HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

MolMarker: A Simple Tool for DNA Fingerprinting Studies and Polymorphic Information Content Calculation

Jahnke, Gizella

Multidisciplinary Digital Publishing Institute
2022-06-19

Jahnke, G.; Smidla, J.; Poczai, P. MolMarker: A Simple Tool for DNA Fingerprinting Studies and Polymorphic Information Content Calculation. Diversity 2022, 14, 497.

http://hdl.handle.net/10138/349366

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

diversity

MDPI

Article

MolMarker: A Simple Tool for DNA Fingerprinting Studies and Polymorphic Information Content Calculation

Gizella Jahnke [1,*], József Smidla [2,*] and Peter Poczai [3]

[1] Institute for Viticulture and Oenology, Buda Campus, Hungarian University of Agriculture and Life Sciences, Villányi Str. 29-43, H-1118 Budapest, Hungary
[2] Institute of Informatics and Mathematics, University of Sopron, H-4010 Debrecen, Hungary
[3] Botany Unit, Finnish Museum of Natural History, University of Helsinki, P.O. Box 7, FI-00014 Helsinki, Finland; peter.poczai@helsinki.fi
* Correspondence: gyorffyne.jahnke.gizella@uni-mate.hu (G.J.); smidla.work@gmail.com (J.S.)

Abstract: Molecular markers and mapping are used to analyze an organism's genes. They allow the selection of target genetic areas based on marker genotype (and not trait phenotype), facilitate the study of genetic variability and diversity, create linkage maps, and follow individuals or lines carrying certain genes. They may be used to select parental genotypes, remove linkage drag in back-crossing, and choose difficult-to-measure characteristics. Due to a lack of genetic variety in crops, the gene pools of wild crop relatives for future agricultural production have been examined. The invention of RFLP (Restriction Fragment Length Polymorphism) for linkage mapping allowed for the creation of other traditional approaches such as RAPD (Random Amplified Polymorphic DNA) and AFLP (Amplified Fragment Length Polymorphism). Accordingly, the need to describe the polymorphic information content (PIC) of the ideal marker has been raised. Marker selection reliability depends on the marker's relationship to the genomic area of interest. Although informativeness must be estimated for genetic study design, there are no readily available tools. Earlier, PICcalc was developed to calculate heterozygosity (H) and PIC to simplify molecular investigations. These two values were corrected for dominant and co-dominant markers (binary and allelic data) to determine polymorphism quality. Due to the popularity of PICcalc web, we developed a downloadable version called MolMarker with extra functionality to reduce server maintenance.

Keywords: PIC; heterozygosity; pedigree analyses; molecular marker; biochemical marker; genetic marker; SSR; isozyme

check for updates

Citation: Jahnke, G.; Smidla, J.; Poczai, P. MolMarker: A Simple Tool for DNA Fingerprinting Studies and Polymorphic Information Content Calculation. *Diversity* **2022**, *14*, 497. https://doi.org/10.3390/d14060497

Academic Editor: Michael Wink

Received: 25 May 2022
Accepted: 16 June 2022
Published: 19 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The primary means to study the genetic features of an organism rely on genetic markers and mapping. Molecular markers are the major tools to identify genomic regions involved in the control of traits of interest. They also facilitate selection for the target genomic regions on the basis of marker genotype rather than the phenotype of the concerned trait [1]. For example, these markers play a key role in studies on genetic variability and diversity, construction of linkage maps, and tracking individuals or lines carrying particular genes. They can be used to select and pair parental genotypes or eliminate linkage drag in back-crossing and select traits that are difficult to measure using phenotypic assays [2]. Molecular markers have many other applications, including in phylogenetics and systematics, conservation biology, molecular ecology, developmental biology, forensics, disease testing, and paternity assessment [3].

The pivotal role of molecular markers can be seen in plant breeding, where developing improved varieties is crucial for food security on a global scale. Given the continuously increasing human population, declining agricultural resources, and the stresses generated by climate change, plant breeding is expected to make greater contributions in increasingly shorter time frames [1]. In some cases, due to the lack of genetic diversity in crops, efforts

have been made to explore the gene pools of wild species for potential utilization in meeting the future challenges of crop production. Thus, the main aim of breeding programs nowadays is to trace diversity and to find new traits, particularly genes conferring resistance to diseases and pests present in wild genetic resources. This is done to maintain current levels of agricultural productivity, and molecular markers are essential tools in this process.

In recent years, many promising new alternative molecular marker techniques have been developed. This was largely due to rapid growth in genomic research, which initiated a trend away from random DNA markers toward gene-targeted functional markers. Due to the rapid expanse of several public genomic databases and next-generation sequencing technologies, the development of such functional markers located in or near candidate genes of interest has become relatively simple. With the advent of genome sequencing projects, high throughput genotyping-by-sequencing (GBS) methods eliminated the need to create individual genetic markers [4]. However, numerous species lack sufficient genome data for GBS methods, and in these cases, the use of PCR amplification remains an important tool for marker development.

The development of restriction fragment length polymorphism (RFLP) for linkage mapping in humans by Botstein et al. [5] not only created the possibility for the development of other classical methods, such as random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP), but also pinpointed the measures of an ideal marker by describing polymorphic information content (PIC). The reliability of marker selection depends mainly on the strength of linkage between the marker and the genomic region of interest. For the accurate design of genetic studies, such estimates must be calculated to describe the informativeness of the markers. However, there are currently no easily accessible calculators for that purpose. To simplify the work of molecular studies, we previously developed a useful online tool PICcalc [6] for the calculation of heterozygosity (H) [7] and PIC. These two values were adjusted for both dominant and co-dominant markers (both binary and allelic data) to measure the quality or informativeness of the polymorphism of the genetic marker. Currently, PICcalc is the only accessible program that can easily calculate these values for genetic studies in various organisms [8–10]. Due to the popularity and high demand for PICcalc web, we sought to develop a downloadable version with additional features that could operate independently of continuous server maintenance procedures. In addition, MolMarker has an easy-to-learn user-friendly graphical user interface (GUI). Java was used as a programing language, which provides platform independence. The software consists of a core application and joint plugins, which makes the software suitable for built-in new algorithms. The core application is responsible for the service and display of the GUI, the projects, and the data, as well as for some simple computations. The plugins carry out the following operations: PIC and H calculation, database editing, construction of dendrograms, calculation of parent-offspring relations, and null allele estimation.

Here, we present our software MolMarker v1.0 (Jahnke G. and Smidla J.; Veszprém, Hungary) (Figure 1) which integrates the key features of PICcalc and also provides various novel functions for genetic marker analyses based on DNA fingerprinting techniques. The user-friendly software has a graphical user interface (GUI) and is platform-independent (Java application).

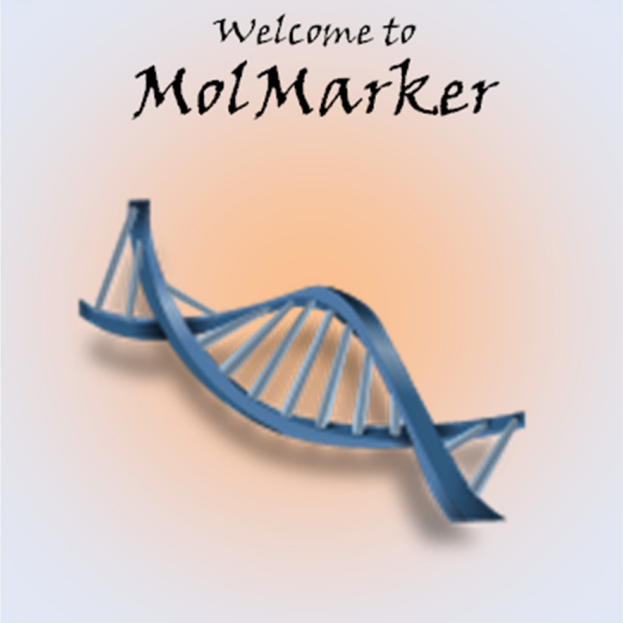

Figure 1. The MolMarker welcome screen.

2. Methods

2.1. Programming Language and IDE

Java is a general-purpose, object-oriented programming language. Object-oriented means that the basic units of the software developed are the so-called objects which allow the modular structure of the program and its subsequent further development. Another major advantage of this programming language is its platform independence, which means that the software developed can be run on any operating system simply by installing the appropriate "Java Virtual Machine" (JVM) on the computer (operating system) [11,12]. The JVM is available for the vast majority of operating systems in use today.

To develop the software on the Windows Vista operating system, the NetBeans 8.0 integrated development environment was used. This IDE allows programmers to write, compile, test, and debug applications, and then profile and deploy the programs [13]. NetBeans supports not only Java but other programming languages. NetBeans IDE can be extended with other modules [14], is free to use, has no restrictions on its use, and effectively supports the creation of GUI applications, allowing the development of user-friendly software [15].

2.2. Main Implemented Algorithms

2.2.1. UPGMA Algorithm

The Unweighted Pair Group Method with Arithmetic Mean algorithm (UPGMA) [16,17] is used to reconstruct phylogenic trees (dendrograms) using a similarity matrix as the input, which is a simple hierarchical clustering procedure.

This method is the simplest for constructing phylogenic trees. Its main drawback is that it assumes the same evolutionary rate for all lineages, i.e., the mutation rate is constant over time (molecular clock theory) [18]. This means that the final apices (leaves) are equidistant from the tree root. As it is highly unlikely that each branch will have the same mutation rate, UPGMA often generates a tree with faulty topology. The algorithm generates a rooted, ultrametric tree and has a run time of $O(n^2)$ [19].

2.2.2. Neighbor-Joining Algorithm

The neighbor-joining algorithm [20,21] is also used to reconstruct phylogenic trees, but also determines the length of the different branches. In each cycle, the "nearest vertices"

of the tree are selected, called neighbors. This is performed recursively in each cycle until all vertices are paired [22].

The algorithm takes a distance matrix as input and sequentially modifies the original star topology tree while minimizing the sum of branch lengths, thus approximating the so-called minimum-evolution method [23–25]. The algorithm has a run time of $O(n^3)$.

2.2.3. The Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm was first formulated by Dempster and colleagues [26]. This algorithm is an iterative method designed to provide maximum likelihood estimates of the parameters of statistical models where the model itself depends on missing or hidden data. The EM iteration consists of the following two steps:

Step 1, E (Expectation): in this step, the missing data are calculated by training a conditional expected value based on the estimated values of the parameters.

Step 2, M (Maximization): Based on the data calculated in the previous step and the existing data, a new estimate of the model parameters is made by maximizing the likelihood function.

The iterations are continued until the difference between the previous and the current value of the likelihood function is less than a predefined, sufficiently small value.

The EM algorithm can be used to estimate the frequency of null alleles in PCR-based genetic markers. In this case, heterozygotes carrying the null allele are indistinguishable from homozygotes carrying the detectable allele, so in this case, the null allele can be considered hidden data. The other problem is that if no product (missing data) is obtained in the PCR reaction, there are two possible reasons for this. It is possible that the tested individual is homozygous for the null allele at the locus or the genotyping failed due to some other error.

In the MolMarker software, the EM algorithm developed by Kalinowski and Taper [27] was implemented to estimate null alleles.

3. Results

Description of the Software and Its Functionalities

The menu structure of MolMarker is provided in Table 1. After installation, new projects can be created or input files can be read by the software. MolMarker employs semicolon-delimited files as input, described as 'molecular', for isozymes or other types of biochemical markers, or 'genetic' type input files, coded in binary (presence/absence) format. Input files are further described in the manual and example files are provided in the software package.

Table 1. Menu structure of MolMarker.

Project	**Data**	**Display**		**Save**		**Help**
New Project	Input Data	Data		Save Project		Open Manual
Rename Project	Read Data From File	Similarity Matrix		Save Project As . . .		Support
Change Active Project	Read Data From Database	Summary Statistics >	Genetic	Similarity Matrix		About
Open Project	Write Data To Database		Molecular	Summary Statistics >	Genetic	
Close Project		Phylogeny >	Paretage		Molecular	
Merge Projects			Dendogram	Phylogeny >	Paretage	
Exit					Dendogram	

During data management, it is possible to upload the data entered into an online database (Figure 2). The MolMarker.sql file, which is available online (also attached to this article as Supplementary Material), is used to create the web SQL database.

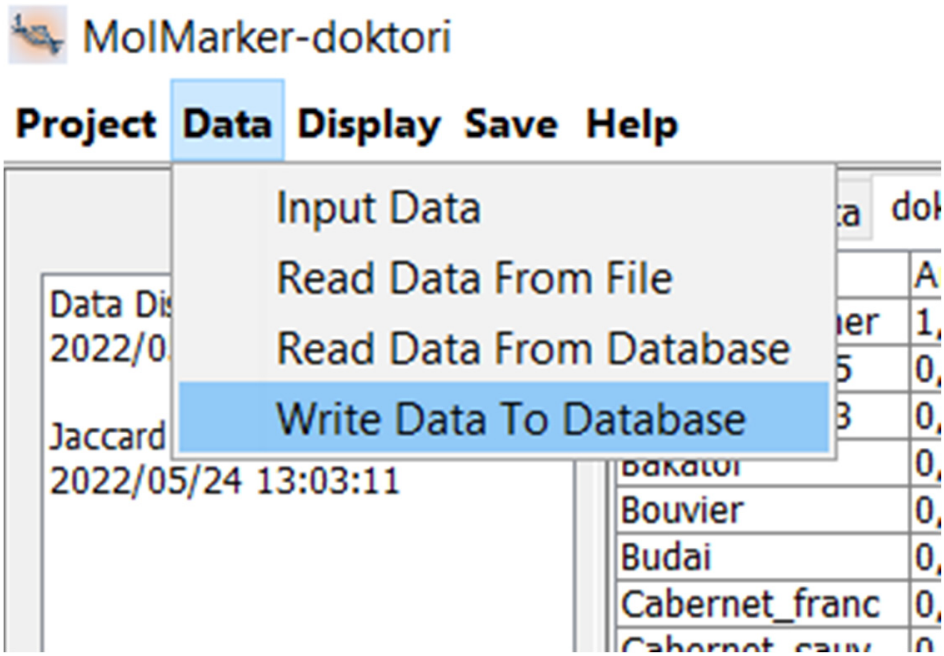

Figure 2. Menu items to upload or download data.

Summary statistics, including allele frequencies, H, and PIC can be displayed or also saved under 'Display/Summary Statistics' or 'Save/Summary Statistics'. Before the allele frequencies are displayed, it is necessary to indicate in which loci a null allele is possible (Figure 3). For example, a screenshot of the summary statistics display is shown in Figure 4.

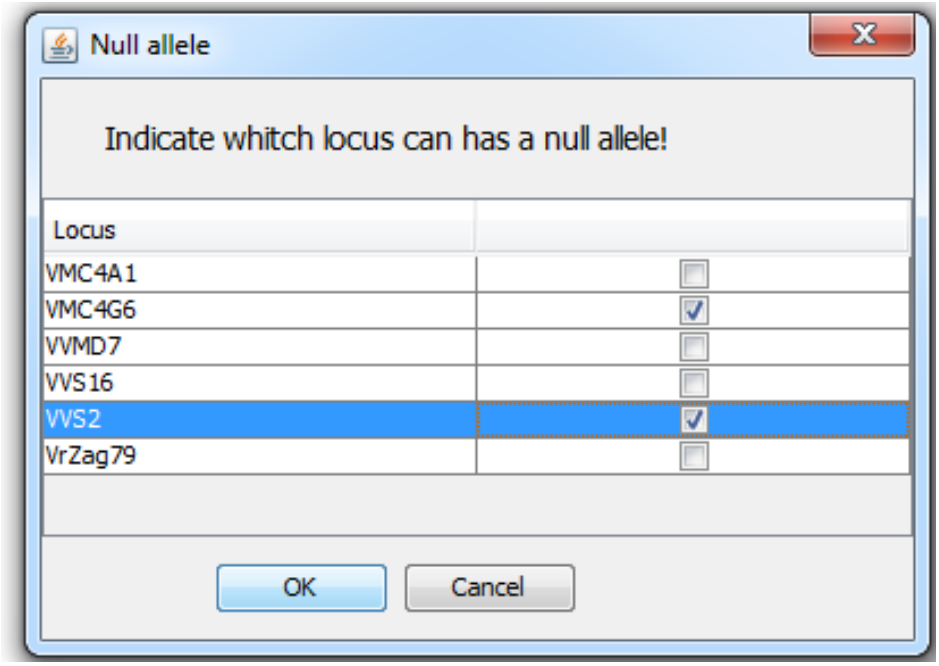

Figure 3. Input dialog to indicate which loci have a null allele.

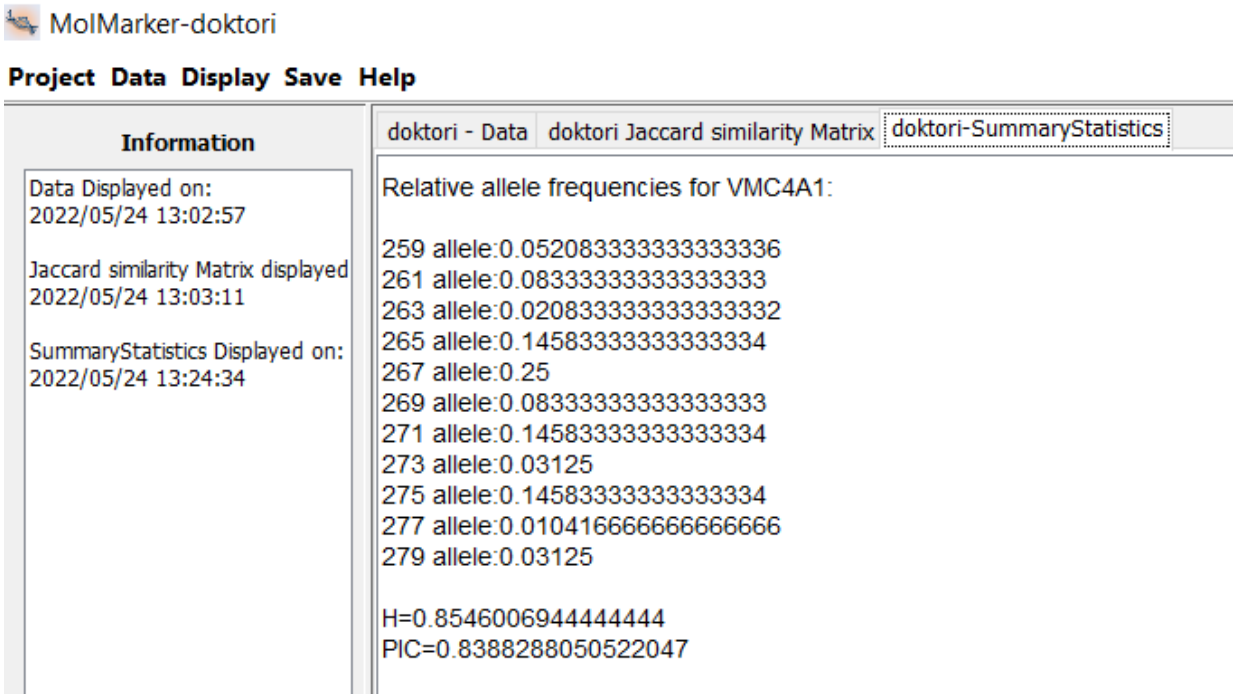

Figure 4. Screenshot of the display of 'Summary Statistics'.

Similarity matrices can also be obtained based on Jaccard similarity, simple matching (SM) [17,28] and the Czekanowski–Dice [29–31] and Ochiai [32] coefficients (Figure 5).

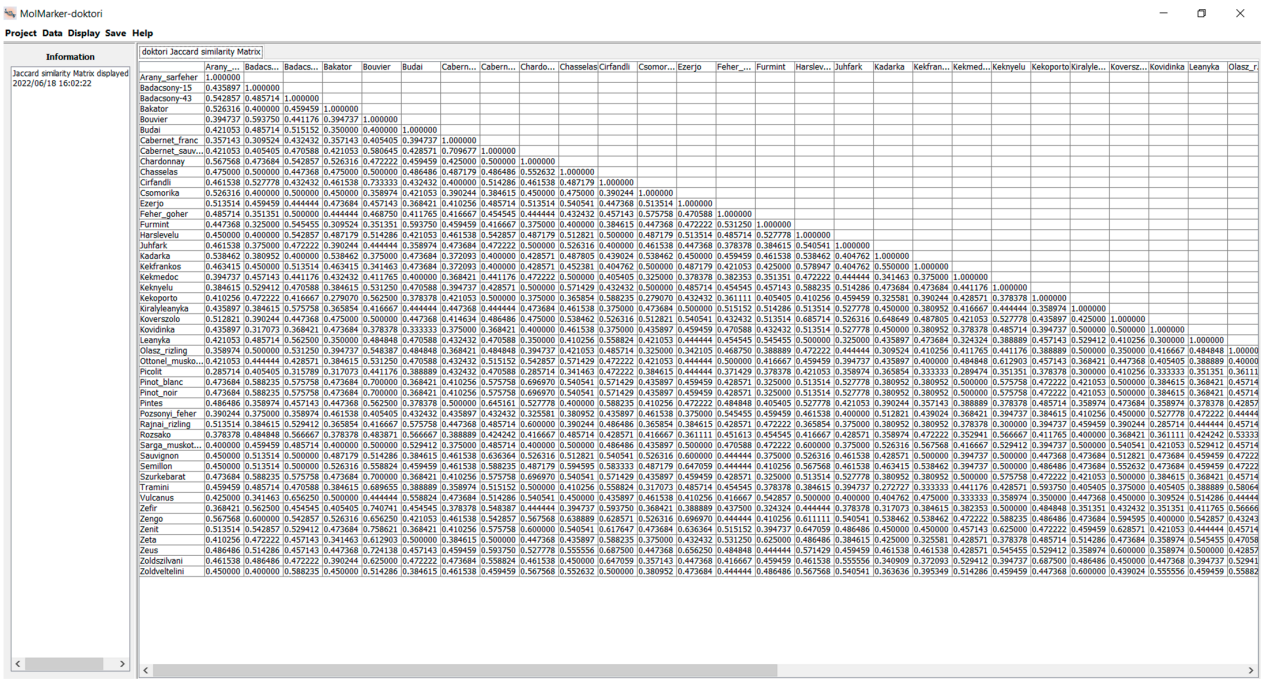

Figure 5. Similarity matrix using Jaccard similarities of SSR data.

The first snapshot of relationships among samples is displayed by the UPGMA and Neighbor-Joining methods (Figure 6). As these methods are preceded by other methods, we recommend subjecting the data set to more rigorous analysis with other programs and using MolMarker for data exploration.

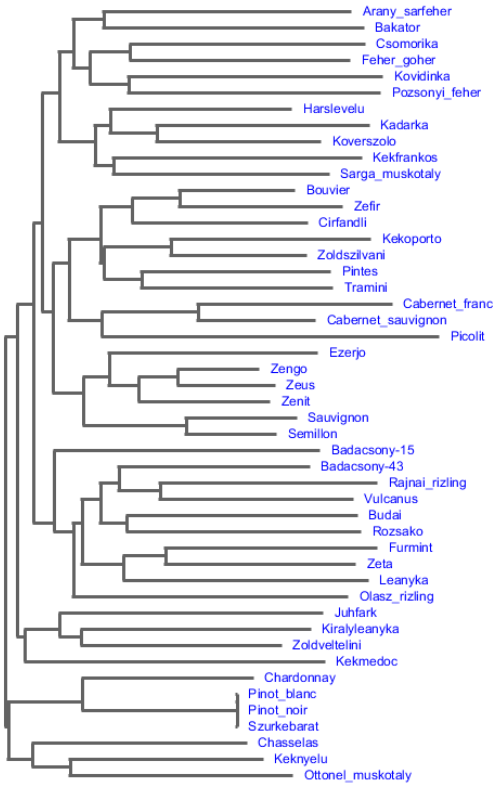

Figure 6. Neighbor-Joining dendrogram using Jaccard similarities displayed by MolMarker.

The parentage analyses option provides a list of possible parent-offspring and likelihood ratio statistics corresponding to the detected combinations.

Using the intuitive graphical user interface, basic marker statistics for genetic studies can be obtained with MolMarker. The software is open source and can be downloaded for free [33]. The software has been downloaded 351 times since its registration (Figure 7).

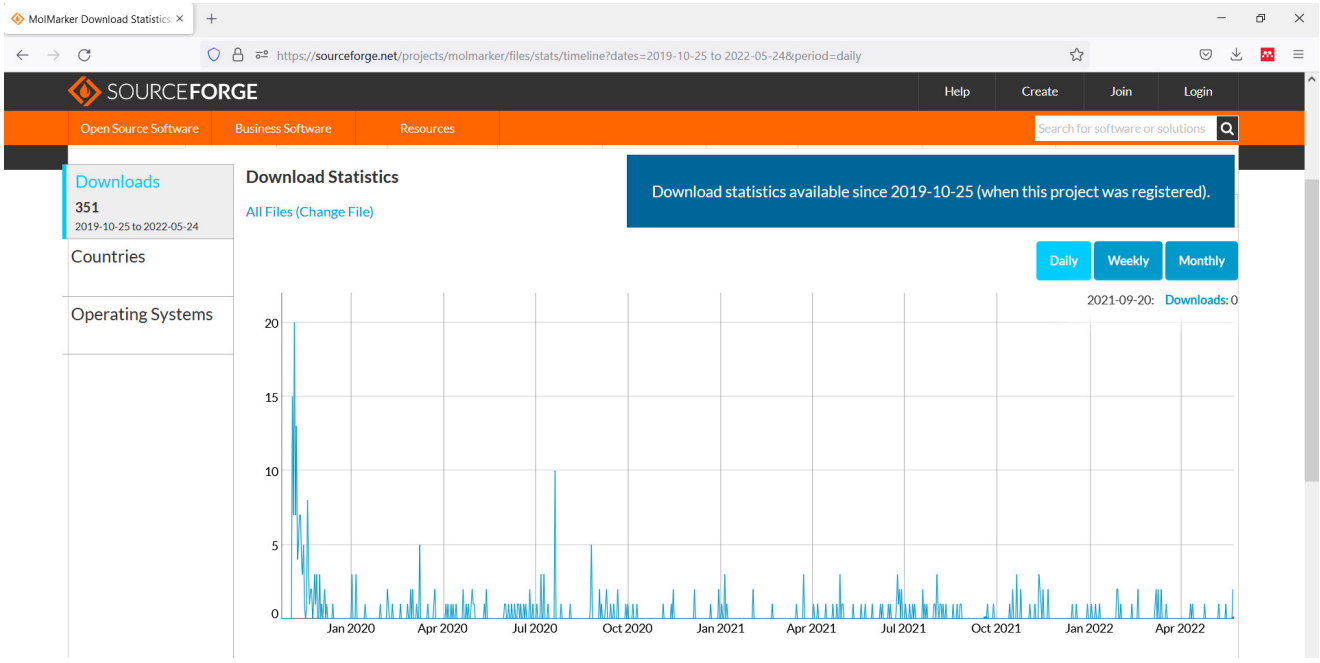

Figure 7. Download statistics for MolMarker (accessed from 25 October 2019 till 24 May 2022).

4. Discussion

Research studies based on molecular markers frequently use a large number of samples, or if the sample size is small, multiple alleles of a single molecular marker are implemented to increase the reliability of the study making it almost inconceivable to evaluate the results without computational support. Although the currently available software is able to process specific data (sets), it is often required to compare and evaluate research data belonging to various different types of markers from several perspectives. Currently, there is no such software available, researchers use numerous (usually 5–10) different programs—many of which are general-purpose spreadsheets or statistical programs—and there is a strong demand for an "*all-in-one*" downloadable software.

For example, a general-purpose spreadsheet (e.g., MS Excel) is most commonly used to calculate marker summary statistics, while the calculation of the similarity matrix and dendrograms is carried out in the statistical software package SPSS [34]. For parentage analyses, Identity 1.0 [35] is often employed, which calculates a wide range of statistics in a non-user-friendly way allowing for high error rates to accumulate during data entry, which is especially cumbersome for large sample sizes. The highly popular web-based application PICcalc [6] was previously used to calculate PIC and H values. For maximum likelihood-based null allele estimation [36], ML-NULL [27] is often used also suffering from data entry difficulties.

5. Conclusions

The primary aim of this study was to develop an open-access software with a user-friendly graphical interface, which is suitable for the multi-objective evaluation of molecular marker datasets. The goals were achieved using Java programming language, while further development can be achieved by the integration of new plugins.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/d14060497/s1, File S1: Molmarker.zip—The compressed file for MolMarker installation; File S2: MolMarker.sql S2—The sql file for the creation of online SQL database. Refs. [6,17,20,23,24,37,38] are cited in the supplementary materials.

Author Contributions: Conceptualization, G.J. and P.P.; methodology, G.J. and J.S.; software, G.J. and J.S.; validation, G.J. and P.P.; writing—original draft preparation, P.P. and G.J.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References and Note

1. Singh, B.D.; Singh, A.K. *Marker-Assisted Plant Breeding: Principles and Practices*; Springer: Berlin/Heidelberg, Germany, 2015.
2. Appleby, N.; Edwards, D.; Batley, J. New Technologies for Ultra-High Throughput Genotyping in Plants. *Methods Mol. Biol.* **2009**, *513*, 19–39. [CrossRef] [PubMed]
3. Poczai, P.; Varga, I.; Laos, M.; Cseh, A.; Bell, N.; Valkonen, J.P.T.; Hyvönen, J. Advances in Plant Gene-Targeted and Functional Markers: A Review. *Plant Methods* **2013**, *9*, 6. [CrossRef] [PubMed]
4. Kim, J.H.; Lee, C.; Hyung, D.; Jo, Y.J.; Park, J.S.; Cook, D.R.; Choi, H.K. CSGM Designer: A Platform for Designing Cross-Species Intron-Spanning Genic Markers Linked with Genome Information of Legumes. *Plant Methods* **2015**, *11*, 30. [CrossRef] [PubMed]
5. Botstein, D.; White, R.L.; Skolnick, M.; Davis, R.W. Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am. J. Hum. Genet.* **1980**, *32*, 314.
6. Nagy, S.; Poczai, P.; Cernák, I.; Gorji, A.M.; Hegedűs, G.; Taller, J. PICcalc: An Online Program to Calculate Polymorphic Information Content for Molecular Genetic Studies. *Biochem. Genet.* **2012**, *50*, 670–672. [CrossRef]
7. Liu, B.H. *Statistical Genomics: Linkage, Mapping, and QTL Analysis*; CRC Press: Boca Raton, FL, USA, 1998; ISBN 9780367400743.
8. Schwartz, J.J.; Roach, D.J.; Thomas, J.H.; Shendure, J. Primate Evolution of the Recombination Regulator PRDM9. *Nat. Commun.* **2014**, *5*, 4370. [CrossRef]

9. Wang, H.-L.; Yang, J.; Boykin, L.M.; Zhao, Q.-Y.; Wang, Y.-J.; Liu, S.-S.; Wang, X.-W. Developing conversed microsatellite markers and their implications in evolutionary analysis of the Bemisia tabaci complex. *Sci. Rep.* **2014**, *4*, srep06351. [CrossRef]
10. Martins, S.; Simões, F.; Matos, J.; Silva, A.P.; Carnide, V. Genetic Relationship among Wild, Landraces and Cultivars of Hazelnut (Corylus Avellana) from Portugal Revealed through ISSR and AFLP Markers. *Plant Syst. Evol.* **2014**, *300*, 1035–1046. [CrossRef]
11. Write Once, Run Anywhere? Computer Weekly. 2002. Available online: http://www.computerweekly.com/Articles/2002/05/02/186793/write-once-run-anywhere.htm (accessed on 25 May 2022).
12. Gosling, J.; McGilton, H. *The Java Language Environment*; Sun Microsystems Computer Company: Hongkong, China, 1996.
13. Wielenga, G. Java Editor. In *Beginning NetBeans IDE*; Apress: Berkeley, CA, USA, 2015; pp. 31–83. [CrossRef]
14. Wielenga, G. Putting the Pieces Together. In *Beginning NetBeans IDE*; Apress: Berkeley, CA, USA, 2015; pp. 103–123. [CrossRef]
15. Wielenga, G. *Beginning Netbeans Ide: For Java Developers*; Apress: Berkeley, CA, USA, 2015. [CrossRef]
16. Sneath, P.H.A.; Sokal, R.R. Numerical Taxonomy. *Nature* **1962**, *193*, 855–860. [CrossRef]
17. Sokal, R.; Michener, C. A Statistical Method for Evaluating Systematic Relationships. *Univ. Kansas Sci. Bull.* **1958**, *38*, 1409–1438; ISBN 0001948000237.
18. Vinay, S. *Text Book of Bioinformatics*; Rakesh Kumar Rastogi for Rastogi Publications: New Delhi, India, 2008.
19. Gronau, I.; Moran, S. Optimal Implementations of UPGMA and Other Common Clustering Algorithms. *Inf. Process. Lett.* **2007**, *104*, 205–210. [CrossRef]
20. Saitou, N.; Nei, M. The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425. [CrossRef] [PubMed]
21. Studier, J.A.; Keppler, K.J. A Note on the Neighbor-Joining Algorithm of Saitou and Nei. *Mol. Biol. Evol.* **1988**, *5*, 729–731. [CrossRef] [PubMed]
22. Gronau, I.; Moran, S. Neighbor Joining Algorithms for Inferring Phylogenies via LCA Distances. *J. Comput. Biol.* **2007**, *14*, 1–15. [CrossRef]
23. Rzhetsky, A.; Nei, M. A Simple Method for Estimating and Testing Minimum-Evolution Trees. *Mol. Biol. Evol.* **1992**, *9*, 945. [CrossRef]
24. Gascuel, O.; Bryant, D.; Denis, F. Strengths and Limitations of the Minimum Evolution Principle. *Syst. Biol.* **2001**, *50*, 621–627. [CrossRef]
25. Kuhner, M.K.; Felsenstein, J. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Mol. Biol. Evol.* **1994**, *11*, 459–468. [CrossRef]
26. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–22. [CrossRef]
27. Kalinowski, S.T.; Taper, M.L. Maximum Likelihood Estimation of the Frequency of Null Alleles at Microsatellite Loci. *Conserv. Genet.* **2006**, *7*, 991–995. [CrossRef]
28. Hardy, G.H. Mendelian Proportions in a Mixed Population. *Science* **1908**, *28*, 49–50. [CrossRef]
29. Czekanowski, J. Zarys Metod Statystycnck Anthr. *Anz* **1913**, *9*, 227–249.
30. Czekanowski, J. *Zarys Metod Statystycznych W Zastosowaniu Do Antropologii*; Towarzystwo Naukowe Warszawskie: Warszawa, Poland, 1913.
31. Dice, L. Measures of the Amount of Ecological Association between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
32. Batzer, D.P.; Rader, R.B.; Wissinger, S.A. *Invertebrates in Freshwater Wetlands of North America: Ecology and Management*; John Wiley & Sons: Hoboken, NJ, USA, 1999; p. 1100.
33. MolMarker Download | Source Forge. Available online: https://sourceforge.net/projects/molmarker/ (accessed on 21 May 2022).
34. IBM SPSS Statistics. Available online: https://www.ibm.com/products/spss-statistics (accessed on 13 June 2022).
35. Wagner, H.; Sefc, K. IDENTITY 1.0 Centre for Applied Genetics. 1999. Available online: https://boku.ac.at/zag/forsch/identity.htm (accessed on 7 January 2019).
36. Kalinowski, S. ML-Null Freq-Steven Kalinowski | Montana State University. Available online: https://www.montana.edu/kalinowski/software/null-freq.html (accessed on 12 May 2022).
37. Bowers, J.E.; Meredith, C.P. The parentage of a classic wine grape, Cabernet Sauvignon. *Nat. Genet.* **1997**, *16*, 84–87.
38. Speed, T. Neighbour Joining Method (Saitou and Nei, 1987). 2006.