. However, there is a continuum between both extreme meiotic behaviors that can be modeled by preferential pairing factor \((\theta)\), which expresses the increased probability of pairing between homologous chromosomes (Bourke et al. 2017). In a generic case with \( \frac{h}{2} \) sub-genomes, where \( h \) is the ploidy level, there are \( \left( \begin{array}{c} h \\ 2 \end{array} \right) = \frac{h(h-1)}{2} \) possible recombination between all chromosomes. pSBVB allows modeling meiotic pairing via a recombination \( h \times h \) matrix.
where $\theta_i^j + \theta_i^j \forall i,j$ in Equation 2 represents the probability of pairing between $i$ and $j$ chromosomes, assuming chromosomes $(1, 2), (3, 4), (h - 1, h)$ are the homologous pairs.

For example, the matrix for a strict auto-tetraploid is:

$$R = \begin{pmatrix}
0 & \frac{1}{h-1} + \theta_{12} & \frac{1}{h-1} + \theta_{13} & \cdots & \frac{1}{h-1} + \theta_{1h} \\
\frac{1}{h-1} + \theta_{12} & 0 & \frac{1}{h-1} + \theta_{23} & \cdots & \frac{1}{h-1} + \theta_{2h} \\
\frac{1}{h-1} + \theta_{13} & \frac{1}{h-1} + \theta_{23} & 0 & \cdots & \frac{1}{h-1} + \theta_{3h} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{h-1} + \theta_{1h} & \frac{1}{h-1} + \theta_{2h} & \frac{1}{h-1} + \theta_{3h} & \cdots & 0
\end{pmatrix}$$ (2)

And for a strict allopolyploid would be:

$$R = \begin{pmatrix}
0 & 1/3 & 1/3 & 1/3 \\
1/3 & 0 & 1/3 & 1/3 \\
1/3 & 1/3 & 0 & 1/3 \\
1/3 & 1/3 & 1/3 & 0
\end{pmatrix}$$ (3)

Statistical model for Genomic prediction

There are currently numerous statistical methods that address the large p small n problem and use genome-wide markers to predict breeding values (e.g., de los Campos et al. 2009, Hayes et al. 2009). pSBVB does not compute genomic breeding values but can produce genomic relationship matrices suitable to obtain GBLUP (VanRaden 2008), as detailed above. Otherwise, pSBVB outputs genotypes of all or a subset of markers and any desired GS algorithm can be applied. R scripts are provided in GitHub that performs GBLUP.

DATA AVAILABILITY

The source codes and the documented functions are distributed from GitHub: https://github.com/lauzingaretti/pSBVB. The manual includes a full tutorial of all functions at the program and a user guide with the installation guidelines and examples to simulate polyploid organisms. The software is accompanied by R scripts (R Core Team 2017) to generate a pedigree file, create the numerator relationship matrix, perform GBLUP (VanRaden 2008) or assess predictive ability (PA). Examples showing the software capabilities with alternative parameter options are also available.

RESULTS

In order to illustrate the software capabilities, we have used dataset from two polyploids species: autotetraploid potato (Solanum tuberosum, $2n = 4x = 48$) and allopolyploid strawberry (Fragaria × ananassa $2n = 8x = 56$).

Potato genotypes

The availability of an 8,300 SNP array has allowed the development of GS studies in potato, one of the most important crops worldwide (e.g., Sverrisdóttir et al. 2017, Enciso-Rodriguez et al. 2018). To illustrate our tool, here we used a subset of 407 SNPs and 150 individuals from Enciso-Rodriguez et al. (2018). SNP positions were obtained from Rosyara et al. (2016). We used these genotypes to generate a vcf file where genotypes were coded between 0 and 4 (the potato ploidy level), phases were randomly generated.

Next, to generate linkage disequilibrium in the randomly phased dataset, we included additional dummy founders using the “EXPAND_BASEPOP” statement in the parameter file (see reference manual, https://lauzingaretti.github.io/pSBVB/). With this option, new base population individuals are obtained via randomly generated pedigrees. A new base population with 100 founders was obtained. The total pedigree size was 700, with 250 founders was obtained. The total pedigree size was 700, including 250 founders (150 initial individuals and the 100 new base population individuals) and four generations with 100, 100, 100 and 150 observations, respectively.

Phenotypes were simulated using 140 randomly chosen QTNs and heritability ($h^2$) was set to 0.5. As numerous studies suggest that allele distribution is highly leptokurtic (Garcia-Dorado et al. 1998; Eyre-Walker and Keightley 2007) with many near-zero effects and a few large effects, we used a gamma $\Gamma(\alpha = 0.2, \beta = 5)$ distribution to simulate additive effects as in Caballero et al. (2015). G matrix was computed assuming that all markers are known without error, since the potato chip ensures that the true genotype can be obtained. Finally, to illustrate GS performance, which was assessed removing the 150 individuals from the last generation and computing the correlation between predicted and observed phenotypes of these 150 individuals. Figure 3 plots the observed vs. predicted phenotypes in training (400 individuals) and test (150 individuals) population. In this example, PA was reasonably high ($\rho = 0.52$), and illustrates that reasonable accuracies can be obtained.
even with small population sizes provided linkage disequilibrium and $h^2$ are relatively high.

The pedigree and the numerator relationship matrix files were generated using the pedigree.R and RelationshipMatrix.R functions, respectively; breeding values were predicted with GBLUP using GBlupFunction.R script. The whole source code and scripts to run this example are available at GitHub site.

**Application to strawberry GBS data**

We also applied our program to octoploid strawberry *F. x ananassa*. In the absence of a reasonable number of strawberry sequenced genomes, we used unpublished data obtained with GBS (Genotyping by Sequencing) from 47 strawberry cultivars. Genotype-by-Sequencing libraries were prepared by Heartland Plant Innovations (http://www.heartlandinnovations.com/). Samples were multiplexed and sequenced 92 cycles on the Illumina MiSeq at the Oklahoma Medical Research Foundation. Data quality was checked by FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To obtain reasonably realistic genotypes based on these data, we applied the following pipeline. GBS reads were aligned against *Fragaria vesca* (diploid strawberry) reference genome (F. vesca-genome.v2.0.a1), bam files were filtered setting minimum base and mapping qualities to 37 and 20, respectively, and parsed with snapE (https://github.com/EmanueleRaineri/snapE-pooled, Raineri *et al.* 2012), a SNP caller developed for pools.

This software requires as input the number of diploid individuals in the pool, which was set to four. Polymorphic positions with fewer than 20 high quality reads were removed, as well as those where more than 60% of the cultivars were not covered. Logically, only allele counts 0, 1, to 8 are allowed in an octoploid genome SNP, whereas the number of reads per position follows a quasi-continuous distribution. To convert number of reads to genotype score, we computed the fraction of alternative allele reads divided by the total number of reads ($f$) and inferred its genotype from the nearest possible integer to $f \times 8$. This was done for each SNP and cultivar. Missing genotypes were sampled according to the genotype frequency in the non-missing positions for that SNP. We assumed independence to perform the assignments. A total of 50,609 variant positions were obtained (5779, 7985, 7328, 6362, 8282, 9012, 5862 in linkage groups GL1, GL2, GL3, GL4, GL5, GL6 and GL7, respectively). These markers were used as genetic file input for the program. Among those SNPs, ~36%, 37%, 14% and 13% variants were classified as segregating in 1, 2, 3 or all sub-genomes: 2x, 4x, 6x and 8x, respectively.

Strawberry breeding programs are based on evaluating crosses between elite lines. Traditional crop breeding is expensive and time consuming and GS can accelerate strawberry improvement if only a subset of these crosses were fully tested in the field. To mimic this scenario, we generated a pedigree file with five generations of intercrossing starting with the 53 base population lines. Each generation was made up of 100 lines. In the last generation, 1000 crosses with unknown phenotype were generated from the 100 current parental lines. As measure of predictive accuracy, we computed the correlation between observed and predicted phenotypes of the 1000 crosses, when the phenotypes from these 1000 crosses were removed. One hundred replicates were run per case.

To simulate the phenotypes, we considered a range of genetic architectures with a focus on sugar content:

- Random QTNs in sugar associated Pathways (RQP): 100 SNPs were randomly chosen as causal among the SNPs in the sugar pathway associated genes ≤ 10 kb.
- Diploid QTNs in sugar associated Pathways (DQP): 100 SNPs were randomly chosen as causal among the diploid SNPs in the sugar pathway associated genes ≤ 10 kb.
- Random QTNs Genome-wide chosen (RQG): 100 SNPs were randomly chosen as causal among all detected SNPs.

In the first two architectures, we aimed at mimicking a trait of economic interest such as sucrose content. The gene information was...
obtained from FragariaCyc (http://pathways.cgrb.oregonstate.edu, Naithani et al. 2016). In total, there were 159 genes containing 499 SNPs associated with these pathways. Within each of the three architectures, phenotypes were simulated according to two extreme gene actions: fully additive and complete dominance ($\phi = 1$, Figure 1). Heritability was set to 0.5.

For each architecture, phenotypes were simulated according to two extreme gene actions: fully additive and complete dominance. In the dominant approach, we set $\Gamma(a = 0.2, \beta = 5)$ (Figure 1). Each phenotype was generated from its genotypic value adding an environmental effect, where was adjusted such that heritability was $h^2 = 0.5$.

Simulated PAs are in Figure 4. We estimated the PA using the following matrices:

- **GT**: The true genotype, i.e., number of copies of the alternative allele, was known without error and all SNPs were used. In this approach Equation $M$ (Equation 1) has elements varying between 0 and 8.
- **G2**: Only diploid SNPs were used, and genotypes were known without error. $M$ (Equation 1) has elements ranging between 0 and 2.
- **G2**: All SNPs were employed but only genotypes of diploid SNPs were known without error, whereas for the remaining, although the organism was polyploid, Genomic matrix is computed using diplodiploid. $M$ (Equation 1) has elements ranging between 0 and 2.
- **Numerator Relationship Matrix (P-BLUP)**: The breeding values were predicted using the pedigree relationship matrix.

Figure 4 shows the obtained accuracies across genetic architectures and for each evaluation method. Overall, these results indicate that performance of GS in polyploids may critically depend on the underlying genetic architecture. Unsurprisingly, accuracy also drops when dominance exists compared to the additive scenarios. Several additional observations of interest can be drawn from Figure 4. First, there were no differences in the ranking of methods irrespective of whether QTNs were scattered throughout the genome (RQG) or localized in given segments (RQP). This was observed for both additive and dominant architectures. Second, using the true genotype values to build $G$ (GT) did not always outperform the rest of GBLUP methods considered. In fact, this was observed only when the architecture was fully additive and the QTNs were segregating in more than one homeolog group. In these cases, GT-BLUP was ~4–8% better than G2-BLUP or G2–BLUP. G2, which employs only diploid SNPs, should be preferred to GT-BLUP only if QTNs are exclusively diploid. A relevant result is that G2–BLUP, which treats markers as dominant, was a quite robust strategy, in particular with complete dominance and with the exception of DQP scenario (i.e., when all QTNs were diploid).

Finally, note that the advantage of GBLUP over P-BLUP is not always guaranteed. At least in the breeding scenario analyzed here, G2-BLUP might actually perform worse than P-BLUP when QTNs segregate randomly (RQP and RQG) and genic action is additive. If true SNP genotypes could be known without error (GT), the increase in accuracy compared to P-BLUP would vary between ~7% and 18%. As for using G2–BLUP, increase in accuracy was between ~3% and ~16% across all cases examined here. The advantage, though, would diminish if genic action were additive and QTN would segregate in more than one homeolog group. In these cases, GT-BLUP might actually perform worse than G2-BLUP or G2–BLUP.

The genetic file used as input includes 1500 SNPs from the whole vcf file. More examples combining a set of different parameters (additive and dominance effects, Genetic Matrix calculation, pedigree and Genomic Relationship Generation, among others) are available on GitHub.

**DISCUSSION**

Certainly, polyploid sequence data will be increasingly available, which will be used to achieve a better understanding of complex trait genetics and to optimize GS strategies. To help in the latter task, here we have developed an extension of SBVB software (pSBVB) that feeds from real sequence data of polyploid organisms. It uses efficient forward algorithms and allows simulating meiosis in polyploid species, suited for both auto and allopolyploid organisms. Further, pSBVB generalizes

---

**Figure 4** Predictive Ability ($PA = cor(y, \hat{y})$) of GBLUP and P–BLUP models for each of the three genetic architectures considered in strawberry simulated dataset: random QTNs in sugar associated pathways (RQP), diploid QTNs in sugar associated pathways (DQP) and genome-wide chosen (RQG) and each of the three GBLUP models. Three GBLUP models were compared: In GT, genetic matrix $G$ was computed assuming SNP allele frequencies were known without error; in G2, only diploid SNPs were used, and genotypes were known without error; and in G2, $G$ Genomic relationship matrix is computed assuming than only presence or absence of the alternative allele could be known for the remaining, i.e., although the organism was polyploid, Genomic relationship matrix is computed assuming than only presence or absence of the alternative allele can be ascertained. (a) additive architecture; (b) dominant architecture.
genetic modeling in polyploids to generate phenotypes and incorporates several options to compute predefined molecular relationship matrices that are specific to polyploid organisms. Note though that, since pSBVB can print the whole SNP dataset, any custom-made G can be computed and any alternative GS method can be evaluated. There are some limitations though. An important one is that epistasis cannot be modeled like the one described here. Among the available forward-time simulation tools, only simuPOP (Peng and Kimmel 2005; Peng and Amos 2008) and PedigreeSim (Voorrips and Maliepaard 2012) consider polyploids. Compared to simuPOP, pSBVB allows simulating both auto and allo-polyploids organisms, accepting as input a recombination phase and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81: 1084–1097. https://doi.org/10.1016/j.ajhg.2017.05.006


R Core Team, 2017 R: A Language and Environment for Statistical Computing.


Communicating editor: G. de los Campos