BinMat: A molecular genetics tool for processing binary data obtained from fragment analysis in R

Clarke van Steenderen ‡

‡ Centre for Biological Control, Department of Zoology and Entomology, Rhodes University, Grahamstown/Makhanda, South Africa

Corresponding author: Clarke van Steenderen (vsteenderen@gmail.com)

Academic editor: Zachary Foster

Abstract

Processing and visualising trends in the binary data (presence or absence of electropherogram peaks), obtained from fragment analysis methods in molecular biology, can be a time-consuming and often cumbersome process. Scoring and analysing binary data (from methods, such as AFLPs, ISSRs and RFLPs) entail complex workflows that require a high level of computational and bioinformatic skills. The application presented here (BinMat) is a free, open-source and user-friendly R Shiny programme (https://clarkevansteenderen.shinyapps.io/BINMAT/) that automates the analysis pipeline on one platform. It is also available as an R package on the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/web/packages/BinMat/index.html). BinMat consolidates replicate sample pairs of binary data into consensus reads, produces summary statistics and allows the user to visualise their data as ordination plots and clustering trees without having to use multiple programmes and input files or rely on previous programming experience.

Keywords

AFLP, binary data scoring, GUI, ISSR, R package, R Shiny

Introduction

Fragment analysis is a method in molecular biology that encompasses the processes by which fragments of DNA are separated by size in order to generate characteristic band profiles. Bands are detected and scored through either the traditional method of viewing them on polyacrylamide gels (Bassam et al. 1991) or through the use of fluorescent markers (such as FAM^{TM} or ROX) that tag fragments so that they can be detected by capillary electrophoresis (Dresler-Nurmi et al. 2000, Applied Biosystems 2014). There are a number of techniques associated with fragment analysis, including AFLP (Amplified

Fragment Length Polymorphism) (Vos et al. 1995), RAPD (Random Amplified Polymorphic DNA) (Koeleman et al. 1998) and ISSR (Inter-Simple Sequence Repeats) (Wolfe and Liston 1998, Abbot 2001). Fragment analysis offers a wide range of applications, such as DNA fingerprinting, SNP (single nucleotide polymorphism) genotyping and microsatellite profiling (Applied Biosystems 2014), which are used across a broad range of disciplines.

Processing and analysing the binary data, obtained from fragment analysis methods, can quickly become challenging due to the large size of datasets and the time required to organise and format them to suit the needs of different programmes used in analysis pipelines. Common practice is to independently replicate each Polymerase Chain Reaction (PCR) sample in order to consolidate the output into one consensus read per individual (see, for example, Taylor et al. 2011 and Sutton et al. 2017). The term 'consolidate', as used here, refers to the process of checking the binary value scored at each locus position across every replicate pair and creating one representative consensus output for that sample. For example, if both replicates show the presence of a band at a particular locus, a '1' is recorded as 'present' at that locus. If a band was absent in both replicates, a '0' is recorded. If one replicate shows the presence of a band, but the other shows an absence, a '?' is recorded to denote an ambiguous read.

Manually consolidating the replicate pairs of large binary matrices in this way is not only impractical, but it also lends itself to human error. Even after fragments have been scored and processed, the downstream analyses of these data are complex. For example, a number of different programmes are often required for different analyses, each of which require a different input file format. This requires a certain level of computational and/or bioinformatic skills, can be both difficult and time-consuming and can result in further potential errors when changing between file formats.

The R programming language (RStudio Team 2020) is becoming an increasingly popular means of analysing genetic data (Paradis et al. 2004, Schliep 2011, Archer et al. 2017), as it can read in multiple file formats and perform a number of analyses all on one platform. Packages in R can, however, often be challenging to utilise for newcomers to programming. The development of a GUI (Graphical User Interface) can address this by collating multiple processing tools into one place and make complex computational tasks more accessible to researchers (see, for example, Reyes et al. 2019).

Here, I present BinMat, an R package and R Shiny application that automates the analysis of fragment data. Named 'BinMat', from '**Bin**ary **Mat**rix', the application offers researchers a user-friendly, open-source platform that does not require multiple programmes and file input formats (Fig. 1). Moreover, a GUI was developed to make data processing easier and more accessible. BinMat is available on three platforms; namely the shinyapps.io server, GitHub and as an R package on CRAN. The following sections detail the functionality of BinMat, how its output compares to PAST (Hammer et al. 2001) and SplitsTree (Huson 1998) (which are standalone software typically used to analyse genetic data) and how it can be accessed.

R Shiny graphical user interface

The R Shiny application platform allocates a maximum memory of 1 GB and is accessible here. The online version may time-out due to insufficient memory if a particularly large binary data file is uploaded. In such a case, the programme can be run directly from R on the user's local machine by typing

install.packages("shiny")

shiny::runGitHub("BinMat", "clarkevansteenderen")

into the console.

The programme's code is freely available on GitHub.

File input

BinMat reads in binary data that has already been processed from raw electropherograms using programmes such as GeneMarker (SoftGenetics) and RawGeno (Arrigo et al. 2012). This needs to be uploaded as a CSV (comma-separated values) file in the format shown in Table 1. Column headings are required, but are not limited to the exact labels shown in the example. If the data consist of replicate pairs, these need to be organised so that they appear consecutively, with a unique name for each sample. It is important to check the data to ensure that there are no single samples without their replicate. When the 'Consolidate matrix' button is clicked, each replicate pair in the dataset is consolidated into a consensus output.

Table 2 shows the output if the data in Table 1 were used as input. The resulting consolidated binary matrix can be downloaded as a CSV file using the 'Download Matrix' button once the message 'COMPLETE. READY FOR DOWNLOAD' appears on the screen. The 'Check my data for unwanted values' button checks the data for any values in the dataset other than a '1', '0', or '?' and returns the column and row index for the unwanted character/s.

Output overview

Once the data have been consolidated, the user can view and download information in the 'SUMMARY' tab at the top of the window; showing the average number of peaks (± standard deviation (sd)), the maximum and minimum number of peaks and the total number of loci. The 'ERROR RATES' tab shows the Euclidean (EE) (± sd) and Jaccard (JE) (± sd) error rates. See Bonin et al. 2004, Pompanon et al. 2005 and Holland et al. 2008 for detailed reviews regarding error rates and their calculation.

The 'Remove samples with a jaccard error greater than' button removes samples with a Jaccard error (ranging from 0 to 1) greater than or equal to a specified value. This can give

the user an idea of how filtering their data can affect overall error rates. The default value is set at zero.

Clustering methods, such as the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and neighbour-joining, are frequently used in the analyses of fragment data to create dendrograms (e.g. Van Eldere et al. 1999, Ticknor et al. 2001, Liu et al. 2009, Timm et al. 2010). Additionally, ordination methods, such as those offered by non-metric multidimensional scaling (nMDS) plots, are also often used (see, for example, Denaro et al. 2005, Zhang et al. 2008, Vašek et al. 2017).

Hierarchical clustering tree: UPGMA

The 'UPGMA TREE' tab in BinMat allows the user to upload a consolidated binary matrix as a CSV file (in the format shown in Table 2), specify the number of bootstrap replications and download the resulting hierarchical clustering tree as a scalable vector graphics (SVG) file. This function makes use of the pvclust function in the pvclust package (Suzuki and Shimodaira 2006) and uses the UPGMA clustering method. The uploaded binary data are converted into a distance matrix applying the Jaccard transformation (dJ_i) (Jaccard 1908) shown below. $\mathbf{f_{11}}$ represents the total number of times that a band occurred at the same locus in both samples, $\mathbf{f_{00}}$ represents the shared absence of bands and $\mathbf{f_{10}}$ and $\mathbf{f_{01}}$ represents the number of times that a band was present in only one of the two sample replicates. The Jaccard transformation was applied using the *dist* function, applying the 'binary' method. This transformation was prefered because it does not treat the shared absence of bands as being biologically meaningful.

$$dJi = \frac{f01 + f10}{f01 + f10 + f11}$$

Ordination: nMDS Plot

The 'nMDS PLOT' tab allows the user to upload a consolidated binary matrix with grouping information as a CSV file. The input file format is shown in Table 3, where grouping information needs to appear in the second column.

The distance methods available are 'binary' (Jaccard's distance), 'euclidean', 'maximum', 'manhattan', 'canberra' and 'minkowski'. The 'No. of dimensions (k)' option can be set at '2' or '3' and can be determined using the 'nMDS Validation' tab using the 'Scree plot' and 'Shepard plot' buttons. The resulting distance matrix can be downloaded as a CSV file and the plot itself as a SVG file. Once the user has uploaded their data, an editable table will appear to allow for the selection of colours and symbols for each group. The user can adjust symbol size and can select whether sample labels should appear on the graph or not. The nMDS plot is created using the isoMDS function in the MASS package (Venables and & Ripley 2002).

Scree plot

The optimal number of dimensions to use for the nMDS plot should minimise the resulting stress value. Clarke 1993 suggests that stress values < 0.05 = excellent, < 0.10 = good,

< 0.20 = usable, > 0.20 = not acceptable, while Dugard et al. 2010 suggest that a stress value below 0.15 represents a good fit for the data. BinMat indicates the 0.15 threshold as a dotted red line on the resulting scree plot.

Shepard plot

Shepard plots are graphical representations of how well the ordination fits the original distance data (Leeuw and Mair 2014). BinMat plots the original Jaccard distances (x-axis) against the transformed distances used to create the nMDS ordination plot (y-axis). R^2 values are shown on the plot for the regression line of best fit.

Filter data

The 'Filter data' tab allows the user to filter their dataset by setting a threshold value for the number of peaks present. The new subsetted data and the removed samples can be downloaded as a CSV file and re-uploaded to create a new nMDS plot and/or hierarchical clustering tree.

Testing BinMat

Comparing BinMat's output to PAST and SplitsTree

Two AFLP datasets were downloaded from the Dryad Digital Repository. These comprised data generated by Arias et al. 2014 and Tewes et al. 2018 for *Heliconius* (Lepidoptera: Nymphalidae) and *Bunias orientalis* L. (Brassicaceae) specimens, respectively. With the authors' permission, a subset of each were used to compare output from BinMat to that of PAST v.4.0 (Paleontological Statistics Software Package for Education and Data Analysis) (Hammer et al. 2001) and SplitsTree v.4.14.6 (Huson 1998) (input data are available in Suppl. materials 1, 2, 3, 4, 5, 6, 7). Replicate pairs were consolidated in BinMat and used to create nMDS plots and UPGMA hierarchical clustering trees (1000 bootstrap repetitions). The lowest number of dimensions were used for nMDS plots (k = 2) and their stress and R^2 values recorded. SplitsTree was used to create a NeighborNet tree applying Jaccard's distance transformation. The nMDS plots created by BinMat and PAST showed comparable clustering patterns (Fig. 2A1, A2, B1 and B2).

The SplitsTree output for the data taken from Tewes et al. 2018 (Fig. 2B4) corroborated the corresponding nMDS plot from the original paper (Fig. 2B3) and from that created by BinMat (Fig. 2B1). Both hierarchical clustering trees using the UPGMA method showed equivalent topologies and bootstrap support values for clades (Fig. 3). BinMat, PAST and SplitsTree perform equally as well for the visualisation of fragment analysis output, where BinMat offers the advantage of a quicker, automated process on one platform.

BinMat as an R package on CRAN

The BinMat R package is available on the Comprehensive R Archive Network (CRAN) and on GitHub and is command-line driven. More information about the package can be obtained by typing

library(help = BinMat)

into the console after it has been installed. This details all the functions available (Table 4).

To cite BinMat, use

citation("BinMat")

There are four example binary matrices embedded in the BinMat package called "BinMatInput_reps", "BinMatInput_ordination", "bunias_orientalis" and "nymphaea" that can be accessed by creating objects such as:

data1 = BinmatInput_reps

data2 = BinmatInput_ordination

These binary matrices can be used to test the various functions as a demonstration example, as shown in the worked example in the vignette supplied with the package. The "BinMatInput_reps" and "BinMatInput_ordination" are small hypothetical datasets, illustrating how BinMat consolidates replicate pairs and then creates an nMDS plot coloured by groups (e.g. populations). The "bunias_orientalis" and "nymphaea" datasets are real-world AFLP and ISSR results from Tewes et al. 2018 and Reid et al. 2021, respectively. These two datasets have already been consolidated and serve as examples for the generation of nMDS plots.

Conclusion

BinMat offers users of fragment analysis methods an efficient and easy-to-use platform to process their binary data matrices, by means of either a graphical user interface or an R package. The programme produces comparable output to other mainstream software, with the benefit of housing all of its functionality on one platform. Suggestions for improvement (for example via pull-requests on GitHub) and feedback from the community, are welcomed.

Acknowledgements

This work was supported by funding from the South African Working for Water (WfW) programme of the Department of Forestry, Fisheries and the Environment: Natural

Resource Management Programmes (DFFE: NRMP). Funding was also provided by the South African Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation (NRF) of South Africa. Any opinion, finding, conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard. Guy Sutton is thanked for his valuable advice and suggestions in the writing of this manuscript. Megan Reid is thanked for providing her *Nymphaea* ISSR dataset, testing BinMat's functionality and providing ongoing feedback. Prof. Iain D. Paterson and Dr. Shelley Edwards are thanked for their assistance throughout the course of my MSc degree.

Hosting institution

Centre for Biological Control, Department of Zoology and Entomology, Rhodes University, South Africa

Conflicts of interest

There are no conflicts of interest.

References

- Abbot P (2001) Individual and population variation in invertebrates revealed by Inter-simple Sequence Repeats (ISSRs). Journal of Insect Science 1 (1): 8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC355892/
- Applied Biosystems (2014) DNA fragment analysis by capillary electrophoresis. Revision B, Publication number 4474504. Thermo Fisher Scientific Inc., 220 pp. URL: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4474504.pdf
- Archer FI, Adams PE, Schneiders BB (2017) stratag: An R package for manipulating, summarizing and analysing population genetic data. Molecular Ecology Resources 17 (1): 5-11. https://doi.org/10.1111/1755-0998.12559
- Arias CF, Salazar C, Rosales C, Kronforst MR, Linares M, Bermingham E, McMillan WO (2014) Phylogeography of *Heliconius cydno* and its closest relatives: disentangling their origin and diversification. Molecular Ecology 23 (16): 4137-4152. https://doi.org/10.1111/mec.12844
- Arrigo N, Holderegger R, Alvarez N (2012) Automated scoring of AFLPs using RawGeno v. 2.0, a free R CRAN library. In: Pompanon F, Bonin A, et al. (Eds) Data production and analysis in population genomics. Springer, 20 pp. https://doi.org/10.1007/978-1-61779-870-2_10
- Bassam BJ, Caetano-Anollés G, Gresshoff PM (1991) Fast and sensitive silver staining of DNA in polyacrylamide gels. Analytical Biochemistry 196 (1): 80-83. https://doi.org/10.1016/0003-2697(91)90120-I
- Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetics studies.

Molecular Ecology 13 (11): 3261-3273. https://doi.org/10.1111/j.1365-294X.2004.02346.x

- Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 18 (1): 117-143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x
- Denaro R, D'auria G, Di Marco G, Genovese M, Troussellier M, Yakimov M, Giuliano L (2005) Assessing terminal restriction fragment length polymorphism suitability for the description of bacterial community structure and dynamics in hydrocarbon-polluted marine environments. Environmental Microbiology 7 (1): 78-87. https://doi.org/10.1111/j.1462-2920.2004.00685.x
- Dresler-Nurmi A, Terefework Z, Kaijalainen S, Lindström K, Hatakka A (2000) Silver stained polyacrylamide gels and fluorescence-based automated capillary electrophoresis for detection of amplified fragment length polymorphism patterns obtained from white-rot fungi in the genus *Trametes*. Journal of Microbiological Methods 41 (2): 161-172. https://doi.org/10.1016/S0167-7012(00)00153-6
- Dugard P, Todman J, Staines H (2010) Approaching multivariate analysis. A practical introduction. 2. Routledge/Taylor & Francis Group, Routledge, New York, 440 pp. [ISBN 9780415645911]
- Hammer O, Harper DAT, Ryan PD (2001) PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica 4 (9).
- Holland BR, Clarke AC, Meudt HM (2008) Optimizing automated AFLP scoring parameters to improve phylogenetic resolution. Systematic Biology 57 (3): 347-36. https://doi.org/10.1080/10635150802044037
- Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics (Oxford, England) 14 (1): 68-73. https://doi.org/10.1093/bioinformatics/14.1.68
- Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bulletin de la Societe Vaudoise des Sciences Naturelles 44: 223-270. https://doi.org/10.5169/seals-268384
- Koeleman JG, Stoof J, Biesmans DJ, Savelkoul PH, Vandenbroucke-Grauls CM (1998) Comparison of amplified ribosomal DNA restriction analysis, random amplified polymorphic DNA analysis, and amplified fragment length polymorphism fingerprinting for identification of *Acinetobacter* genomic species and typing of *Acinetobacter baumannii*. Journal of Clinical Microbiology 36 (9): 2522-2529. https://doi.org/10.1128/JCM.36.9.2522-2529.1998
- Leeuw JD, Mair P (2014) Shepard diagram. Wiley StatsRef: Statistics Reference Online1-3. https://doi.org/10.1002/9781118445112.stat06268.pub2
- Liu L, Meng Z, Wang B, Wang X, Yang J, Peng D (2009) Genetic diversity among wild resources of the genus *Boehmeria* Jacq. from west China determined using inter-simple sequence repeat and rapid amplification of polymorphic DNA markers. Plant Production Science 12 (1): 88-96. https://doi.org/10.1626/pps.12.88
- Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20 (2): 289-29. https://doi.org/10.1093/bioinformatics/btg412.
- Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics 6 (11): 84. https://doi.org/10.1038/nrg1707
- Reid M, Naidu P, Paterson I, Mangan R, Coetzee J (2021) Population genetics of invasive and native Nymphaea mexicana Zuccarini: Taking the first steps to initiate a

biological control programme in South Africa. Aquatic Botany 171 https://doi.org/10.1016/j.aquabot.2021.103372

- Reyes ALP, Silva TC, Coetzee SG, Plummer JT, Davis BD, Chen S, Hazelett DJ, Lawrenson K, Berman BP, Gayther SA, et al. (2019) GENAVi: a shiny web application for gene expression normalization, analysis and visualization. BMC Genomics 20 (1): 745. https://doi.org/10.1186/s12864-019-6073-7
- RStudio Team (2020) RStudio: Integrated development for R. RStudio, PBC, Boston, MA. URL: http://www.rstudio.com/
- Schliep KP (2011) Phangorn: Phylogenetic analysis in R. Bioinformatics 27 (4): 592-593. https://doi.org/10.1093/bioinformatics/btq706
- Sutton GF, Paterson ID, Paynter Q (2017) Genetic matching of invasive populations of the African tulip tree, *Spathodea Campanulata* Beauv. (Bignoniaceae), to their native distribution: Maximising the likelihood of selecting host-compatible biological control agents. Biological Control 114: 167-175. https://doi.org/10.1016/j.biocontrol.2017.08.015
- Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22 (12): 1540-1542. https://doi.org/10.1093/bioinformatics/btl117
- Taylor SJ, Downie DA, Paterson ID (2011) Genetic diversity of introduced populations of the water hyacinth biological control agent *Eccritotarsus catarinensis* (Hemiptera: Miridae). Biological Control 58 (3): 330-336. https://doi.org/10.1016/j.biocontrol.2011.05.008.
- Tewes LJ, Michling F, Koch M, Müller C (2018) Intracontinental plant invader shows matching genetic and chemical profiles and might benefit from high defence variation within populations. Journal of Ecology 106 (2): 714-726. https://doi.org/10.1111/1365-2745.12869
- Ticknor LO, Kolstø A, Hill KK, Keim P, Laker MT, Tonks M, Jackson PJ (2001) Fluorescent amplified fragment length polymorphism analysis of Norwegian *Bacillus cereus* and *Bacillus thuringiensis* soil isolates. Applied and Environmental Microbiology 67 (10): 4863-487. https://doi.org/10.1128/AEM.67.10.4863\T1\textendash4873.2001.
- Timm A, Geertsema H, Warnich L (2010) Population genetic structure of economically important Tortricidae (Lepidoptera) in South Africa: a comparative analysis. Bulletin of Entomological Research 100 (4): 421-43. https://doi.org/10.1017/S0007485309990435.
- Van Eldere J, Janssen P, Hoefnagels-Schuermans A, van Lierde S, Peetermans WE (1999) Amplified- fragment length polymorphism analysis versus macro-restriction fragment analysis for molecular typing of *Streptococcus pneumoniae* isolates. Journal of Clinical Microbiology 37 (6): 2053-2057. https://doi.org/10.1128/JCM.37.6.2053-2057.1999
- Vašek J, Čepková PH, Viehmannová I, Ocelak M, Huansi DC, Vejl P (2017) Dealing with AFLP genotyping errors to reveal genetic structure in *Plukenetia volubilis* (Euphorbiaceae) in the Peruvian Amazon. PLOS One 12 (9). https://doi.org/10.1371/journal.pone.0184259
- Venables WN, & Ripley BD (2002) Modern Applied Statistics with S. 4. Springer, New York. [ISBN 0-387-95457-0] https://doi.org/10.1007/978-0-387-21706-2
- Vos P, Hogers R, Bleeker M, Reijans M, Lee Tvd, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23 (21): 4407-4414. https://doi.org/10.1093/nar/23.21.4407

- Wolfe AD, Liston A (1998) Contributions of PCR-based methods to plant systematics and evolutionary biology. In: Soltis D, Soltis P, Doyle J, et al. (Eds) Molecular Systematics of Plants II. Springer, 43 pp. https://doi.org/10.1007/978-1-4615-5419-6_2
- Zhang R, Thiyagarajan V, Qian P (2008) Evaluation of terminal-restriction fragment length polymorphism analysis in contrasting marine environments. Federation of European Microbiological Societies (FEMS) Microbiology Ecology 65 (1): 169-178. https://doi.org/10.1111/j.1574-6941.2008.00493.x.

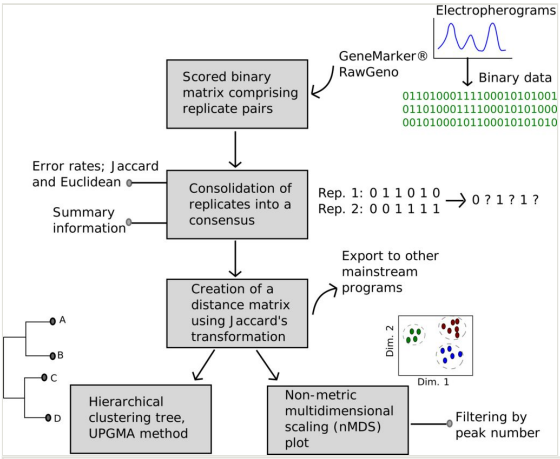

Figure 1.

Flowchart of the utility of the BinMat programme, starting with input that has been processed in programmes such as GeneMarker and RawGeno, to the rapid visualisation of a hierarchical clustering tree and non-metric dimensional scaling (nMDS) plot.

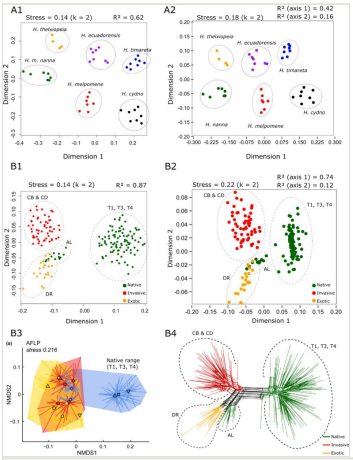

Figure 2.

Comparisons of non-metric multidimensional scaling (nMDS) plots in BinMat (**A1** and **B1**) and PAST (**A2** and **B2**). Both nMDS plots are plotted for k = 2 dimensions. Data were taken from Arias et al. 2014 (A1 and A2) and Tewes et al. 2018 (**B1**, **B2** and **B4**). Stress and R^2 values are shown above each plot. Diagram B3 shows the original nMDS plot presented by Tewes et al. 2018, which depicts the same clustering pattern of the native range samples (T1, T3 and T4). Diagram B4 shows the SplitsTree representation of the same data (NeighborNet, Jaccard distance).

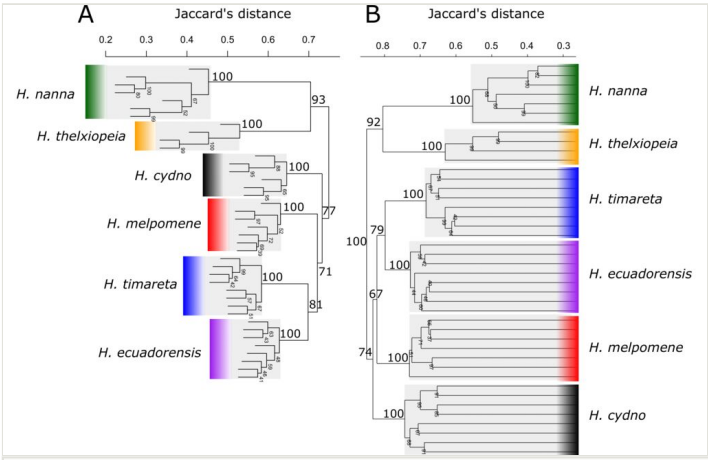

Figure 3.

Comparison of hierarchical clustering trees in A) BinMat and B) PAST using the data taken from Tewes et al. 2018. Both programmes applied Jaccard's transformation to create a distance matrix and used the UPGMA clustering method. Bootstrap probabilities are shown on the branches, resulting from 1000 bootstrap repetitions.

Table 1.

File input for a dataset containing replicate pairs that needs to be consolidated.

Sample label	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5
Sample A rep 1	0	0	1	1	1
Sample A rep 2	0	0	1	1	1
Sample B rep 1	1	1	0	0	0
Sample B rep 2	0	1	0	0	1

Table 2.

A consolidated matrix derived from Table 1 using BinMat.

Sample label	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5
Sample A rep 1 + Sample A rep 2	0	0	1	1	1
Sample B rep 1 + Sample B rep 2	?	1	0	0	?

Table 3.

Data input required for the creation of a non-metric multidimensional scaling (nMDS) plot. Grouping information needs to be in the second column. The data here represents binary replicate pairs that have already been consolidated into consensus reads.

Sample label	Group	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5
Sample A	Africa	0	0	1	1	1
Sample A	Asia	?	1	0	0	?

Table 4.

BinMat R package functions available on CRAN. Typing **?functionName** into the console provides more information about each function.

Function	Description
check.data()	Checks for unwanted characters.
consolidate()	Consolidates replicate pairs. 1 & 1 = 1; 1 & 0 = ?; 0 & 0 = 0
errors()	Calculates Jaccard and Euclidean error rates.
group.names()	Outputs groups in the uploaded binary matrix.
nmds()	Creates a non-metric multidimensional scaling (nMDS) plot.
peak.remove()	Removes samples with peaks equal to, or less than, a specified threshold value.
peaks.consolidated()	Peak summary for a consolidated binary matrix.
peaks.orignal()	Peak summary for replicate data or consolidated data from file.
scree()	Creates a scree plot of stress values vs. ordination dimensions.
shepard()	Creates a shepard plot for goodness-of-fit for ordination data.
upgma()	Draws a hierarchical clustering tree (UPGMA) with bootstrapping.

Supplementary materials

Suppl. material 1: Arias et al. (2014) consolidated binary matrix with groups

Authors: Adapted from Arias et al. (2014)
Data type: Binary data (AFLP)
Brief description: Consolidated AFLP binary data from Arias et al. (2014), with a grouping column. This is used as input to BinMat for the creation of an nMDS plot.
Download file (102.92 kb)

Suppl. material 2: Arias et al. (2014) consolidated binary matrix without groups

Authors: Adapted from Arias et al. (2014)
Data type: AFLP binary data
Brief description: Consolidated AFLP binary data from Arias et. al. (2014), without a grouping column.
Download file (178.97 kb)

Suppl. material 3: Arias et al. (2014) raw AFLP binary data

Authors: Adapted from Arias et al. (2014)
Data type: AFLP binary data
Brief description: Raw AFLP binary data from Arias et al. (2014), before replicates have been consolidated.
Download file (178.97 kb)

Suppl. material 4: Tewes et al. (2018) AFLP binary data for native and invasive Brunias orientalis species

Authors: Adapted from Tewes et al. (2018)
Data type: AFLP binary data in NEXUS format
Brief description: A NEXUS file containing AFLP binary data for native and invasive *Brunias orientalis* species from the Tewes et al. (2018) study. This file is used as input for the SplitsTree programme.
Download file (38.27 kb)

Suppl. material 5: Tewes et al. (2018) consolidated binary matrix with groups

Authors: Adapted from Tewes et al. (2018)
Data type: Consolidated AFLP binary data
Brief description: Consolidated AFLP binary data from Tewes et. al. (2018), with a grouping column. This is used as input to BinMat for the creation of an nMDS plot.
Download file (38.99 kb)

Suppl. material 6: Tewes et al. (2018) consolidated binary matrix without groups

Authors: Adapted from Tewes et al. (2018)

Data type: Consolidated AFLP binary data
Brief description: Tewes et al. (2018) consolidated AFLP binary data without a grouping column.
Download file (38.33 kb)

Suppl. material 7: Tewes et al. (2018) raw binary AFLP data

Authors: Adapted from Tewes et al. (2018)
Data type: Binary AFLP data
Brief description: Raw AFLP binary data from Tewes et al. (2018), before replicates have been consolidated.
Download file (73.61 kb)