Annals of Agri-Bio Research **28** (2): 466-471, 2023 **Impact Score : 0.28 (Scopus)**

# Comparison of Softwares for Molecular Population Genetics Using *Oryza sativa* Varieties

BHARTI GUPTA, NEERU REDHU*, JYOTI TAUNK[1], SHIKHA YASHVEER[2] AND SUNITA JAIN

*Department of Bioinformatics and Computational Biology, CCS Haryana Agricultural University, Hisar-125 004 (Haryana), India*
**(e-mail: redhuneeru95@hau.ac.in; Mobile: 94668 33480)*

(Received: February 27, 2023; Accepted: March 25, 2023)

**ABSTRACT**

Numerous programs for studying population genetics using molecular data have been developed. The present study compared the features of such freely available softwares viz., STRUCTURE, Arlequin, PopGene, PowerMarker and Winboot. In this comparative study, the molecular data from 28 rice varieties at 42 loci obtained using simple sequence repeat markers were used. Even though the results produced by all the softwares were found to be in corollary, but there were marked differences in their attributes as analysis each software offered, its graphical interface, type of data it supported and method it was based on. The study demonstrated that among all the studied softwares, PowerMarker was best for analyzing genetic relationship using codominant genotypic data. STRUCTURE was best for finding genetic structure, whereas Arlequin was appropriate for haplotypic data. PopGene should be used only with closely related species represented by codominant data. Winboot should also be used with codominant data only.

**Key words:** Arlequin, PopGene, PowerMarker, STRUCTURE, Winboot

## INTRODUCTION

Molecular markers are specific DNA fragments within the genome which can be utilized by scientists to investigate the allele origins, their genetic diversity and structure in population as well as to describe genetic relationships among species/populations. By testing marker loci variation, the populations could be classified genetically leading to better understanding of changes during the course of evolution in their improvement (Flanagan and Jones, 2019). Numerous molecular markers like simple sequence repeats (SSRs), restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs) and amplified fragment length polymorphisms (AFLPs), etc. that could be used in genetic diversity studies.

Genetic diversity could be assessed by analyzing the variation at marker loci using an explicit technique or an amalgamation of techniques. However, correct analysis depends on the selection of marker loci with high polymorphic information content (PIC) value. PIC value includes two distinct components (a) richness i.e. more the number of alleles, and (b) evenness i.e. more evenly the alleles are distributed (Salem and Sallam, 2016).

To find the genetic structure, clustering methods were broadly divided into two types i.e., distance-based methods and model-based methods. Former calculated the distance among every pair of entities generating a pairwise distance matrix (Hua *et al.*, 2017). Methods used for graphical representation of distance matrix were unweighted pair group method with arithmetic mean (UPGMA), neighbour-joining (NJ), principal component analysis (PCA), principal coordinate analysis (PCoA) and analysis of molecular variance (AMOVA). Model based methods presume that observations from each cluster were randomly drawn from certain parametric model. The inference for the parameters corresponding to each cluster was performed in combination with inference for each individual in cluster, by regular statistics like maximum likelihood

[1]Department of Biotechnology, University Centre for Research and Development, Chandigarh University, Mohali-140 413 (Punjab), India.
[2]Department of Molecular Biology, Biotechnology and Bioinformatics, CCS Haryana Agricultural University, Hisar-125 004 (Haryana), India.

or bayesian schemes (Zaharias and Warnow, 2022).

Different softwares may be based on different methods: distance method or model method or may be specialized for specific datatype like dominant or codominant and haplotypic or genotypic data and may also differ in the range of analyses they provide. For example, PowerMarker, based on distance method, provided a large range of statistical analysis, Arlequin was used specifically for haplotypic data, PopGene and Winboot provided only UPGMA tree, while STRUCTURE used a model-based approach (Perez *et al.*, 2015; Dearfield *et al.*, 2017; Domínguez *et al.*, 2017; Meirmans *et al.*, 2018; Zhang *et al.*, 2021).

Availability of such a large number of softwares each with some or other special feature leaves user with uncertainty of choosing the correct software for available dataset. Therefore, this study compares the current features of above-mentioned softwares for molecular marker data analysis, including the operating system they run on, data forms they process and assays they execute, so that these freely available softwares could be used optimally to obtain the correct inference from a given molecular dataset.

## MATERIALS AND METHODS

From the large number of softwares available freely, five were chosen for the present study as these softwares could study diverse molecular markers through varied kinds of evaluations (Table 1). Data of 42 SSR markers from 28 rice varieties (Fig. 4) were used to compare the results of chosen software. Data were analyzed using all the five softwares.

**STRUCTURE ver. 2.2:** The data in plain text format with individuals in rows and loci in column were used. The data were analyzed for 10000, 1000000, 10000000 iterations after burn-in length of 10000. 1000000 and 10000000 iterations after burn-in length of 1000000 and also 10000000 iterations after burn-in length of 10000000 for population number 3, 4 and 5 each.

**Arlequin:** Standard indices like gene diversity and effective allele number, AMOVA analysis, population differentiation and population distance were computed by selecting specific nodes on tree.

**Winboot:** The data file was prepared in plain text format with each individual in one row and each allele in column. Allele presence/absence was marked as 1/0. The computation was performed using default desired distance coefficient and bootstrap number.

**PopGene:** Plain text input file where alleles were coded as alphabets were analyzed for diploid codominant data. All loci and population were kept for the analysis.

**PowerMarker ver. 3.25:** Dataset was prepared in plain text format with data for individuals in rows and loci in column. A variety of analyses were performed.

## RESULTS AND DISCUSSION

The present study compared the main features of five freely available softwares viz., STRUCTURE, Arlequin, Winboot, PopGene and PowerMarker for molecular marker data analysis.

Using Bayesian clustering approach, STRUCTURE assigned individuals to populations so that population groupings obtained were not in disequilibrium. Result of run for 1,00,00,000 iterations after the burn-in length of 1,00,00,000 for four populations was chosen on the basis of value of $\alpha$ (allele frequency distribution considering admixture model) which converged well and highest value of Inprob(X|K) (posterior probability of proportion of genotype × originating from population K) i.e. -2045.5. STRUCTURE clustered the varieties Basmati (Bas) 370, Taraori Basmati, Haryana Basmati (HBC) 19, Dehradun Basmati Type (Type) 111, Super, CSR 30, Kernel Basmati (Kernel) and Haryana Kaul Rice (HKR) 94-416 in one population i.e. Basmati; varieties Ranbir Basmati (RanBas) and Kasturi in between Basmati and *Japonica*, varieties; Pusa Basmati (PB)1, Basmati (Bas)

**Table 1.** Softwares used in the present study

| Softwares | Operating system | URL |
|---|---|---|
| Popgene | Windows | *http://www.ualberta.ca/\|fyeh/index.htm* |
| Arlequin | Windows | *http://anthropologie.unige.ch/arlequin* |
| PowerMarker | Windows | *http://www.powermarker.net* |
| STRUCTURE | Windows, Linux | *http://pritch.bsd.uchicago.edu/software* |
| Winboot | Windows | *http://www.irri.org/science/software/winboot.asp* |

217, Nippon bare (Nippon), New Plant Type (NPT) 11, Azucena (Azu) and Della in one population i.e. *Japonica,* while varieties Sharbati, Sabarmati, Indica Rice (IR)70423, Improved (Imp) Sabarmati in one population i.e. cross-bred and varieties Pokkali, HKR 120, IR 72, IR 36, CSR 10, IR 24 and IR 64 were clustered together as *Indica* (Fig. 1A). The divergence among population was found to be less than the divergence estimated within the population (Fig. 1B).

Statistically, STRUCTURE was defined as the most robust method. It also dealt very well with admixture population, linked markers, dominant markers and null alleles. Being model-based method, it allowed incorporation of prior information and predicted the migrant history of population. On the other hand, large numbers of iterations had to be done to obtain the correct results, which consumed lot of time and computing power. Also, the results had to be obtained for different population number to get converged $\alpha$ and maximum probability, therefore, some prior information regarding population structure was required, especially when population was large. Also, it did not provide PIC value, so to choose subset of markers with high polymorphism user had to obtain PIC value using some other source. STRUCTURE only clustered species into different clusters; it did not provide relationship between each species. Also, its dealing with the dominant data was quite doubtful.

Arlequin provided results for large set of basic methods and statistical tests and created a result directory with the extension *.res. This directory contained all the result files, including a table of standard indices with expected heterozygosity i.e. the probability that two randomly chosen haplotypes, allelic range, allele number i.e. polymorphism shown by each locus, G-W stat i.e. a small value of G-W stat proposed that the population to be going through bottleneck (Fig. 2).

For AMOVA analysis, genetic structure definition was required by Arlequin. In spite of defining the structure based on geographic distribution or other assumptions, results obtained from software "STRUCTURE ver.2.2" were used as input. AMOVA hierarchical model employed "within population" ($\sigma_c$), "Among population/within group" ($\sigma_b$), "Among groups" ($\sigma_a$) components of diversity. The AMOVA outcomes depicted that most of the genetic diversity attributed to $\sigma_b$ (67.02) i.e. diversity found among populations within the group was maximum, the appreciable amount 30.65 still separated regions, however, the difference within population was small 2.33. Null distribution of variance components at each level i.e. $\sigma_b$, $\sigma_a$, $\sigma_c$ also predicted the same structure as no other value was larger or even nearby the input structure.

Exact test of population differentiation, through testing of the null hypothesis for random genotypes dispersal within populations, also predicted populations as significantly differentiated on the basis of P-value which was smaller than the significance level.

Arlequin was one of the pioneer softwares for genetic diversity assessment. It provided a very large set of statistical tests including computation of heterozygosity, gene diversity, polymorphism, neutrality tests, population differentiation, population comparison, AMOVA,

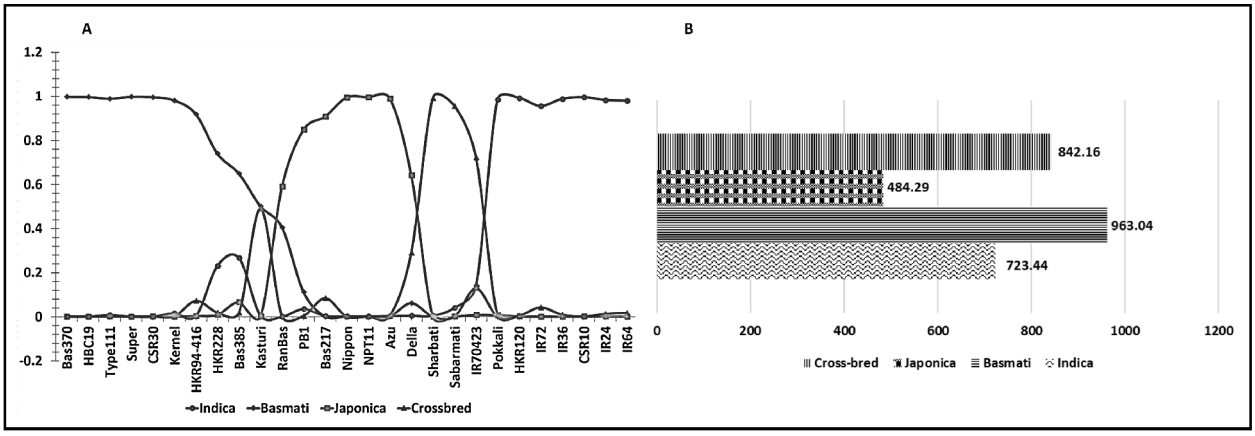


Fig. 1. (A) STRUCTURE assignment of individuals to distinct populations and (B) Divergence within population.

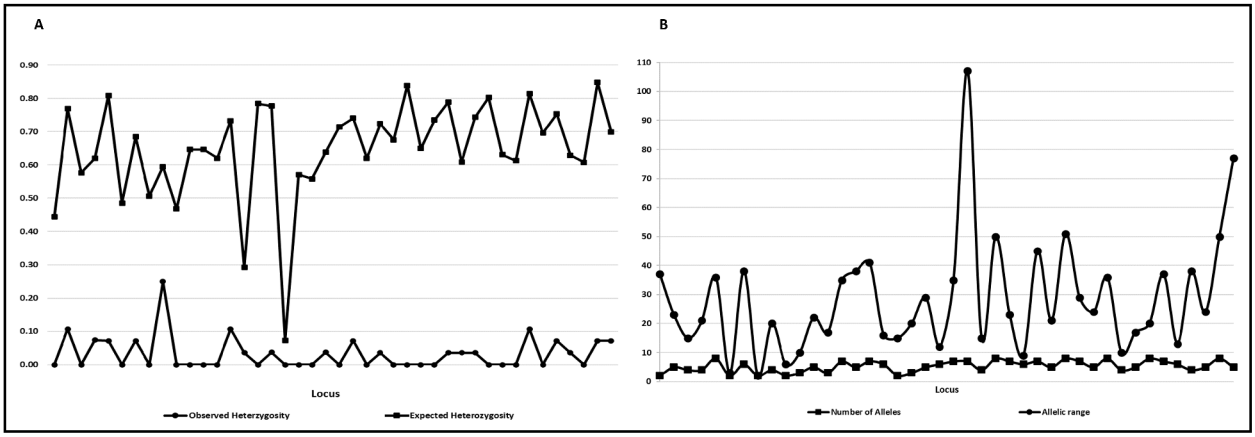


Fig. 2. Standard indices estimated by Arlequin for 42 loci (A) Observed and expected heterozygosity and (B) Number of alleles and allelic range.

Hardy-Weinberg equilibrium and Mantel test. In spite of such a large variety of tests, Arlequin was designed for haplotypic data, therefore, was rarely used with genotypic, some of its application like neutrality tests and minimum spanning tree were available for haplotypic data only. Application of Arlequin was also complex and cumbersome for e.g. Mantel test provided by Arlequin found correlation only between its own estimated $F_{ST}$ matrix and distance matrix provided by user in input file and also for AMOVA analysis it required some prior information regarding genetic structure, without providing genetic structure AMOVA did not run.

Winboot estimated genetic relationship between populations on distance-based method. It provided variety of genetic distance measure and graphically represented the matrix using UPGMA method. UPGMA tree obtained for the Rogers' distance measure after 100 bootstraps indicated Bas370, Type111, Super, HBC19, CSR30, HKR94-416, Kernel, HKR228, Bas 385, Kasturi, Della as more closely linked to each other, while Pokkali, CSR10, Sabarmati, Imp. Sabarmati, Sharbati, IR70423, HKR120, IR72, IR36, IR64 and IR24 more linked to each other. Bas217, Nippon, NPT11, Azu, RanBas and PB1 were found to be close to each other (Fig. 3A). Winboot also provided information regarding the number of times a grouping has been observed by giving the respective number at the node.

Winboot constructed UPGMA tree and obtained confidence interval by bootstrapping. But it took binary data in input file due to which its dealing with null alleles was doubtful. It provided no information regarding the PIC value, heterozygosity, Hardy-Weinberg equilibrium, F-statistics, etc. and availability of UPGMA tree only allowed analysis of only closely related species.

PopGene gave results for large number of tests. It found that none of the loci was in Hardy-Weinberg equilibrium as calculated by likelihood and chi-square test. Results for expected and observed heterozygosity and homozygosity also predicted very less heterozygosity for all loci. F-statistics estimation predicted inbreeding for large number of loci and near about zero gene flow. UPGMA tree predicted nearly same genetic relationship as by Winboot (Fig. 3).

PopGene was simple and easy to apply. It provided a large set of statistical tests like polymorphism, gene diversity, gene flow, Hardy-Weinberg equilibrium, neutrality tests, genetic distance and dendrogram. But it provided only Nei's genetic distance and UPGMA tree, which made it suitable for analysis of closely related species using codominant data only.

Summary statistics table provided by PowerMarker gave information regarding the gene diversity, heterozygosity and PIC value. It also provided maximum likelihood estimates of allele frequency and genotype frequency. As predicted by Arlequin and PopGene, PowerMarker also found that none of the loci was in Hardy-Weinberg equilibrium. The genetic distances estimated by PowerMarker among the regions (given by structure) were also found to be in agreement with results obtained by STRUCTURE (Table 2). The F-statistics results produced by PowerMarker were also in corollary with results produced by PopGene showing in-breeding between sub-

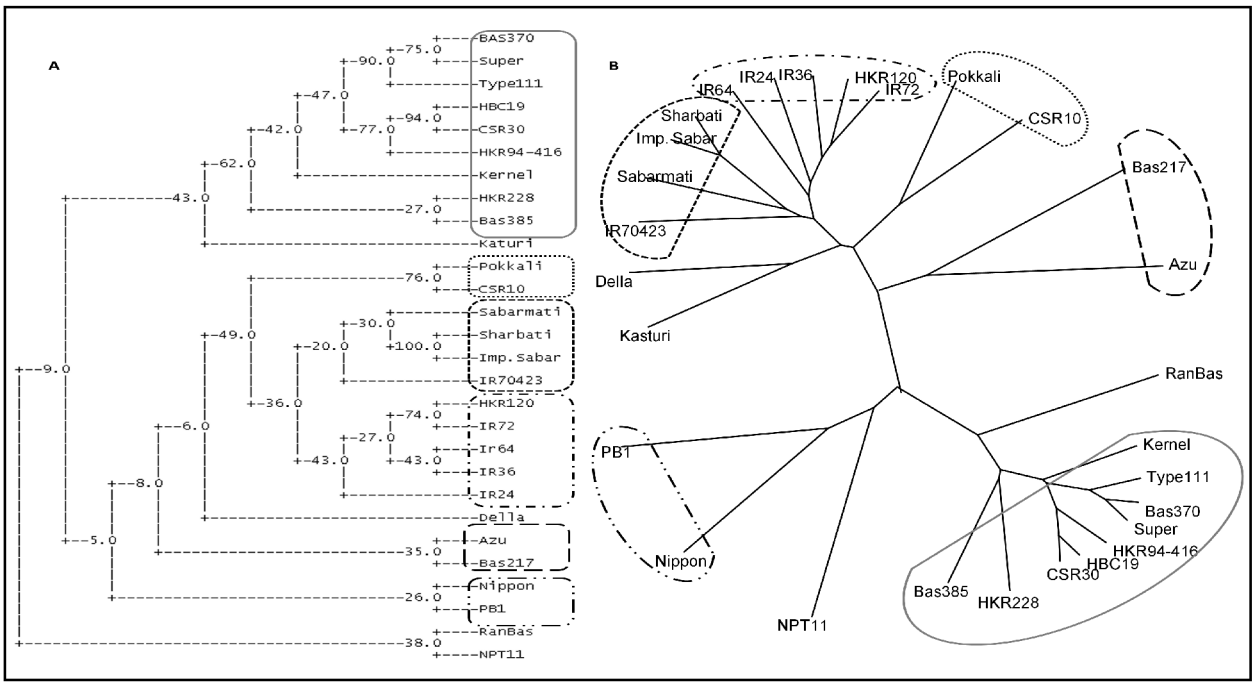


Fig. 3. (A) UPGMA dendrogram constructed for Rogers' distance by Winboot. The numbers at the forks show the % of times the group consisting of the species which were to the right of that fork occurred and (B) Genetic relationship estimated by PopGene using genetic distance and UPGMA method. Radial tree was generated by Treeview.

populations. The tree constructed using UPGMA (Fig. 4) method considering the genetic distance produced the same result as by PopGene. But the software clearly demarcated the difference between UPGMA and NJ (Fig. 4) method. As it was well known that UPGMA produced better result than NJ for closely related species (Varón-González *et al.*, 2020), the present data also strengthened the fact.

PowerMarker has a very good graphical interface. Tests were done step by step and results could be seen in PowerMarker or Excel. It provided large variety of statistical tests including PIC value, heterozygosity, Hardy-Weinberg equilibrium, F-statistics, number of genetic distance measure (e.g., Jacard, Nei's, Rogers', etc.), UPGMA or NJ tree, Mantel test and obtained confidence level by bootstrapping. Number of different genetic distance measure allowed analysis of data based on different assumptions regarding origin of species. Availability of both UPGMA and NJ tree allowed analysis of both closely related as well as distant species. Mantel test found correlation between two or more distance matrices, and helped user to infer whether different matrices gave same result. But PowerMarker was based on distance method, where result depends heavily on type of distance measure and graphical representation chosen. Though PowerMarker allowed user to attach information regarding marker or species in separate table and used this information to choose subset, but this extra information could not be used in analysis, as it was used in model-based method. One of the major shortcomings of PowerMarker was that it analyzed codominant markers only.

Even though the results obtained from all the softwares when applied to chosen data were in similar lines, but software study revealed marked differences in their attributes like analysis each software offered, their graphical interface, types of data they supported and methods software were based on. One major shortcoming present in all the softwares discussed above was their non-availability of 2-3 D scatter plots i.e. PCA or PCoA, which could be obtained by NTSYS-pc. This availability of

**Table 2.** Genetic distance estimated by PowerMarker between the regions which were estimated by STRUCTURE

| OTU | Basmati | Cross-bred | Indica | Japonica |
|---|---|---|---|---|
| Basmati | 0.0000 | 1.0513 | 0.7907 | 0.6200 |
| Cross-bred | 1.0513 | 0.0000 | 0.3020 | 0.7772 |
| Indica | 0.7907 | 0.3020 | 0.0000 | 0.5768 |
| Japonica | 0.6200 | 0.7772 | 0.5768 | 0.0000 |

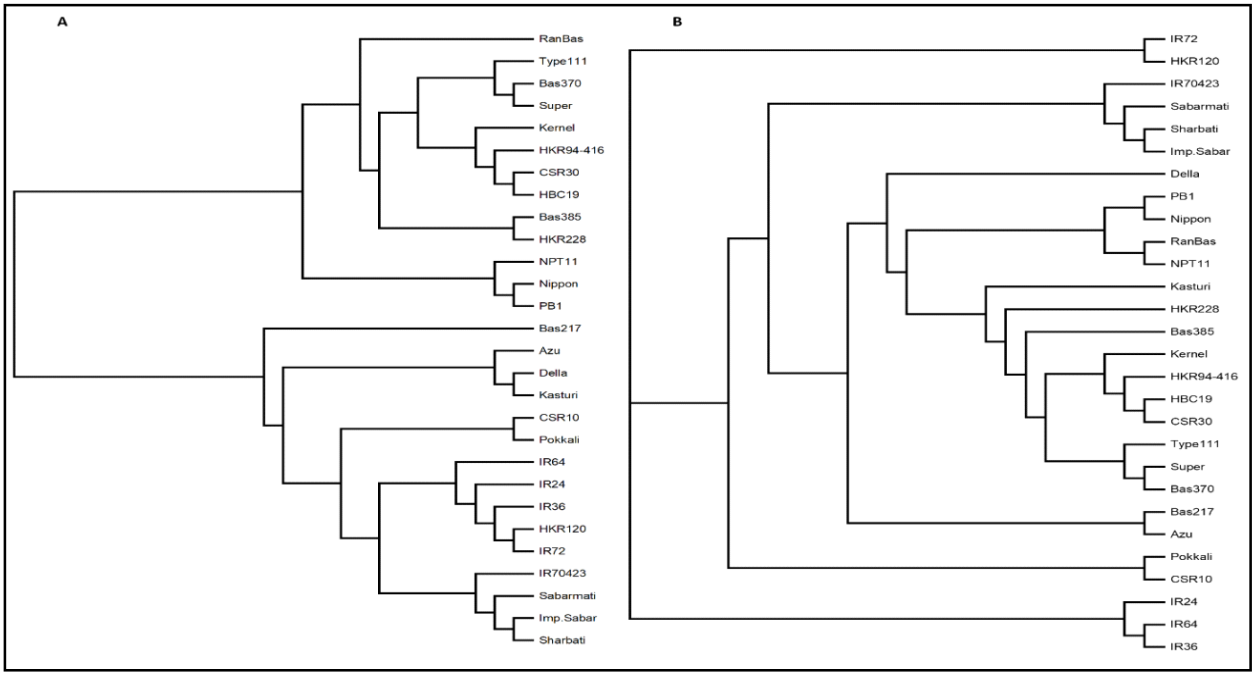


Fig. 4. Tree constructed by PowerMarker using Nei's genetic distance measure using (A) UPGMA and (B) Neighbour Joining methods.

PCA or PCoA made NTSYS-pc the most commonly used software though it was paid software.

## CONCLUSION

From this study, it can be concluded that for finding the genetic relationship with codominant genotypic data PowerMarker was the best among the above discussed softwares and if the requirement was to identify the genetic structure specifically then the STRUCTURE was best. Arlequin was most suitable for haplotypic data and PopGene should be used for closely related species only, that's too when their genotype was represented by codominant molecular markers and similarly Winboot can't resolve null allele properly so Winboot should be used with codominant molecular markers only.

## REFERENCES

Dearfield, K. L., Gollapudi, B. B., Bemis, J. C., Benz, R. D., Douglas, G. R., Elespuru, R. K., Johnson, G. E., Kirkland, D. J., LeBaron, M. J., Li, A. P., Marchetti, F., Pottenger, L. H., Rorije, E., Tanir, J. Y., Thybaud, V., van Benthem, J., Yauk, C. L., Zeiger, E., and Luijten, M. (2017). Next generation testing strategy for assessment of genomic damage: A conceptual framework and considerations. *Environ. Mol. Mutagen.* **58**: 264-283.

Domínguez, C., Heras, J., Mata, E., Pascual, V., Vázquez-Garcidueñas, M. S. and Vázquez-Marrufo, G. (2017). Extending Gel J for interoperability: Filling the gap in the bioinformatics resources for population genetics analysis with dominant markers. *Comp. Meth. Program. Biomed.* **140**: 69-76.

Flanagan, S. P. and Jones, A. G. (2019). The future of parentage analysis: From microsatellites to SNPs and beyond. *Mol. Ecol.* **28**: 544-567.

Hua, G. J., Hung, C. L., Lin, C. Y., Wu, F. C., Chan, Y. W. and Tang, C. Y. (2017). MGUPGMA: A fast UPGMA algorithm with multiple graphics processing units using NCCL. *Evolution. Bioinfor. online* **13**: 1176934317734220. *https://doi.org/10.1177/1176934317734220.*

Meirmans, P. G., Liu, S. and van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *J. Heredity* **109**: 283-296.

Perez, A., MacCallum, J. L. and Dill, K. A. (2015). Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Nat. Acad. Sci. United States of Am.* **112**: 11846-11851.

Salem, K. F. and Sallam, A. (2016). Analysis of population structure and genetic diversity of Egyptian and exotic rice (*Oryza sativa* L.) genotypes. *Comptes Rendus Biologies* **339**: 01-09.

Varón-González, C., Whelan, S. and Klingenberg, C. P. (2020). Estimating phylogenies from shape and similar multidimensional data: Why it is not reliable. *Syst. Biol.* **69**: 863-883.