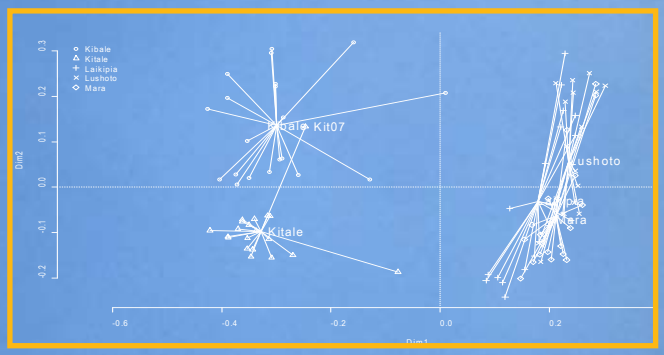
Technical Manual no.13

Molecular Markers for Tropical Trees

Statistical Analysis of Dominant Data



World Agroforestry Centre

World Agroforestry Centre
TRANSFORMING LIVES AND LANDSCAPES

The World Agroforestry Centre, an autonomous, non-profit research organization, aims to bring about a rural transformation in the developing world by encouraging and enabling smallholders to increase their use of trees in agricultural landscapes. This will help to improve food security, nutrition, income and health; provide shelter and energy; and lead to greater environmental sustainability.

We are one of the 15 centres of the Consultative Group on International Agricultural Research (CGIAR). Headquartered in Nairobi, Kenya, we operate six regional offices located in Brazil, Cameroon, India, Indonesia, Kenya, and Malawi, and conduct research in eighteen other countries around the developing world.

We receive our funding from over 50 different investors. Our current top ten investors are Canada, the European Union, Finland, Ireland, the Netherlands, Norway, Denmark, the United Kingdom, the United States of America and the World Bank.

FOREST & LANDSCAPE

FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

Forest & Landscape Denmark (FLD) is part of the Faculty of Life Sciences at the University of Copenhagen. FLD is the national centre for research, education and advisory services within the fields of forest and forest products, landscape architecture and landscape management, urban planning and urban design. FLD houses a major programme on use, improvement and conservation of genetic resources of trees and forests in the tropics. This programme is supported by Danida and includes the former Danida Forest Seed Centre.

Molecular Markers for Tropical Trees

by
Roeland Kindt
Alice Muchugi
Ole Kim Hansen
Hillary Kipruto
Jane Poole
Ian Dawson
Ramni Jamnadass

Edited by: Ian Dawson and Ole Kim Hansen

Statistical Analysis of Dominant Data

Citation: Kindt R, Muchugi A, Hansen OK, Kipruto H, Poole J, Dawson I, Jamnadass R. 2009. Molecular Markers for Tropical Trees: Statistical Analysis of Dominant Data (eds. Dawson I, Hansen OK). ICRAF Technical Manual No. 13. Nairobi: World Agroforestry Centre, Frederiksberg: Forest and Landscape Denmark.

Titles from the Technical Manual series synthesise the outcomes and practical implications of agroforestry research activities and projects that are ready for scaling up.

ICRAF Technical Manual No. 13

ISBN: 978-92-9059-262-4

Editors: Ian Dawson and Ole Kim Hansen
Cover photos: Jan Beniest and Tony Simons, © World Agroforestry Centre
Design and layout: Kris Vanhoutte

Suggestions for additions and revisions for subsequent editions of this guide are welcome. Please write to: The molecular laboratory, World Agroforestry Centre, P.O. Box 30677 - 00100, Nairobi, Kenya; email: icrafmolecularlab@cgiar.org

World Agroforestry Centre
United Nations Avenue, Gigiri,
P. O. Box 30677-00100, Nairobi,
Kenya
Tel: (+254 20) 722 4000
Fax: (+254 20) 722 4001
Email: icraf@cgiar.org
www.worldagroforestry.org

Forest and Landscape Denmark
Rolighedsvej 23
DK-1958 Frederiksberg C
Denmark
Tel: (+45 3533) 1500
Fax: (+45 3533) 1508
Email: sl@life.ku.dk
www.en.sl.life.ku.dk

Acknowledgements

We are grateful to the Programme for Cooperation with International Institutes (SII), the Education and Development Division of the Netherlands' Ministry of Foreign Affairs, for contributing funding to the publication of this guide and to Jan Beniest for stimulating this effort. We equally appreciate the donors who have funded molecular research at ICRAF, especially the Swedish International Development Cooperation Agency (SIDA), the United States Agency for International Development (USAID), the Danish International Development Agency (Danida), the Department for International Development UK (DFID) and the European Union (EU). We further acknowledge the support of Forest and Landscape Denmark, Kenyatta University and the Scottish Crop Research Institute. We are also grateful to the participants and resource persons at a SII-funded workshop on molecular markers titled *Agroforestry and Tree Genetics: Making Markers Meaningful* who tested a beta version of this guide and provided important feedback on content: Achille Ephren Assogbadjo, Abel Gari, Demissew Serte Desta, Bi Irie Arsene Zoro, Cosmas Sorngmenenye Abengmeneng, Daniel Aninagyei Ofori, Otto George Dangasuk, Eric Bertrand Kouam, Joseph Mwangi Muchua, George Edward Mamati, Josiah Chemulanga Chiveu, Francis Wachira, Victor Wasike, Cecilia Mbithe Mweu, Weston Fredrick Mwase, Yogeshwarsingh Parmessur, Rami Sirelkhatem Habeballa Ahamed, Pauline Aluka, Samson Gwali, Judith Nantongo, Sammy Olal, Eric Ndenga, Noel Onyango Ochieng, Kennedy Owuor Olale and Joanne Russell. Finally, our thanks to Nelly Mutio, Rita Mulinge, Margaret Hanson and Agnes Were, who provided administrative and further support in the development of this guide.

Contents

Contributors

Roeland Kindt
Community ecologist
World Agroforestry Centre
Email: r.kindt@cgiar.org

Alice Muchugi
Population geneticist
World Agroforestry Centre and Department of Biochemistry and Biotechnology,
Kenyatta University, Nairobi, Kenya
Email: a.muchugi@cgiar.org

Ole Kim Hansen
Assistant Professor
Forest and Landscape Denmark, Faculty of Life Sciences, University of Copenhagen, Denmark
Email: OKH@life.ku.dk

Hillary Kipruto
Biostatistician
World Agroforestry Centre
Current address: The World Health Organization, Nairobi, Kenya
Email: kiprutoh@gmail.com

Jane Poole
Applied Statistician, Co-Lead ILRI-ICRAF Research Methods Group
World Agroforestry Centre and International Livestock Research Institute (ILRI),
Nairobi, Kenya
Email: j.poole@cgiar.org

Ian Dawson
Research Fellow
World Agroforestry Centre
Email: iankdawson@aol.com

Ramni Jamnadass
Molecular geneticist and team leader of ICRAF Global Research Project 1
World Agroforestry Centre
Email: r.jamnadass@cgiar.org

Why a guide, and for whom?

Purpose and audience

In the last fifteen years, there has been an enormous increase worldwide in the use of molecular markers to assess genetic diversity in trees. These approaches are able to tell us how genetic variation is structured in natural, managed and cultivated tree stands, and they can provide significant insights into the defining features of different species.

Molecular techniques can provide more detailed information than phenotypic studies of genetic variation are able to do, knowledge that can then, in theory, be applied to devise more optimal management strategies for trees within natural and human landscapes, in order to benefit users and the environment. Proper genetic management is crucial as trees are planted to combat poverty, fight malnutrition, provide medicines and fulfil other needs, such as the mitigation of climate change and the prevention of soil degradation. As very little information has been available on how genetic variation is structured in the majority of tropical trees, modern molecular methods provide clear opportunities for the quantification of diversity.

Despite evident potential, a survey of the literature indicates that the implementation of practical, more optimal management strategies based on results from molecular marker research is to date very limited for tropical trees, both in farmland and forest settings. To explore why this is the case, in 2006 the World Agroforestry Centre (ICRAF) undertook a survey of molecular laboratories in low-income countries in the tropics. Problems in application highlighted by surveyed scientists included a lack of knowledge on the different procedures available for molecular genetic studies, and the absence of guidance on how best to apply methods specifically to tropical trees.

To help meet these needs, in 2008 ICRAF published a practical protocol guide on molecular marker methods for tropical trees (Muchugi et al. 2008a; available for

download from ICRAF's web site and from the CD-ROM on which this statistical guide is provided). This practical protocol guide addresses the basics of population genetics (e.g., the processes of mutation, migration, recombination, selection and drift), the issues to consider before embarking on molecular marker studies, the design of field sampling strategies for tree species, and the strengths and disadvantages of different laboratory techniques. The guide also provides detailed practical protocols for various marker methods.

To help further in addressing identified needs, in 2008 ICRAF developed a one week 'training of trainers' course for African scientists titled *Agroforestry and Tree Genetics: Making Markers Meaningful*. This course was designed to explore the links between molecular marker methods and global production and conservation challenges for trees in the context of smallholder agroforestry systems. The materials developed for this course are available as a CD-ROM from ICRAF's training unit (see ICRAF's contact address at the start of this guide), and we encourage all those interested in the practical application of markers on trees to make use of these resources.

Another constraint that ICRAF's 2006 survey identified for the proper application of molecular markers is the effective handling and analysis of data sets once they have been generated. This current guide has been designed to address this need. It has been created especially for students (MSc, PhD) and other researchers in developing countries that find themselves isolated from their peers and – when faced with an apparently bewildering array of options – find it difficult to settle on appropriate methods for analysis. Most benefit will be obtained from this guide if it is used together with the companion volume on practical protocols, and so we recommend that scientists read both before proceeding further.

What this guide does, and does not, cover

This guide deals with the 'population genetic' analysis of dominant markers and does not consider how to analyse information from co-dominant methods (although we intend that this will be the subject of a later publication). The meanings of the terms 'co-dominant' and 'dominant' were addressed in the practical protocol guide and so we provide only a brief explanation below.

For co-dominant markers – and when dealing with a diploid organism – each individual's score at a particular locus consists of two numbers or characters, one for each chromosome state. Complete genotypic information, including on heterozygosity, is available, and allele frequency distributions among sampled individuals are provided directly. Dominant markers are however different. In this case, each locus (or band position, we use the terms interchangeably here) is scored in a 'binary' fashion, as 'product presence' [1] or 'product absence' [0]. For a diploid individual, only one number can be recorded at a locus, even though heterozygote 'present-absent' [1, 0] states may well be present, as these heterozygotes are indistinguishable from homozygote 'present-present' [1,1] conditions. Full genotypic information is clearly not available from dominant markers: instead, population allele frequencies must be estimated indirectly, based on assumptions described in this guide.

The inability to score heterozygotes restricts the types of data analysis that are possible for dominant markers. As trees are generally highly heterozygous, out-crossed organisms, the inability to directly observe heterozygote conditions would therefore appear to be a distinct disadvantage. Certainly, in an ideal world, co-dominant markers would be the method of choice for assessing genetic variation in tree species. However, many of the molecular marker methods most used on tropical trees, especially in laboratories in low income-nations, are dominant approaches. One reason for this apparent discrepancy is that, unlike co-dominant techniques, dominant methods do not normally require prior sequence knowledge on the organism being tested, information that is often not available for tropical trees. In addition, dominant marker techniques are often cheaper and easier to apply than co-dominant approaches. They can reveal data at many positions of the genome and on many individuals quickly.

A strong reliance on dominant markers in laboratory analysis – especially currently the use of the amplified fragment length polymorphism (AFLP) technique – explains the focus of this guide. Often, analysis of co-dominant data will be possible with the same software packages referred to in this publication, but extra or alternative steps in analysis will frequently be required, in addition to those described here.

Analysis of data involves describing the variation revealed by molecular markers at individual, population and other levels of geographic or taxonomic structure. It involves calculating the relationships between different levels of structure and

expressing these in ways that are clear – numerically and, ideally, visually – to other researchers and (if data are to be used practically) to field managers. Fundamental to analysis are calculations of genetic diversity and genetic differentiation, and both are addressed here. Methods such as cluster analysis and ordination are also key to visualise results, and, again, both of these approaches are covered in this guide.

As subsequent sections of the guide will show, there are a multitude of software packages available for analysis. These packages tend to have a mix of common and unique functions. When functions are in common, which software to use can be a question of user preference – users will tend to use the package they were first introduced to, unless there is good reason for change. We however encourage users to evaluate the assumptions behind the approaches used to make calculations in different packages, to determine whether they are really appropriate for addressing the question at hand. Of the packages discussed here, for newcomers we would suggest that GenAlEx (Peakall and Smouse 2006) provides a good starting point for building skills and confidence. Users can then move on to other programs with more sophisticated methods.

Whatever the software used, there is no substitute for understanding the basic procedures involved in analysis. Furthermore, rather than the complexity of the analysis, it is the interpretation of results in the context of the biological processes shaping variation that is fundamental in determining the relevance of molecular marker studies.

Quality issues

The old adage 'garbage in, garbage out' most definitely applies during the analysis of molecular marker data, and adoption of the most sophisticated approaches to analysis will not provide meaningful results if the initial quality of data is low. Molecular genetic studies can only ever be as good as the collection strategy adopted during field sampling. A poor sampling strategy will mean that it is impossible to say anything meaningful about the biology of the species in question. Worse, conclusions that are inaccurate and possibly misleading may be the result.

Based on this concern and the observation that molecular genetic differentiation between natural populations of most tree species is low, various scientists have suggested sampling a minimum of 30 individuals to represent a tree population. While we would not be this prescriptive – sample numbers will depend on the question being addressed, and it is often not practically possible to include so many individuals when assessing a range of stands – we raise this issue to illustrate the importance of comprehensive sampling. This topic was addressed in the practical protocol guide, where further information can be found.

A second element of quality relates to how genetic variation is first identified in the laboratory, as if this is done incorrectly it can lead to problems in later interpretation. In laboratory studies, researchers will almost always first carry out preliminary screens for variation on a 'test panel' of their samples (often, 8 or 16 individuals). This is in order to identify those primers (in polymerase chain reaction [PCR] studies) that will work well in revealing variation in wider collections of material. While this is a useful approach, it does create potential for bias. This can be understood by considering a situation where primers are screened for variation in only one of many populations. Naturally enough, this will select for those primers that reveal high variation in the particular tested stand. Due to population differentiation, however, there is no guarantee that the same primers will respond equally well in revealing variation in other stands, even though, intrinsically, these are as diverse as the first population. In this instance, screening (or ascertainment) bias will result in one particular stand artificially appearing more polymorphic than others, and a false picture of the structuring of diversity will be the result. To prevent bias, the test panel must be properly constructed to be representative of the entire collection being studied. Again, further information on this topic can be found in the practical protocol guide.

Finally, data quality also depends on how well the laboratory protocol has been optimised for the species being tested. Researchers familiar with different marker techniques know well how the quality of results can vary between species and even between different experiments or 'runs' on the same species. It is important then to invest time in optimising laboratory approaches so that the results obtained are as clear as possible, with polymorphisms well resolved from each other and easy to score. Attention to this may be of more importance than the details of the methods that are subsequently used to analyse results.

How this guide is structured

This guide is divided into four main sections, as described below. The meanings of the terms used in the below descriptions will become evident in the main body of the guide.

Part 1 (Chapter 1) describes the steps involved in preparing data for analysis. Covered in this section are the basics of scoring, storing and handling data, including getting data into the right format for subsequent analyses.

Part 2 (Chapters 2 to 5) relates different methods for analysis at the population level. It considers the measurement of diversity at each individual locus and across loci. It explains how to calculate genetic distances between populations, and how to visualise distances in summary form by clustering and ordination techniques. It assumes that the organism being analysed is diploid.

Part 3 (Chapters 6 and 7) considers different methods for analysis at the individual level. Included here is the measurement of genetic distances between individuals and the expression of these distances through ordination.

Part 4 (Chapters 8 and 9) relates some additional and more modern techniques for analysis not described in previous sections. Included here is the analysis of molecular variance approach (known as AMOVA) and the STRUCTURE method.

Appendix I provides a range of mathematical formulae that – in order to make the body of the guide more readable – are not included in the main text.

Appendix II provides detailed information on how to install the various software packages referred to in this guide and how to format data for analysis in them. It also gives similar information for some others programs that are not considered in detail in the guide, but which users' may want to explore further.

Appendix III gives step-by-step instructions on how to do analyses in particular software packages. Instructions are given following the same layout of chapters as in the main body of the guide.

The CD-ROM accompanying this guide provides a collection of input data spreadsheets formatted for different software packages and types of analysis. Also included here are the corresponding results files produced by different programs. The purpose of these files is to allow users of this guide to experiment directly with different approaches to analysis. These input files can be modified through cutting and pasting to include users' own data sets. For each of the subsequent sections of this guide, the content of the CD-ROM should be investigated to see if it contains relevant files for further exploration. As with Appendix III, the folder structure on the CD-ROM follows the layout of the chapters of the main body of the guide.

Throughout this guide, we base test analyses on an example AFLP data set from the African medicinal tree *Warburgia ugandensis*. This data was collected by one of the authors (Alice Muchugi) during her PhD studies, as part of a wider study on *Warburgia* species (as reported in Muchugi et al. 2008b).

We assume that users of this guide will have access to Microsoft Excel for the initial input and formatting of data.

References

Muchugi A, Kadu C, Kindt R, Kipruto H, Lemurt S, Olale K, Nyadoi P, Dawson I, Jamnadass R (eds. Dawson I, Jamnadass R) (2008a) Molecular Markers for Tropical Trees: A Practical Guide to Principles and Procedures. ICRAF Technical Manual No. 9. The World Agroforestry Centre, Nairobi, Kenya.

Muchugi A, Muluvi GM, Kindt R, Kadu CAC, Simons AJ, Jamnadass RH (2008b) Genetic structuring of important medicinal species of genus *Warburgia* as revealed by AFLP analysis. *Tree Genetics and Genomes* 4: 787-795.

Peakall R, Smouse PE (2006) GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6: 288-295.

Part I

Data preparation

Chapter 1. Getting data ready for analysis

Key points

This section describes the steps involved in preparing data for analysis. Topics covered are the basics of scoring and storing data, including getting information into the right format for use.

We suggest that data spreadsheets should be generated using a standard format, with rows representing individuals and columns representing loci. Rows can also contain geographic data on individuals, such as the populations and/or regions that they come from.

An illustration of data formatting is based on an AFLP data set collected on the African medicinal tree *W. ugandensis*. Appendix II shows how this data set can be imported into the various software packages used in this guide.

Data spreadsheets should be printed and kept in paper format for archiving purposes.

1.1. Scoring and storing data

For dominant data, each locus is scored for an individual in a 'binary' way, as [1] or [0] (presence or absence of a product, respectively). If an observation is missing because, e.g., a PCR amplification has not worked well, then that data point should be recorded in some other way than as [0], e.g., as [NA] (not available). Such missing scores will need to be treated differently from product absences in analysis, as we describe in subsequent sections.

Usually, the best way to deal with data is to input it straight into an electronic spreadsheet. This spreadsheet should have been set up in advance by the user and should contain all the relevant column and row labels that describe data points. Labels should give information on the population[1]/location that an individual comes from, and provide a unique sample identifier for each tested accession ('geographic data'). Additional geographic data could include higher level groupings of populations, e.g., the regions and/or countries that they come from. On the same axis, additional labels could, if multiple species are being tested, divide data into taxonomic groups. These extra levels of stratification can be useful in hierarchical analyses such as the analysis of molecular variance (AMOVA, see Chapter 8). On the alternative axis of the spreadsheet, labels should provide information on the particular locus that is being scored – generally the name of the primer/primers used to reveal that band and a unique locus identifier ('molecular data').

In generating spreadsheets, we suggest that a common format is used in which rows always represent 'geographic data' (unique sample identifier, population, region, species, etc.) and columns always represent 'molecular data' (unique locus identifier, primer[s], etc.). Each row thus represents the overall profile of a particular individual for all loci, and each column corresponds to the genetic scores for all individuals at a given locus. The number of rows therefore corresponds to the number of individuals tested, and the number of columns to the number of loci scored. An example of what a spreadsheet looks like is given in section 1.2 below. Once this matrix has been generated, the scores of individual accessions at individual loci can be inserted. The use of Microsoft Excel is ideal for data entry and links seamlessly with the 'bolt on' GenAlEx package for basic data analysis.

[1] Throughout the guide, we use the term population to refer to a collection of individuals from the same location.

Data spreadsheets are the basic building blocks for all subsequent analyses of results. In addition, they are essential for 'archiving' information. For this last purpose, as well as being backed up electronically, spreadsheets should be printed out and stored in a safe place. Ideally, a more detailed description of the experiment from which data were collected should also be printed and attached to the hard copy of the spreadsheet. This annex can give much more information on the experiment in question. This might include information on the purpose of the study, the locations of populations on a map, and the geographic coordinates of individual trees. It may also include data on human management of populations, information on the gender of individual trees (if a dioecious species), and measurements on diameter, fruiting, phenotype, etc.

The proper archiving of data allows information to be returned to in the future, perhaps when more data become available or new ways of interpretation are possible.

1.2. An example data set

Throughout this guide, we base test analyses on an example AFLP data set from the African medicinal tree *Warburgia ugandensis*. The bark extract of this species is used as an anti-malarial treatment. It has been suggested that the chemical composition of the active components in the bark may be different in different stands, and certain populations are also subject to conservation threats because of over-exploitation. *Warburgia ugandensis* is one of the key medicinal tree species that ICRAF is interested in, and has therefore merited molecular marker investigation.

The test *Warburgia* data set is based on 20 individuals sampled from each of five different populations. These populations are Kibale (in Uganda), Kitale (Kenya), Laikipia (Kenya), Lushoto (Tanzania) and Masai Mara (Kenya). Two of these populations were collected to the west of the Rift Valley (Kibale and Kitale) and three to the east (Laikipia, Lushoto and Masai Mara), creating a regional hierarchy between stands that is important for some but not all analyses. Data were collected for 185 AFLP loci or bands. Data are therefore represented by a matrix of 100 individuals (rows) by 185 loci (columns), equaling 18,500 points of information in total (for further information on sampling, laboratory methods and data collection on *Warburgia*, see Muchugi et al. 2008b).

A subset of this data – for the first 45 individuals tested and the first five loci scored – is shown in the spreadsheet below (Table 1.1). In this instance, there was no missing data at loci and so the code [NA] was not required during scoring. As can be seen, the first column of the data is a unique identifier for each individual that also includes a population abbreviation, the second column gives the population, and the third the region from which a population came.

It is worth noting that for unique identifiers it is better to use the labels '01', '02', '03' (or '001', '002', '003') rather than '1', '2', '3', etc., since this allows proper ordering of samples when more than nine individuals are present in a population. Otherwise, when sorting data in spreadsheet manipulations, '10' will order after '1' rather than after '9'. (The same applies when labelling loci.) By including a population reference in the unique individual identifier, it is easier to keep track of geographic origins in subsequent analyses based on individuals. Note that, in this instance, locus identifiers are given as single numbers only and do not extend to include the combination of primers used to detect the AFLP (this would have been another option).

The entire *Warburgia* data set is provided in various formats on the CD-ROM that accompanies this guide (basic data set as *warburgiabase.xls* or *warburgiabase.txt*). Guidelines for formatting data for different software packages are provided in Appendix II.

Table 1.1. A subset of AFLP data (for 45 individuals and 5 loci) collected for *Warburgia ugandensis*, showing the appropriate format for inputting results into a spreadsheet. [1] represents product presence and [0] product absence during scoring.

Individual	Population	Region	Locus001	Locus002	Locus003	Locus004	Locus005
Kit01	Kitale	west	0	0	0	0	0
Kit02	Kitale	west	0	1	0	1	0
Kit03	Kitale	west	0	0	0	1	0
Kit04	Kitale	west	0	0	0	1	0
Kit05	Kitale	west	0	0	0	1	0
Kit06	Kitale	west	0	0	0	1	0
Kit07	Kitale	west	0	1	0	1	0
Kit08	Kitale	west	0	0	0	1	0
Kit09	Kitale	west	0	0	0	0	0
Kit10	Kitale	west	0	0	0	1	0
Kit11	Kitale	west	0	0	0	1	0
Kit12	Kitale	west	0	0	0	1	0
Kit13	Kitale	west	0	0	0	0	0
Kit14	Kitale	west	0	0	0	0	0
Kit15	Kitale	west	0	0	0	0	0
Kit16	Kitale	west	0	0	0	0	0
Kit17	Kitale	west	0	0	0	0	0
Kit18	Kitale	west	0	0	0	0	0
Kit19	Kitale	west	0	0	0	0	0
Kit20	Kitale	west	0	0	0	0	0
Kib01	Kibale	west	0	0	0	0	0
Kib02	Kibale	west	0	0	0	0	0
Kib03	Kibale	west	0	0	0	0	0
Kib04	Kibale	west	0	0	0	1	0
Kib05	Kibale	west	0	0	0	0	0
Kib06	Kibale	west	0	0	0	0	0
Kib07	Kibale	west	0	1	0	1	0
Kib08	Kibale	west	0	0	0	1	0
Kib09	Kibale	west	0	0	0	1	0
Kib10	Kibale	west	0	1	0	1	0
Kib11	Kibale	west	0	0	0	0	0
Kib12	Kibale	west	0	0	0	0	0
Kib13	Kibale	west	0	0	0	0	0
Kib14	Kibale	west	0	0	0	1	0
Kib15	Kibale	west	0	0	0	0	0
Kib16	Kibale	west	0	0	0	0	0
Kib17	Kibale	west	0	1	0	1	0
Kib18	Kibale	west	0	0	1	1	0
Kib19	Kibale	west	1	0	0	0	0
Kib20	Kibale	west	1	0	0	1	0
Lai01	Laikipia	east	0	1	0	1	0
Lai02	Laikipia	east	0	1	0	1	0
Lai03	Laikipia	east	0	0	0	0	0
Lai04	Laikipia	east	0	0	0	0	0
Lai05	Laikipia	east	0	0	0	1	0

1.3. References

Muchugi A, Muluvi GM, Kindt R, Kadu CAC, Simons AJ, Jamnadass RH (2008b) Genetic structuring of important medicinal species of genus *Warburgia* as revealed by AFLP analysis. *Tree Genetics and Genomes* 4: 787-795.

Part 2

Analysing data at the population level

Chapter 2. Measuring diversity

Key points

This section considers the measurement of genetic diversity for individual loci in populations, and then how information is summed across loci.

For dominant markers, the standard methods for calculating diversity rely on estimating allele frequency distributions from product frequency distributions. This involves making assumptions about genetic structure within populations. There are a number of ways in which this can be done. We recommend using a Bayesian approach.

Once allele frequencies have been estimated for populations, Nei's unbiased diversity statistic (*H*) is the standard way to express the level of variation. Once estimates have been calculated for individual loci, the arithmetic mean of these values provides an overall estimate of diversity for a population.

2.1. Estimating allele frequencies from product frequencies

Genetic diversity can be quantified in terms of the richness and evenness of distribution of molecular marker variation within populations (or, indeed, within other defined groups of individuals). Estimating the genetic diversity of a population from a dominant marker normally involves first converting product frequencies to allele frequencies before further analysis is undertaken. Remembering that dominant markers are unable to give complete genotypic information (as they do not discriminate between heterozygote 'present-absent' [1, 0] and homozygote 'present-present' [1, 1] states; see the Introduction), this means making assumptions about the structure of genetic variation within populations.

To estimate allele frequencies from product frequencies, it is often assumed that populations are at Hardy-Weinberg equilibrium, which means that they function as groups of individuals that mate completely randomly with each other (we discuss this assumption further below). In this case, allele frequencies for a diploid organism are determined by the equation $1 = p^2 + 2pq + q^2$, where p and q are the frequencies of the presence and absence alleles (rather than products), respectively. Because q^2 is the frequency of occurrence of product absence in a population, the value for q and then p $(1 - q)$ can be generated accordingly. This approach, known as the 'square-root' method, can be illustrated with our example *Warburgia* data set (Table 2.1; see further information in Chapter 1).

Table 2.1. Allele frequencies for the first 10 AFLP loci for the Kitale population, estimated from marker frequencies using the square-root method (*N* = number of individuals sampled from the population). Results were obtained with the GenAlEx 6.2 software.

Locus	*N*	Product frequency (*f*)	*p*, presence allele frequency ($p = 1 - q$)	*q*, absence allele frequency ($q = \sqrt{(1 - f)}$)
Locus001	20	0.0	0.000	1.000
Locus002	20	0.1	0.051	0.949
Locus003	20	0.0	0.000	1.000
Locus004	20	0.5	0.293	0.707
Locus005	20	0.0	0.000	1.000
Locus006	20	0.2	0.106	0.894
Locus007	20	0.0	0.000	1.000
Locus008	20	0.0	0.000	1.000
Locus009	20	0.0	0.000	1.000
Locus010	20	0.3	0.163	0.837

Table 2.2. Allele frequencies for the first 10 AFLP loci for the Kitale population, estimated from marker frequencies using a Bayesian method with non-uniform priors (*N* = number of individuals sampled from the population). Results were obtained with the AFLP-SURV 1.0 software.

Locus	*N*	Product frequency	*p*, presence allele frequency	*q*, absence allele frequency
Locus001	20	0.0	0.003	0.997
Locus002	20	0.1	0.054	0.946
Locus003	20	0.0	0.003	0.997
Locus004	20	0.5	0.293	0.707
Locus005	20	0.0	0.003	0.997
Locus006	20	0.2	0.108	0.892
Locus007	20	0.0	0.003	0.997
Locus008	20	0.0	0.003	0.997
Locus009	20	0.0	0.003	0.997
Locus010	20	0.3	0.165	0.835

Because trees are generally out-crossing they are more likely to mate randomly within a population than other types of plants are, and the assumption of Hardy-Weinberg equilibrium is therefore more valid. In practice, however, few populations are likely to be fully at Hardy-Weinberg equilibrium, and this approximation will therefore result in some bias in allele frequency estimation. Other approaches have been developed in efforts to minimise this bias, and one such method is based on Bayesian estimation techniques (Zhivotovsky 1999). A Bayesian method that uses non-uniform priors[1] can produce good estimates of allele frequencies (Bonin et al. 2007). This therefore is the approach that we recommend. Results based on estimating allele frequencies in this manner are shown in Table 2.2. Comparing the results shown in Tables 2.1 and 2.2, it can be seen that values are similar in magnitude, although with the Bayesian method estimates are never zero.

[1] Bayesian methods generate *posterior probabilities* that indicate the probability that a certain hypothesis is true. Bayes' rule is used to combine *prior probabilities* with the information contained in data to generate the posterior predictions. Zhivotovsky (1999) developed Bayesian estimation methods for different types of prior probabilities for the distribution of allele frequencies: (i) a method based on *uniform* prior distribution of allele frequencies and (ii) methods based on *non-uniform* prior distribution of allele frequencies (see Appendix I). Approaches based on non-uniform priors are expected to provide the most reliable estimates.

Alternative methods to estimate allele frequencies based on some degree of inbreeding (selfing) are also available (see Box 2.1). For interested readers, these and other more sophisticated methods for estimating allelic frequency can be explored in the references at the end of this chapter and through the electronic files on the CD-ROM accompanying this guide. For example, methods that include only the most informative markers (that have product frequencies within a certain range of values) have been developed (Nybom 2004). These methods are of potential value and we encourage their further consideration, even though they have not been applied widely in the analysis of data (perhaps because of a perception of complexity).

2.2. Calculating diversity from allele frequencies

Once allele frequencies have been estimated, the normal means by which to express diversity is through calculating *H*, the Nei diversity statistic (Nei 1978). Either the standard statistic, or an unbiased estimate that corrects for small population sizes, can be used (Nei developed one method of correction, Lynch and Milligan [1994] an alternative; see Appendix I). We recommend that the unbiased estimate is used. In practice it makes little difference if sample sizes for all populations are 50 or more (see more below). The standard and unbiased values of *H* for the same loci listed in Table 2.2 are given in Table 2.3.

Table 2.3. Standard and unbiased Nei diversity (*H*) estimates for the first 10 AFLP loci from the Kitale population, based on allele frequencies estimated using a Bayesian method (see Table 2.2). Allele frequencies were obtained with the AFLP-SURV 1.0 software.

Locus	*p*, presence allele frequency	*q*, absence allele frequency	Standard *H*	Unbiased *H*
Locus001	0.003	0.997	0.006	0.006
Locus002	0.054	0.946	0.102	0.105
Locus003	0.003	0.997	0.006	0.006
Locus004	0.293	0.707	0.414	0.425
Locus005	0.003	0.997	0.006	0.006
Locus006	0.108	0.892	0.192	0.197
Locus007	0.003	0.997	0.006	0.006
Locus008	0.003	0.997	0.006	0.006
Locus009	0.003	0.997	0.006	0.006
Locus010	0.165	0.835	0.275	0.283

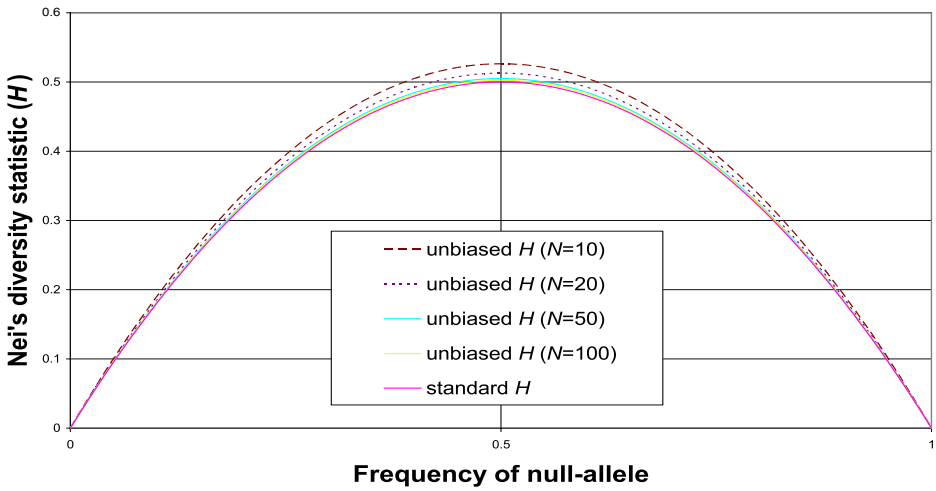


Figure 2.1. The relationship between *H* and allele frequency, and the effect of varying populations sizes on estimates of diversity, using Nei's methods. *H* is at its maximum value when allele frequencies are balanced. The larger the number of individuals sampled then the less difference it makes when correcting for population size in unbiased estimates. The data and figure were created in Microsoft Excel.

As can be seen from Table 2.3, the more balanced the allele frequencies at a locus, then the greater the value of *H*, up to a maximum of 0.5 for a dominant marker with equal frequencies of the two allele states. The relationship between *H* and allele frequency is shown in Fig. 2.1. The differences between standard and unbiased estimates as a function of population size are also shown (the difference between the two estimates is low for larger populations). Other methods of calculating diversity can be explored in Appendix I and through the electronic files on the CD-ROM. Again, we encourage the users of this guide to consider the application of these other approaches.

2.3. Summarising diversity across loci

Once individual locus estimates of diversity have been obtained, normal practice (unless interested in specific markers for some reason, perhaps because selection is expected in a known mapped region of the genome) is to summarise information across a group of loci. This then gives an overall picture of the level of genetic

variation within a population. Data can be summarised simply by calculating the arithmetic mean of standard *H* or unbiased *H* values. In the case of the loci listed in Table 2.3, this provides mean standard *H* and unbiased *H* values of 0.102 and 0.105, respectively. Because the number of individuals sampled for analysis is relatively high ($N = 20$), the difference between the two estimates is only marginal.

It is obvious from our *Warburgia* example that individual locus estimates can vary greatly, and so a relatively large number of loci – certainly more than the 10 used in our test analysis – are required to provide a reasonable overall estimate of diversity. Estimates based on all 185 AFLP loci scored for the five *W. ugandensis* populations that constitute our example data set (this number of loci is likely to be more than sufficient to provide good overall estimates) are given in Table 2.4. Note that the difference in estimates between certain populations (e.g., Kitale and Laikipia) is small and is unlikely to be statistically significant (a method to test for this is given in Box III.1 of Appendix III).

An alternative means to summarise population diversity across loci would simply be to estimate the fraction or percentage of markers that are polymorphic. In such summations, researchers will sometimes set certain criteria for considering whether a locus is polymorphic or not (e.g., a product must have a minimum frequency of 0.05 or 0.01 before being considered polymorphic). We however prefer not to specify limits, but rather to score any level of product variation within a population as polymorphic during summations. (Note that in this instance we are referring to product rather than allele frequencies. The issue of Bayesian procedures providing non-zero allele frequencies from zero value product frequencies in populations is therefore not relevant.)

Table 2.4. Standard and unbiased Nei diversity (*H*) estimates for all 185 AFLP loci for each of five *W. ugandensis* populations, based on allele frequencies estimated using a Bayesian method. Allele frequencies were obtained with the AFLP-SURV 1.0 software.

Method	Kitale	Kibale	Laikipia	Mara	Lushoto
Standard *H*	0.147	0.140	0.151	0.132	0.117
Unbiased *H*	0.155	0.148	0.159	0.138	0.123

As noted above, different methods are available for estimating allele frequencies and diversity at individual loci, and of course this influences the results then obtained when summing information across loci. We encourage readers to explore this further through the equations in Appendix I and the electronic files on the CD-ROM. If the results from different calculation methods do not cause rank differences between populations in diversity estimates, then the approach used is unlikely to be of significant concern. If different methods give significant rank differences, however, then more thought to the most appropriate approach needs to be given.

2.4. References

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology* 16: 3737-3758.

Kremer A, Caron H, Cavers S, Colpaert N, Gheysen G, Gribel R, Lemes M, Lowe AJ, Margis R, Navarro C, Salgueiro F (2005) Monitoring genetic diversity in tropical trees with multilocus dominant markers. *Heredity* 95: 274-280.

Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology* 3: 91-99.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583-590.

Nybom H (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology* 13: 1143-1155.

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* 8: 907-913.

2.5. Suggested software

A wide range of software packages produce diversity estimates (e.g., GenAlEx, PopGene, AFLP-SURV, FAMD, TFPGA), but we recommend using AFLP-SURV or FAMD since these packages allow Bayesian estimations methods for allele frequencies and should therefore provide the most reliable results (see Appendix III).

Box 2.1. Estimating allele frequencies and diversity based on a degree of selfing

Kremer et al. (2005) suggested undertaking sensitivity analyses on dominant data diversity estimates assuming a degree of inbreeding in tested populations when calculating allele frequencies (that is, when F_{IS}, the inbreeding coefficient, > 0).

They suggested that results be compared for F_{IS} = 0 (the standard estimate) and F_{IS} = 0.1. If population diversity estimates rank the same for both values of F_{IS}, then more confidence can be assigned to results.

Using our example *Warburgia* data set, we therefore calculated diversity for populations according to different values of F_{IS}, as shown in Table 2.5.

Our results show that for both 'square-root' and Bayesian estimation methods, the ranking of population diversity is not always the same when F_{IS} > 0: rather, the ranking of Kitale and Kibale as second or third depends both on the method used and assumptions of inbreeding (see values in bold where Kibale is more diverse than Kitale). Based on this difference in ranking, we conclude that caution should be exercised when comparing estimates.

Table 2.5. Unbiased Nei diversity (*H*) estimates for all 185 AFLP loci for each of five *W. ugandensis* populations, based on 'square-root' and Bayesian allele frequency estimation methods and different levels of inbreeding (F_{IS}). Results were obtained with the AFLP-SURV 1.0 software.

Method to estimate allele frequencies	Assumed F_{IS}	Laikipia	Kitale	Kibale	Mara	Lushoto
Bayesian	0.0	0.15570	0.15305	0.14457	0.13618	0.12112
	0.1	0.15308	0.14790	0.14360	0.13245	0.11767
	0.5	0.16258	**0.14116**	**0.15544**	0.13453	0.12028
Square-root	0.0	0.14535	0.14005	0.13715	0.12343	0.11140
	0.1	0.14744	**0.13869**	**0.13944**	0.12448	0.11182
	0.5	0.15523	**0.13347**	**0.14829**	0.12772	0.11447

Chapter 3. Measuring genetic distance between populations

Key points

This chapter explains how to calculate genetic distances between pairs of populations.

For dominant markers, the calculation of distances normally relies, as when calculating genetic diversity, on estimates of allele frequency distributions from population product frequencies.

Once allele frequencies have been estimated, we suggest the use of Nei's unbiased measure to calculate genetic distances between populations.

Once obtained, genetic distances can be visualised by clustering and ordination analyses, as explained in Chapters 4 and 5.

3.1. Calculating genetic distances from allele frequencies

Once allele frequencies have been calculated at loci (see Chapter 2), these can be compared between pairs of populations to generate a genetic distance matrix among stands. Nei's (1978) measure of genetic distance is one of the most used methods, and this can be calculated as a standard or unbiased value (just as Nei's diversity estimate can; see Chapter 2), depending on whether small population sizes are corrected for. We recommend that the unbiased method be generally adopted. Unbiased values for Nei's genetic distance for the five *W. ugandensis* populations of our example data set, for all 185 AFLP loci scored (see further information in Chapter 1), are given in Table 3.1.

As can be seen from this matrix, the distance of a population to itself is always zero and each matrix is symmetrical: that is, the values above a diagonal from top left to bottom right through the matrix correspond with the values in the lower part of the matrix. As a result, the matrix shown in Table 3.1 can be expressed in the triangular format shown in Table 3.2, in which only four columns and rows represent results.

Table 3.1. Unbiased Nei distance estimates (calculated according to Lynch and Milligan 1994) between five populations of *W. ugandensis*, based on allele frequencies estimated using a Bayesian method. Results were obtained with the AFLP-SURV 1.0 software.

Unbiased Nei distance	Kitale	Kibale	Laikipia	Mara	Lushoto
Kitale	0.0000	0.0428	0.0920	0.1002	0.1071
Kibale	0.0428	0.0000	0.0539	0.0676	0.0594
Laikipia	0.0920	0.0539	0.0000	0.0062	0.0104
Mara	0.1002	0.0676	0.0062	0.0000	0.0060
Lushoto	0.1071	0.0594	0.0104	0.0060	0.0000

Table 3.2. Unbiased Nei distance estimates taken from Table 3.1 but expressed as a diagonal matrix only.

Unbiased Nei distance	Kitale	Kibale	Laikipia	Mara
Kibale	0.0428	-	-	-
Laikipia	0.0920	0.0539	-	-
Mara	0.1002	0.0676	0.0062	-
Lushoto	0.1071	0.0594	0.0104	0.0060

As noted in Chapter 2, different methods are available for estimating allele frequencies at loci (we recommended a Bayesian approach), and of course this influences the results then obtained when calculating genetic distances. Furthermore, a whole range of other distance measures in addition to Nei's coefficient can be calculated, using a range of software packages (see more in Box 3.1). We encourage readers to investigate these methods further through the references at the end of this chapter (Kosman and Leonard 2007, Nagamine and Higuchi 2001, Reif et al. 2005). Among these alternative measures, Wright's *F*-statistics are the most important (we refer to these later in Chapter 8). The formulae for a range of distance measures are given in Appendix I and the use of these can be explored further through the electronic files on the CD-ROM.

Because of the number of methods available, when writing up results it is always advisable to specify which distance measure was adopted and which software package was used to calculate it. In practice, it is only if the results from different methods cause large differences in how populations place (when the relationships among stands are visualised through, e.g., clustering or ordination; see Chapters 4 and 5) that the measure used becomes a matter of significant concern. As always in these matters, the greater challenge lies not in the methods used, but in the sensible interpretation of data in a biological context.

3.2. References

Kosman E, Leonard KJ (2007) Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction. *New Phytologist* 174: 683-696.

Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology* 3: 91-99.

Nagamine Y, Higuchi M (2001) Genetic distance and classification of domestic animals using genetic markers. *Journal of Animal Breeding and Genetics* 118: 101-109.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583-590.

Reif JC, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science* 45: 1-7.

3.3. Suggested software

A wide range of software packages produce genetic distance matrices (e.g., GenAlEx, PopGene, AFLP-SURV, TFPGA and FAMD), but we recommend using AFLP-SURV since this packages allows estimation of allele frequencies by Bayesian methods, and then provides distance matrices based on Nei's genetic distance (see Appendix III).

Box 3.1. Comparing results from different coefficients used to calculate distance matrices

Methods are available to graphically compare the genetic distances produced by different distance measures, to see how closely they correspond (see also Box III.2 in Appendix III). Based on our example *Warburgia* data set, Fig 3.1 shows Nei's genetic distances between pairs of populations plotted against the pairwise coancestry distance between populations. The 10 data points represent the number of pairwise comparisons between populations that are possible for each measure. As expected, estimates increase together, although the relationship is not entirely linear, showing that the method used can influence relative estimates among stands.

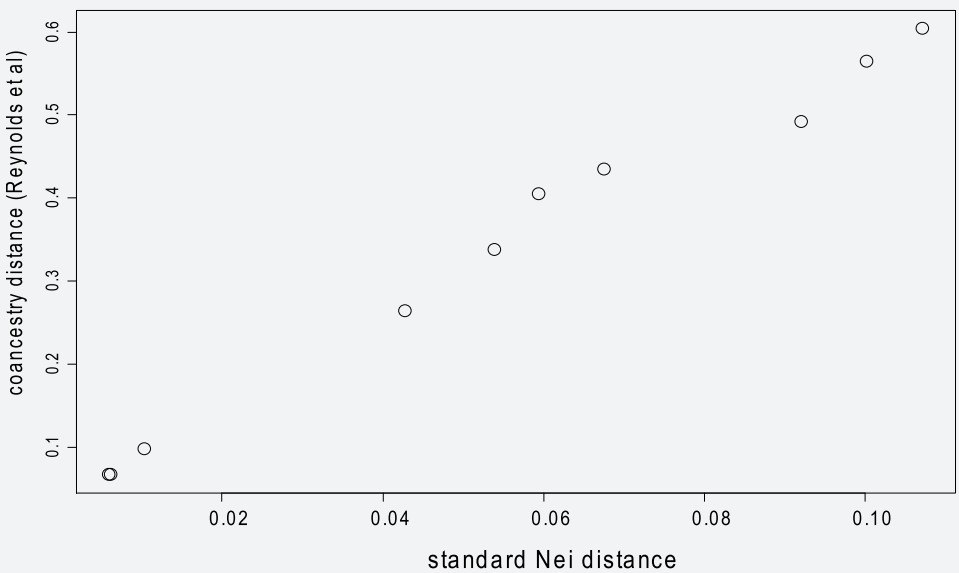


Figure 3.1. Plot of Nei's standard genetic distances against pairwise coancestry distances for five *W. ugandensis* populations. Nei's genetic distances were calculated with the AFLP-SURV 1.0 package, coancestry distances by the TFPGA 1.3 package. The figure was created with the BiodiversityR package.

Chapter 4. Visualising genetic distances by cluster analysis

Key points

This section explains how to visualise genetic distances between populations by cluster analysis. It relies on having first calculated a distance matrix (see Chapter 3).

In interpreting the results of cluster analysis, it is important to remember that it only provides an incomplete overview of the relationships between populations.

The most common approach to cluster analysis is to use an unweighted pair group method with arithmetic averaging (UPGMA).

4.1. Cluster analysis of genetic distances

Once genetic distances have been determined (see Chapter 3), the relationships between populations can be summarised visually through cluster analysis or ordination. In this chapter we describe the first option, while Chapter 5 relates the second approach. With both methods, it is important to remember that the techniques used are unable to fully express the relationships among populations. They therefore provide an overview of structure only. This is an important point to remember in interpretation, as not all significant features of data may be evident.

The most common approach in cluster analysis is to use an unweighted pair group method with arithmetic averaging (known as UPGMA). This means that the distance at which a cluster is formed corresponds to the average of all the pair-wise distances between populations that are joined together in that particular step (thus not including pair-wise differences of populations that had already been joined in earlier steps). An example of a genetic distance matrix from our *Warburgia* data set (see information in Chapter 1) was shown in Tables 3.1 and 3.2 (see Chapter 3). The result of a UPGMA cluster analysis based on this matrix, known as a phenogram, is shown in Fig. 4.1.

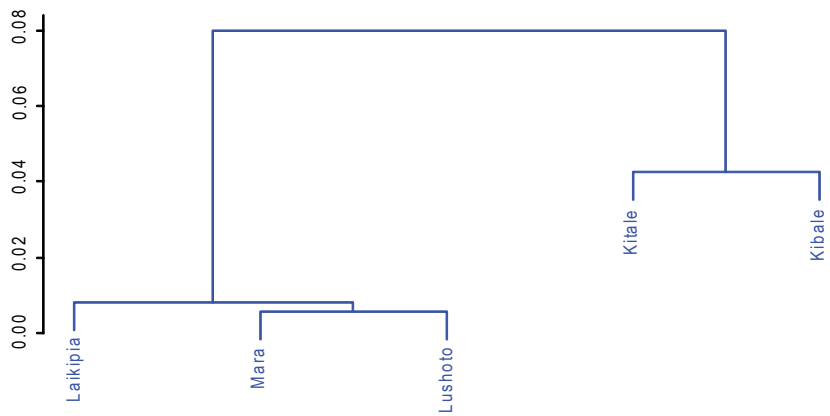


Figure 4.1. A phenogram showing cluster analysis of five *W. ugandensis* populations, based on the UPGMA clustering method and Nei's unbiased genetic distances (distances taken from Table 3.2). The vertical axis shows the genetic distance at which populations cluster. The figure was created with the BiodiversityR package.

Our example shows that the first clustering step is between Mara (Masai Mara) and Lushoto populations (at a distance of about 0.005). This is followed by a second step that joins the Laikipia population to these two populations (at a distance of about 0.01). Kitale and Kibale populations join in a separate cluster (at a distance of about 0.045). Finally, the two clusters (of [Kitale + Kibale] and [Laikipia + Mara + Lushoto]) join into a single large cluster. The pattern revealed corresponds with the regional distribution of populations, with Laikipia, Lushoto and Masai Mara on the east side of the Rift Valley, and Kibale and Kitale on the west (see Chapter 1). However, the populations on the east side are genetically much closer to each other than those on the west side are.

It is important to note that spinning the dendrogram around any vertical branch provides an equally valid representation of cluster memberships. For example, the positions of Kitale and Kibale could be interchanged on the vertical axis, as could Masai Mara and Lushoto. In other words, it should not be concluded from Fig. 4.1 that, e.g., Lushoto and Kitale populations are more similar to each other than Lushoto and Kibale are.

Cluster analysis can be undertaken in a variety of other ways using a range of software packages. We encourage users of this guide to explore methods further by reading the manual of Kindt and Coe (2005), which is given on the accompanying CD-ROM. The use of alternative approaches can also be tested through the electronic files on the CD-ROM. It makes sense to test a range of methods and see if these generate the same pattern of results. If they do not, more thought needs to be given to what is the best approach. We tested our *Warburgia* data set with a range of clustering methods and, although these approaches resulted in some changes in branch lengths, the overall pattern of differentiation revealed among populations was the same (see Fig. 4.2). Because of the range of approaches available, when writing up results it is advisable to specify which method or methods were applied, and which software package was used.

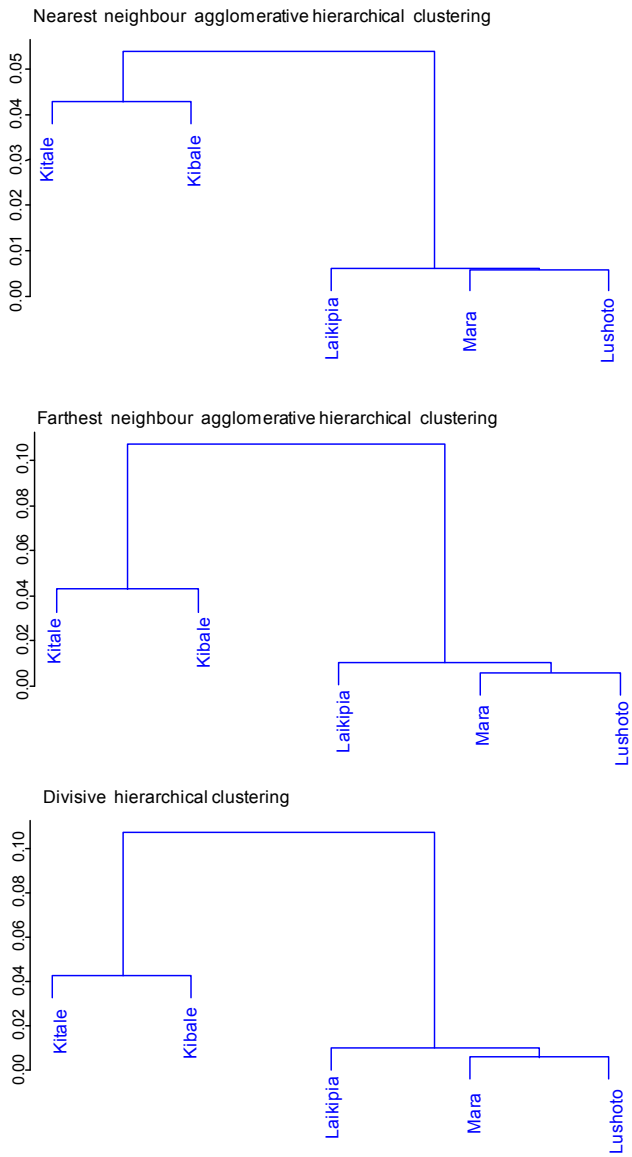


Figure 4.2. Phenograms showing cluster analysis of five *W. ugandensis* populations using three different methods. All are based on Nei's unbiased genetic distances (distances taken from Table 3.2). The vertical axis shows the genetic distance at which populations cluster. In this example, the method used makes no difference to the pattern observed in clustering. The figure was created with the BiodiversityR package.

4.2. Assigning levels of significance to relationships by bootstrap analysis

Using bootstrap analysis, it is possible to obtain information about the influence of the number of loci scored in data sets on the pattern of clustering observed (a method to do this is given in Appendix III). Bootstrapping is based on sampling collections of loci at random from data and calculating how often the same phenograms are obtained. This allows confidence to be assigned to the 'overall' pattern of relationships (e.g., as in Figs. 4.1 and 4.2) observed among populations. A method of selection-with-replacement is used to construct many new molecular data sets from scored loci with the same dimensions as the original data set. A locus can be included more than once or not at all when a new data set is generated.

A bootstrap analysis based on 10,000 random data sets generated from our *Warburgia* example demonstrated that we can be confident that Kibale and Kitale cluster uniquely. On the other hand, least confidence is assigned to the positioning of Mara and Lushoto into a group, as in only 52% of cases did these populations cluster first with each other rather than with another population.

4.3. References

Kindt R, Coe R (2005) Tree Diversity Analysis: A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies. The World Agroforestry Centre, Nairobi, Kenya. Available from the CD-ROM and at: www.worldagroforestry.org/treesandmarkets/tree_diversity_analysis.asp

4.4. Suggested software

A wide range of software packages provide cluster results (e.g., PopGene, TFPGA, FAMD, BiodiversityR and AFLP-SURV, the last in combination with PHYLIP). The combination of AFLP-SURV with PHYLIP allows estimation of allele frequencies by Bayesian methods, provides distance matrices based on Nei's measure, and allows bootstrap analysis of cluster relationships (see Appendix III).

Chapter 5. Visualising genetic distances by ordination

Key points

This chapter explains how to visualise genetic distances between populations by ordination. It relies on having first calculated a distance matrix (see Chapter 3).

In interpreting the results of ordination, it is important to remember that it only provides an incomplete overview of the relationships between populations.

It is possible to superimpose onto ordination diagrams further information such as clustering data that can aid in interpretation.

The most common approach to ordination is principal coordinate analysis (PCoA).

5.1. Ordination of genetic distances

Once genetic distances between populations have been determined, relationships among populations can be summarised visualised through cluster analysis or ordination. In Chapter 4 we described the first option and here we relate the second approach. As with cluster analysis, it is important to remember that ordination is unable to fully express the relationships that exist among populations and it provides an overview only (how good an overview can be tested, see Box 5.1). The ordination approach works best when comparing relationships between populations that are rather different from each other, and when assessing the possibilities for interaction between different genetic entities. For example, ordination is a good approach for detecting hybrid stands, as in analysis such stands may well locate intermediately between aggregations of 'pure' populations.

During ordination, a pairwise distance matrix is subjected to an analysis that expresses observed differences in terms of different positions along a small number of principal axes of variation, positions that can then be visually compared in two- or three-dimensional diagrams. Several ordination techniques, such as principal coordinate analysis (PCoA) and non-metric multidimensional scaling, are appropriate for molecular data. The most common approach used is PCoA. An example of a genetic distance matrix from our *Warburgia* data set (see information in Chapter 1) was shown in Tables 3.1 and 3.2 (see Chapter 3) and the results of a PCoA based on this matrix are shown in Fig. 5.1 (Box 5.2 at the end of this chapter explains how variation is assigned to principal axes).

Our example shows that Laikipia, Lushoto and Mara (Masai Mara) are well separated from Kibale and Kitale on the first principal axis of variation, consistent with clustering (see Chapter 4) that defined two distinct groups of populations based on their regional location (Laikipia, Lushoto and Masai Mara sampled from the east side of the Rift Valley, Kibale and Kitale from the west).

It is important to note that rotating Fig. 5.1 around its horizontal or vertical axis makes no difference to the relationships observed among populations. If different software packages produce mirror images of relationships this is therefore not an issue of concern. Such figures can also be rotated for presentation purposes if this is convenient, e.g., if it makes comparison with a geographic map of sample locations easier.

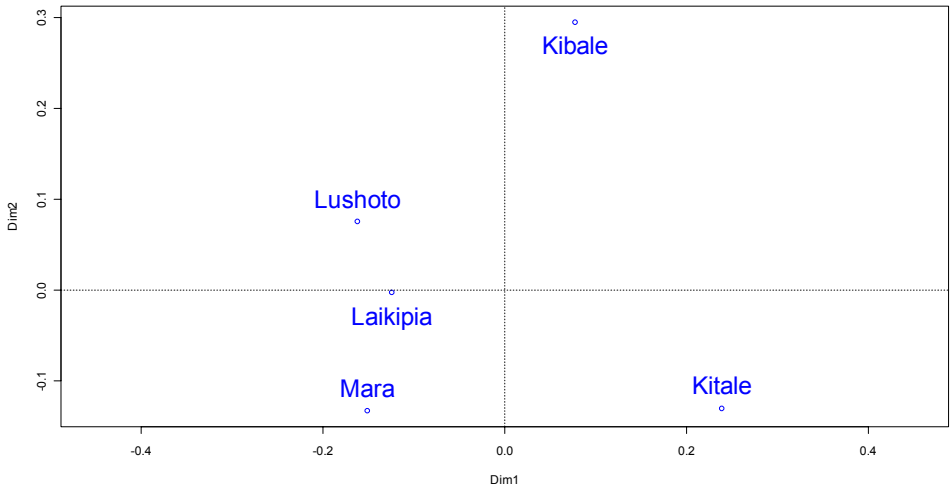


Figure 5.1. PCoA of five *W. ugandensis* populations, based on Nei's unbiased genetic distances (distances taken from Table 3.2). Horizontal and vertical scales represent the first and second principal axes of variation respectively (the two axes that explain the most variation among populations). In this instance, the 1st principal axis represents a large 92.7% of variation, the 2nd a much smaller 6.9%. The figure was created with the BiodiversityR package.

In Fig. 5.1, only the first two principal axes of variation are shown, and these unusually explain the vast majority of overall variation detected in analysis. When the first two axes explain less of the overall variation, however, 3-dimensional figures that provide the 3rd principal axis (into the page) can be generated and are useful. Alternatively, 2-dimensional figures that represent the 1st and 3rd axes, or 2nd and 3rd axes, can also be created and displayed alongside a figure showing the 1st and 2nd axes.

Other ways to assist in interpretation of analysis are also available, such as superimposing a cluster analysis over ordination figures. An example is shown in Fig. 5.2 for our *Warburgia* data set (same PCoA as Fig. 5.1).

When superimposing cluster information in this way, the closer relationship between the Lushoto and Masai Mara (Mara) populations with each other than with Laikipia is more clearly represented than by ordination alone. This illustrates the point made at the beginning of this chapter about the positioning of populations in ordination diagrams only approximating the distance matrices used as inputs for analysis.

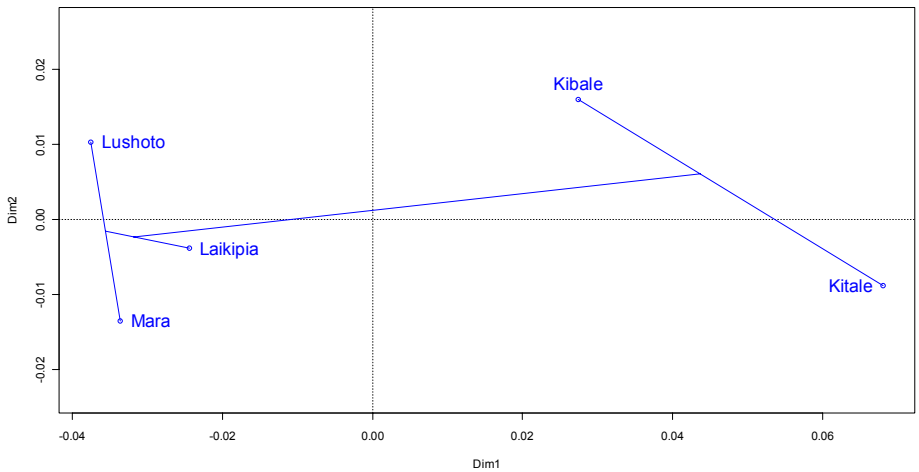


Figure 5.2. PCoA of five *W. ugandensis* populations (PCoA taken from Fig 5.1) with clustering positions superimposed. Clustering was based on the nearest neighbour method (there are statistical reasons why this is the best method to use when superimposing onto ordination diagrams, see Legendre and Legendre 1998). The figure was created with the BiodiversityR package.

We encourage users of this guide to explore the other methods available for ordination by reading Kindt and Coe (2005). The use of alternative approaches can also be tested through the electronic files on the CD-ROM. Just as with other methods described in this guide, it makes sense to test a range of approaches and see if these generate similar results. Because of the range of approaches available, when writing up results it is important to specify which method was used and to indicate which software package was applied.

5.2. References

Kindt R, Coe R (2005) Tree Diversity Analysis: A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies. The World Agroforestry Centre, Nairobi, Kenya (196 pages). Available from the CD-ROM and at: www.worldagroforestry.org/treesandmarkets/tree_diversity_analysis.asp

Legendre P, Legendre L (1998) Numerical Ecology. Developments in Ecological Modelling 20. Elsevier, Amsterdam, The Netherlands, 853 pp.

5.3. Suggested software

Various software packages provide ordination results (e.g., GenAlEx, FAMD and BiodiversityR). We recommend use of the BiodiversityR package as it allows cluster information to be superimposed onto ordination diagrams and has other features that provide added value in analysis (see Appendix III).

Box 5.1. Comparing the results of ordination with genetic distance matrices

Ordination can only provide a summary of the relationships between populations that are more fully expressed in distance matrices. How good a summary is provided by ordination can be explored by plotting distances shown in ordination diagrams against original distances. The results for our test *Warburgia* data set (with a trend-line added) are shown in Fig 5.3. The 10 data points represent the number of pairwise comparisons between populations that are possible for each distance measure.

For our example, it is evident that there is a generally good relationship between ordination results and original distances. The exception is at low genetic distances. This suggests that when populations are genetically quite similar ordination results should be interpreted cautiously.

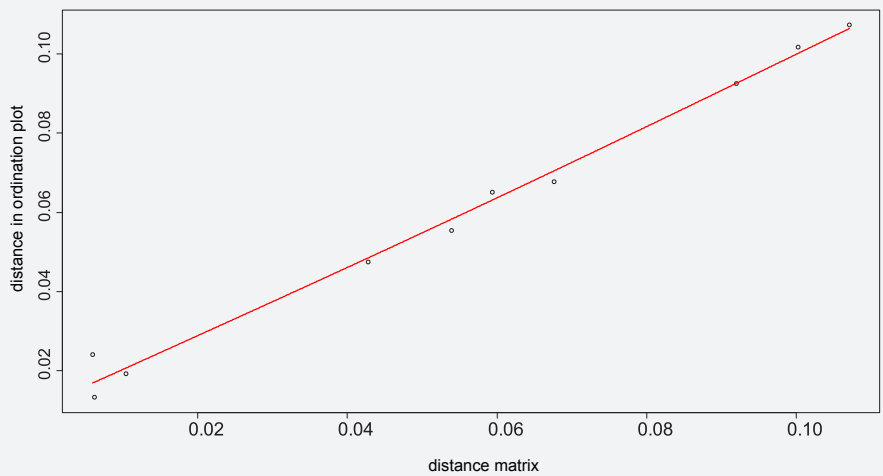


Figure 5.3. Plot of ordination distances (vertical axis) against Nei's unbiased genetic distances (horizontal axis) for five *W. ugandensis* populations. The figure was created with the BiodiversityR package.

Box 5.2. Assigning levels of variation to the principal axes of ordination

Ordination expresses variation in terms of positions along a small number of principal axes of variation. In order to understand what percentage of variation is explained by each axis, the eigenvalue obtained from analysis that corresponds to a particular axis must be divided by the total sum of all eigenvalues (from all axes, the number of which usually corresponds to the number of populations minus one; in the case of our test data set of *Warburgia* [5 populations], there are four eigenvalues).

Negative eigenvalues are sometimes obtained in analysis. This indicates that the results of ordination do not exactly match distances in the original input matrix used for analysis (see Kindt and Coe 2005). The occurrence of negative eigenvalues can be used to tell us something fundamental about the distance matrix. When estimating the percentage of variation explained by different axes, there are different ways of dealing with negative eigenvalues: they can simply be ignored in calculations, or (better) absolute values can be used (i.e., multiply negative values by -1).

A similar procedure for estimating the percentage of variation explained by principal axes can be used when dealing with individuals (see Chapter 7). The only difference is that the total number of axes involved is generally much higher (usually the number of individuals minus 1).

Part 3

Analysing data at the individual level

Chapter 6. Measuring genetic distance between individuals

Key points

This chapter explains how to calculate genetic distances between pairs of individuals.

For dominant markers, incomplete information at loci means that it is not possible to define the full genotypes of individuals. Distance measures must thus rely on comparing the occurrence of product presences and absences between individuals, without considering whether or not individuals are heterozygous.

Common approaches for measuring genetic distances between pairs of individuals include simple mismatching, Jaccard's and Dice's coefficients.

Once genetic distances between pairs of individuals have been estimated, the relationships between individuals can (as for populations) be visualised by other approaches, with ordination being especially useful (see Chapter 7).

6.1. Calculating genetic distances from product distributions

Chapter 3 of this guide described how dominant markers could be used to measure genetic distances between populations. Because dominant markers are unable to directly reveal heterozygous 'present-absent' [1, 0] conditions, this required assumptions in converting product frequencies to allele frequencies. The calculation of genetic distances between pairs of individuals is clearly also compromised by the inability to identify heterozygotes. In this instance, calculations must simply be based on the presence or absence of products without reference to possible heterozygosity. This is a significant assumption when dealing with organisms that are highly heterozygous, as outcrossing trees species generally are.

There are a whole range of measures for calculating genetic distances between pairs of individuals based on dominant markers, but the three most common ones are: (1) simple mismatching (the same as 1 – simple matching!), (2) Jaccard's measure; and (3) Dice's distance. Distance measures like these have been used widely in ecology and as a result the same formulae have frequently been given different names. For example, simple mismatching is sometimes referred to as Hamming's distance, while the Dice distance is also known as the Sørenson coefficient and as the Bray-Curtis measure. Legendre and Legendre (1998) provide an overview of the different measures available and explain some of the 'overlaps' in the naming of approaches. On a practical level, it is always important to report both the name of the measure applied and the software package used in your calculations.

The formulae for different measures are given in Appendix I and we encourage readers to explore the use of these further through the electronic files on the CD-ROM accompanying this guide. Once calculated, distances can be presented in a triangular matrix, as shown in Tables 6.1, 6.2 and 6.3 (each based on one of the three common distance methods described above) for the first five individuals of our example *W. ugandensis* data set (see further information in Chapter 1; values based on all 185 AFLP loci scored).

Table 6.1. Distance estimates between five individuals of *W. ugandensis*, based on the simple mismatching coefficient. Results were obtained with the BiodiversityR software package.

Simple mismatching	Kit01	Kit02	Kit03	Kit04
Kit02	0.141			
Kit03	0.162	0.108		
Kit04	0.205	0.173	0.119	
Kit05	0.108	0.108	0.097	0.141

Table 6.2. Distance estimates between five individuals of *W. ugandensis*, based on Jaccard's distance. Results were obtained with the BiodiversityR software package.

Jaccard's distance	Kit01	Kit02	Kit03	Kit04
Kit02	0.406			
Kit03	0.448	0.294		
Kit04	0.535	0.432	0.314	
Kit05	0.357	0.323	0.290	0.394

Table 6.3. Distance estimates between five individuals of *W. ugandensis*, based on Dice's distance. Results were obtained with the BiodiversityR software package.

Dice's distance	Kit01	Kit02	Kit03	Kit04
Kit02	0.255			
Kit03	0.288	0.172		
Kit04	0.365	0.276	0.186	
Kit05	0.217	0.192	0.170	0.245

6.2. Choosing between distance measures

As can be seen from the above three tables, the coefficient chosen can make a big difference to the absolute values of the pairwise distances generated. This is because these coefficients are calculated in quite different ways. Of particular note, simple mismatching includes shared product absences when making calculations, whereas Jaccard's and Dice's measures do not.

The relationship between the three measures for our total *Warburgia* data set is illustrated in Figure 6.1. As can be seen, the distribution of points is wide along a diagonal both for simple mismatching compared to Jaccard's measure and for simple

mismatching compared to Dice's measure. This indicates that the measures do not compare well in the common ranking of distances between pairs of individuals. On the other hand, when comparing Jaccard's and Dice's measure with each other, a single curve is observed that indicates direct correspondence.

Knowing which distance measure is best to use depends on understanding something about both the technicalities of the laboratory method that is used to reveal variation and the biology of the material being researched. When considering PCR approaches such as AFLP analysis or the random amplified polymorphic DNA (RAPD) procedure, the shared presence of a product between two individuals can generally be assumed to indicate that they are identical at that point in the genome. The shared absence of a product does not however necessarily indicate the same thing, as a whole range of different mutations could lead to product absence.

For this reason, the greater the expected biological differences within the range of material under study, the more emphasis should be placed on using measures that ignore shared absences. For example, when examining different species more credence should be give to results obtained with Jaccard's or Dice's measure than simple mismatching. We suggest reading Kosman and Leonard (2005) for a more thorough discussion on this topic. Of course, if different measures indicate the same basic relationships between individuals when distances are visualised by methods such as ordination (see Chapter 7), then the greater the certainty that can be placed on results.

6.3. Dealing with missing data

When dealing with individual- rather than population-level data, it is often the case that no information will be available for a particular pairwise comparison because of missing [NA] data points. This is especially likely when laboratory analysis uses a technique that is less reliable, e.g., if PCR often fails. For example, the RAPD method tends to fail frequently and therefore leaves more gaps in data sets than AFLP analysis does. Most software packages can handle missing data by excluding particular pairwise comparisons in analysis. We would however suggest that those individuals or loci that have a high proportion of missing data (e.g., when more than 10% of data points are not able to be scored) be completely excluded from individual-level analyses.

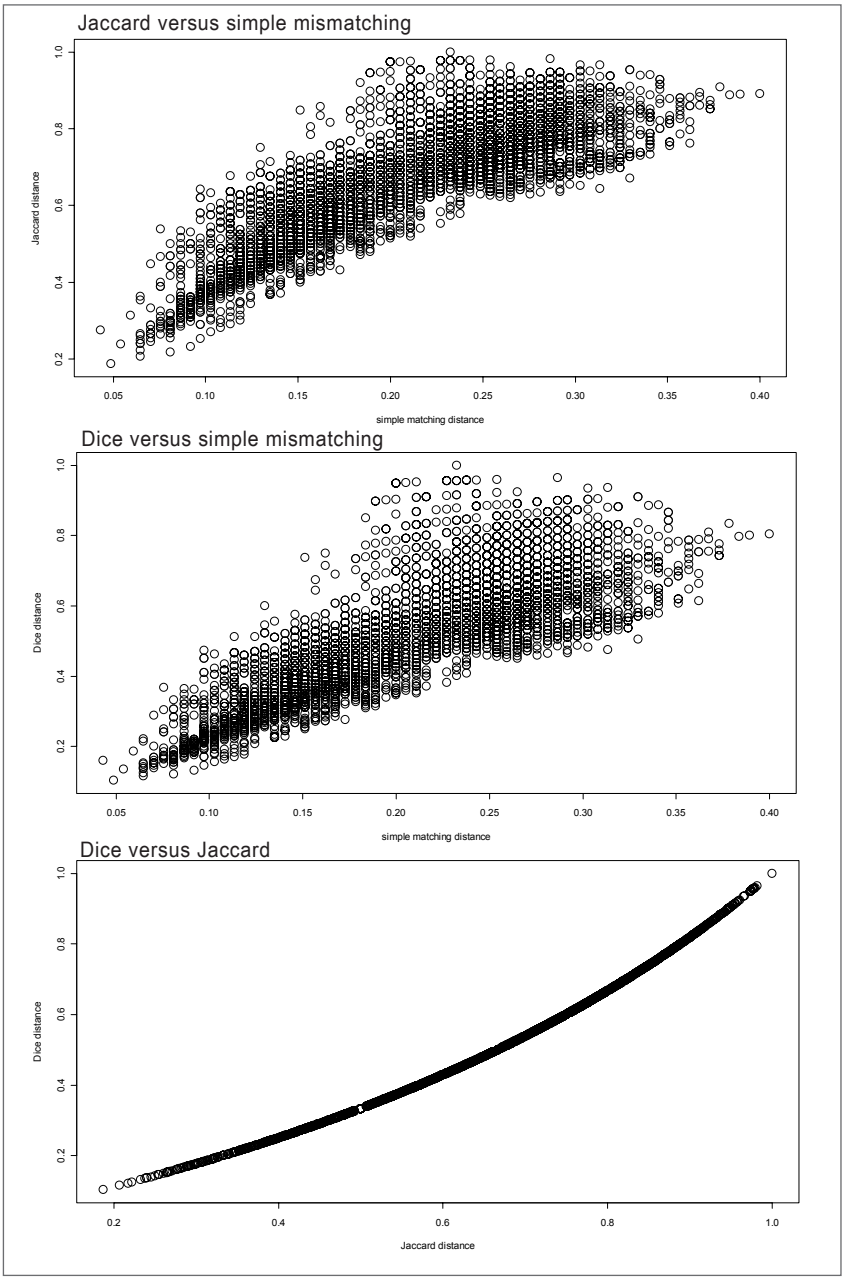


Figure 6.1. Comparisons of simple mismatching, Jaccard and Dice distance estimates between pairs of *W. ugandensis* individuals (100 individuals in total). Each circle represents a single comparison, with the number of comparisons made corresponding to the size of the distance matrix. Comparisons were undertaken and figures generated in the BiodiversityR software package.

6.4. References

Legendre P, Legendre L (1998) Numerical Ecology. Developments in Ecological Modelling 20. Elsevier, Amsterdam, The Netherlands, 853 pp.

Kosman E, Leonard KJ (2005) Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology* 14: 415-424.

6.5. Suggested software

A wide range of software packages can calculate genetic distances between pairs of individuals and generate distance matrices (e.g., GenAlEx, FAMD, BiodiversityR and AFLP-SURV). We recommend the use of the BiodiversityR package as it provides the most options for subsequent ordination analysis, as described in Chapter 7 (see Appendix III).

Chapter 7. Visualising genetic distances by ordination

Key points

This section explains how to visualise genetic distances between individuals by ordination. This relies on having first calculated a distance matrix (see Chapter 6).

In interpreting the results of ordination, it is important to remember that the technique only provides an incomplete overview of the relationships between individuals.

By assessing the spread of distributions of individuals within populations, further insights into species biology are provided.

The most common approach to ordination is principal coordinate analysis (PCoA)

7.1. Ordination of genetic distances

Just as we showed earlier for populations (Chapters 4 and 5), once genetic distances between individuals have been determined, the relationships between them can be summarised visually through cluster analysis or ordination. In the case of tree species, which normally express high genetic variation within populations and relatively low diversity among stands, cluster analysis at the level of individuals is generally not very informative (although there are exceptions, see Jamnadass et al. 2005 for a clear case). More useful is to visualise differences by ordination, and we therefore focus on this approach here. As we related earlier with reference to populations (Chapter 5), it is important to remember that ordination can provide only a summary of the relationships between individuals (how good a summary can be tested, as was shown in Box 5.1 for populations; the same approach can be applied to individuals). Ordination expresses variation in terms of a small number of principal axes in two- or three-dimensional diagrams, and works well when comparing relationships among individuals that are rather different from each other.

As noted in Chapter 5, several ordination techniques can be applied to molecular data, such as principal coordinate analysis (PCoA) and non-metric multidimensional scaling. We encourage users of this guide to explore the additional use of methods such as distance-based redundancy analysis, which combines an 'analysis of variance' with ordination and allows for tests of significance of relationships by randomisation procedures, through reading Kindt and Coe (2005). The use of alternative approaches can also be tested through the electronic files on the CD-ROM. Just as with other methods described in this guide, it makes sense to test a range of approaches. Because of the range of methods available, when writing up results it is important to specify which technique was used and to indicate which software package was applied.

Here, we demonstrate the use of the standard PCoA approach on our test *Warburgia* data set (see information in Chapter 1). The results of a PCoA based on a pairwise genetic distance matrix calculated according to Jaccard's measure (see Chapter 6; based on all 185 AFLP loci scored) is shown in Fig 7.1. (Note that Box 5.2 in Chapter 5 explained how variation is assigned to principal axes during the ordination analysis of populations; the same approach applies for the ordination of individuals).

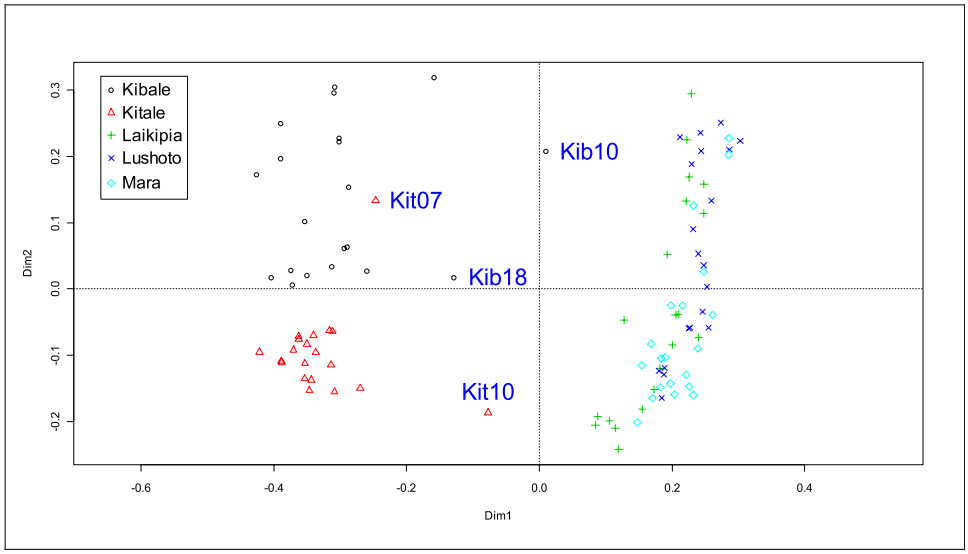


Figure 7.1. PCoA of 100 *W. ugandensis* individuals taken from five populations, based on genetic distances calculated according to Jaccard's method. Individuals from different populations are labelled with different symbols. Four individuals that position unusually in ordination are indicated. Horizontal and vertical scales represent the first and second principal axes of variation, respectively (the two axes that explain the most variation among individuals). In this instance, the 1st principal axis represents 28% of variation, the 2nd 9%. The figure was created with the BiodiversityR package.

Our example shows that individuals from Laikipia, Lushoto and Mara (Masai Mara) overlap in distribution but are generally well separated from individuals from Kibale and Kitale on the first principal axis. This separation is consistent with previous cluster and ordination analyses of populations (see Chapters 4 and 5), which defined two distinct groups of populations that corresponded with regional location (Laikipia, Lushoto and Masai Mara stands on the east side of the Rift Valley, Kibale and Kitale on the west). As noted in Chapter 5, rotating an ordination graph such as Fig 7.1 around its horizontal or vertical axis makes no difference to the relationships observed (meaning that, if convenient, figures can be rotated for presentation purposes). As also noted previously, 3-dimensional figures that provide the 3rd principal axis (into the page) of an ordination analysis can be generated, or 2-dimensional figures that also represent the 1st and 3rd axes, or 2nd and 3rd axes.

Our example shows that an ordination of individuals provides some additional information compared to an analysis of populations, as it provides an overview of the spread of variation within stands, which in this instance is considerable. It is also able to identify particular individuals in certain populations that place unusually and that may merit further investigation – in this case, two individuals from Kibale (Kib10 and Kib18), and two from Kitale (Kit07 and Kit10).

Further analysis of our *Warburgia* example could involve undertaking ordination for pairs of populations only. In addition, the identity of the four unusual individuals identified above could be explored in more detail using only those markers that are diagnostic (e.g., that have a 'product presence' frequency difference of 0.8 or more) between the two geographic regions covered by the study. Might these individuals represent recent introductions into stands? Or, could they be the products of hybridisation between populations, or even with alternate species in the *Warburgia* genus (see Muchugi et al. 2008b for a description of other *Warburgia* species local to the region)? It is clear that further interpretation requires an understanding of the taxonomy and biology of the genus.

Other ways to assist the interpretation of ordination diagrams are also available (e.g., using BiodiversityR). For example, individuals can be connected to the centroid positions (the multivariate average) of their respective populations to generate 'spider' diagrams, as illustrated for our *Warburgia* example in Fig. 7.2. The shorter the 'legs' of the spider are, the smaller the variance within a population is (e.g., in our example, legs are shorter on average in Kitale than Kibale), while the greater the distance between centroids then the more different two populations are. 'Odd' individuals can be identified that plot toward the centroids of populations from which they were not collected (in our example, only one individual, from Kitale [Kit07], is thereby identified). Alternatively, arrows can be superimposed between the most related individuals according to the original distance matrix.

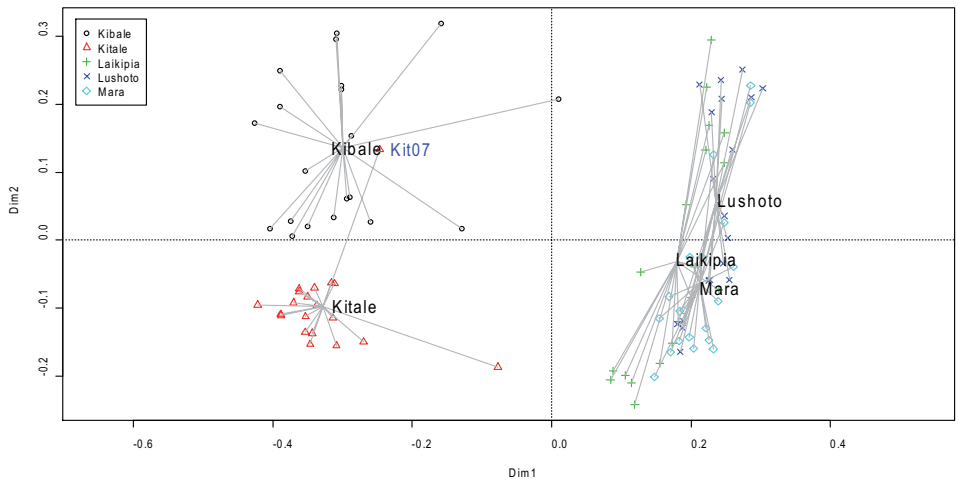


Figure 7.2. PCoA of 100 *W. ugandensis* individuals taken from five populations (PCoA taken from Fig 7.1) with a spider diagram linking individuals to population centroid positions (labelled by name of population) superimposed. The figure was created with the BiodiversityR package.

7.2. References

Jamnadass R, Hanson J, Poole J, Hanotte O, Simons TJ, Dawson IK (2005) High differentiation among populations of the woody legume *Sesbania sesban* in sub-Saharan Africa: implications for conservation and cultivation during germplasm introduction into agroforestry systems. *Forest Ecology and Management* 210: 225-238.

Kindt R, Coe R (2005) Tree Diversity Analysis: A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies. The World Agroforestry Centre, Nairobi, Kenya. Available from the CD-ROM and at: www.worldagroforestry.org/treesandmarkets/tree_diversity_analysis.asp

Muchugi A, Muluvi GM, Kindt R, Kadu CAC, Simons AJ, Jamnadass RH (2008b) Genetic structuring of important medicinal species of genus *Warburgia* as revealed by AFLP analysis. *Tree Genetics and Genomes* 4: 787-795.

7.3. Suggested software

Various software packages allow for ordination of individuals (e.g., GenAlEx, FAMD and BiodiversityR). We recommend use of the BiodiversityR package as it allows additional information to be superimposed onto ordination diagrams and has other features that provide added value in analysis (see Appendix III).

Part 4

Further methods

Chapter 8. Analysis of molecular variance (AMOVA)

Key points

This section explains how to carry out an analysis of molecular variance (AMOVA).

AMOVA is a particular approach that partitions genetic variation among individuals within populations and among populations. It can also be used to partition variation at higher levels of structure in nested analyses (e.g., by geographic region or species).

As is the case with other analyses based on dominant markers, incomplete information at loci means assumptions during analysis. AMOVA of dominant data relies on treating individuals as haploids.

The AMOVA approach has been widely adopted because it is a simple way of expressing variation and it makes comparisons with other studies straightforward.

8.1. Partitioning variation within and among populations

The AMOVA approach as described by Excoffier et al. (1992) generates squared Euclidean distances between pairs of individuals and then partitions this variation at different levels of structure – within and among populations, among regions, among species, etc. – depending on the hierarchy that is available for testing. As noted earlier in this guide (Chapter 6), the calculation with dominant markers of genetic distances between pairs of individuals is compromised by the inability to identify heterozygotes. AMOVA, therefore, estimates distances between pairs of individuals based simply on the presence or absence of products. It can be demonstrated that using the Euclidean distance corresponds to treating individuals as completely inbred (with the same allele frequencies as if the dominant markers were obtained from haploid organisms), which is clearly not realistic for outbreeding trees (see Appendix III).

A powerful feature of AMOVA is that it can accept a number of levels of structure within a single analysis, allowing variance to be 'nested' across levels. As implemented in packages such as GenAlEx and Arlequin, significance values can then be ascribed by permutation tests to the variance which partitions at different levels of structure. Permutation tests compare given results with a situation where no genetic structure in material is assumed, and estimate how often the actual results exceed expectations on this basis.

An AMOVA for our example *W. ugandensis* data set (see further information in Chapter 1; values based on all 185 AFLP loci scored) is shown in Table 8.1. The table expresses the proportion of variance within and among populations based on an analysis nested into two regions, as defined by the locations of stands with respect to the Rift Valley (Laikipia, Lushoto and Masai Mara to the east of the Rift, Kibale and Kitale to the west). For our example, it is evident that a significant proportion of variance (33%) partitions between regions, consistent with earlier illustrations of analyses in this guide (see Parts 2 and 3). The variance partitioning between populations within regions (10%) is less than between regions but is still significant, and this may be accounted for primarily by the difference observed earlier between Kibale and Kitale stands from west of the Rift Valley. A large fraction of variance (56%) is found within stands, consistent with the scatter of individuals within populations in ordination diagrams (see Chapter 7).

Table 8.1. A nested AMOVA (by region) for five populations of *W. ugandensis*. The percentage of variance partitioning at different levels of structure and associated significance values are shown, as are degrees of freedom (d.f.), mean squared deviations (MSD) and the values of the variance components. Results were obtained with the GenAlEx 6.2 software package.

Source of variation	d.f.	MSD	Variance component	% of total	*P*-value*
Between regions	1	442.4	7.90	33	< 0.001
Among populations within regions	3	63.3	2.49	10	< 0.001
Among individuals within populations	95	13.5	13.49	56	< 0.001

* In the GenAlEx output, *P*-values are given in relationship to Phi values of genetic structure. In this instance, significance values were assigned to partitioned variation by 9,999 permutations (one of the given values to choose from in GenAlEx) of genetic structure.

The AMOVA approach also generates F_{ST} values (one of the *F*-statistics such as Weir and Cockerham's 1984 measure, see Appendices I and III). This parameter is calculated as diversity among populations / total diversity. Values close to zero indicate little differentiation among populations (that is, most genetic diversity is within populations), while values close to one indicate high differentiation (that is, most diversity is among populations). Generally, F_{ST} values in the range of 0 to 0.05 are considered low, 0.05 to 0.15 moderate, 0.15 to 0.25 large, and 0.25 to 1.0 very large.

AMOVA can also generate genetic distance values between pairs of populations, thus providing an alternative approach to those methods described in Chapter 3 of this guide. Pairwise distances are calculated by subsequent AMOVA analyses for each pair of populations, interpreting the resulting F_{ST} values as a measure of genetic distance. Results for our example *W. ugandensis* data set are shown in Table 8.2. In our analysis, distance values between Laikipia, Lushoto and Masai Mara (Mara) were lowest, consistent with previous findings (Part 2; see especially Chapters 4 and 5). All pairwise distances between populations were considered to be significant, even for Laikipia versus Lushoto, Lushoto versus Masai Mara and Laikipia versus Masai Mara. The significance values revealed by AMOVA should however be treated with caution because, as noted above, the method used to estimate distances between populations in this case relied on product distributions rather than estimated allelic frequencies. (AMOVA is thus best used for looking at 'relative' rather than 'absolute' differences between stands.)

Table 8.2. Pairwise genetic distances (F_{ST}) values from AMOVA for five populations of *W. ugandensis*. Results were obtained with the GenAlEx 6.2 software package. Levels of significance were assigned to values by 9,999 permutations (one of the given values to choose from in GenAlEx) of genetic structure. All pairwise distance values were considered significant ($P < 0.001$).

AMOVA distances	Kitale	Kibale	Laikipia	Mara
Kibale	0.265			
Laikipia	0.447	0.356		
Mara	0.495	0.405	0.069	
Lushoto	0.526	0.393	0.109	0.077

8.2. References

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479-491.

Weir BS, Cockerham C (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.

8.3. Suggested software

AMOVA can be undertaken in GenAlEx, Arlequin (similar approach to GenAlEx) and FAMD (see Appendix III). GenAlEx calculates squared Euclidean distances and then uses these for further analysis. Although FAMD can provide AMOVA statistics for a wider range of genetic distance measures than GenAlEx, it does not provide tests for statistical significance.

Chapter 9. STRUCTURE analysis

Key points

This section explains how to carry out analysis of data in the software package STRUCTURE.

STRUCTURE analyses genetic differentiation by assigning individuals to a number of assumed groups (*K*), where *K* is set by the user.

STRUCTURE provides useful graphical outputs that allow the genetic composition of an individual to be compared with its given population identity.

9.1. The basis of STRUCTURE

The STRUCTURE approach as described by Falush et al. (2007) infers genetic structure by assigning individuals probabilistically to one or more of a number of assumed groups in analysis (*K*), where *K* is set by the user. Analysis assigns a group membership coefficient profile (*Q*) for each individual. As with other methods in this guide, analysis using dominant markers is based on an approximation, since heterozygotes are unable to be identified. STRUCTURE however includes a Bayesian method for calculating allele frequencies, which should provide reasonable estimates (see Chapter 2).

In Fig. 9.1, we show the results of STRUCTURE for our example *W. ugandensis* data set (see further information in Chapter 1; values based on all 185 AFLP loci scored), setting values of *K* from 2 to 5 (the same information can be obtained from the results file under the heading of "inferred ancestry of individuals"). In our example, setting *K* = 2 divides samples into regions of collection in a manner consistent with previous clustering (Chapter 4), ordination (Chapters 5 and 7) and AMOVA results (Chapter 8). In the figure, Laikipia, Mara (Masai Mara) and Lushoto individuals collected from east of the Rift Valley are indicated predominantly in green, while Kitale and Kibale individuals from the west of the Rift Valley are coloured primarily red. Setting *K* = 3 indicates that the next most significant level of structuring is between Kitale (predominantly in green) and Kibale (predominantly in blue), as was also evident in previous ordination (e.g., see Fig. 7.1 in Chapter 7). Setting *K* = 4 indicates a degree of structure within material collected from east of the Rift Valley, but no clear pattern is evident since each eastern population contains 'green' and 'yellow' elements. The case is similar when *K* = 5. The main difference between results with *K* = 4 and *K* = 5 is that some individuals from Kibale (see blue element) are discriminated from other stands.

An important part of STRUCTURE analysis is to settle on the 'correct' *K* value to use. The suggested approach (Falush et al. 2007) is to adopt the lowest value for *K* for which the LN P(D) values calculated by the package have begun to plateau, and for which results appear biologically meaningful. On this basis, in our example we would settle on either *K* = 3, which explains the main disjunctions in data revealed by other analyses, or *K* = 4, when the LN P(D) values appear to have plateaued.

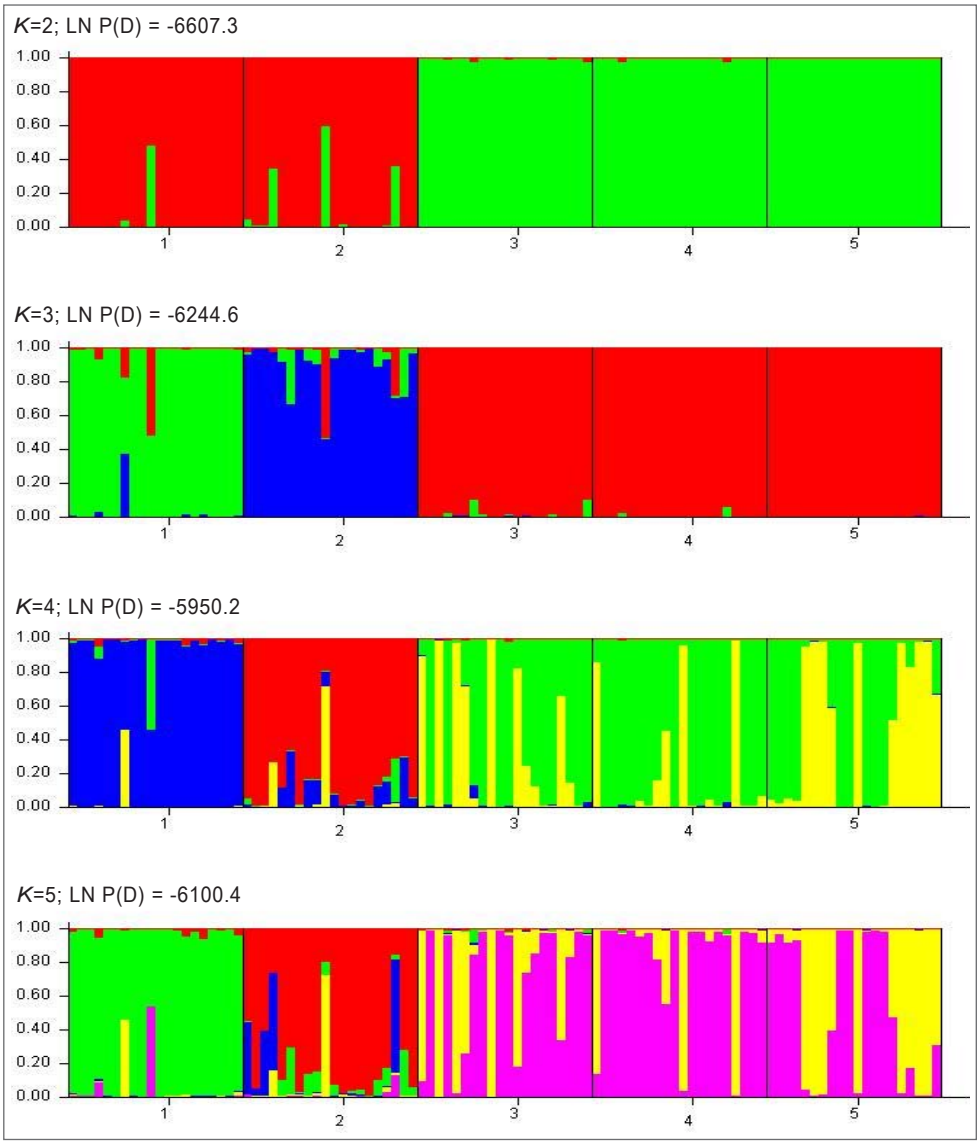


Figure 9.1. STRUCTURE analysis showing *Q* profiles (on the *y* axis, components of *Q* represented by different colours) for 100 *W. ugandensis* individuals taken from five populations, based on different values of *K* (the number of assumed groups in analysis). Individuals are arranged by the population they belong to (1 = Kitale, 2 = Kibale, 3 = Laikipia, 4 = Masai Mara, 5 = Lushoto). LN P(D) values are also shown. Analysis used the admixture ancestry model, the correlated allele frequencies model, and 100,000 steps during burnin, with 100,000 subsequent MCMC steps. Analysis was carried out in the STRUCTURE 2.3.2 software package.

Also important for obtaining meaningful results in STRUCTURE is to decide how long to run the program for, in terms of the number of steps to use in the 'burnin' period and then the number of steps to use in subsequent analysis. The authors of STRUCTURE suggest using between 10,000 and 100,000 steps in both phases. The more steps that are used then the longer analysis will take, and so users may want to minimise the number of steps employed to still obtain accurate results. The number of steps required can be tested by repeating analysis with different number of steps and checking to see when values become stable.

In Fig. 9.2, we show STRUCTURE estimates by iteration (100,000 steps for each of the burnin and subsequent analysis phases) for our example *Warburgia* data set, setting values of *K* at 3 and 4. The figure helps us in concluding that stability was reached well before the burnin phase was completed and suggests that lower values for repetitions could have been employed.

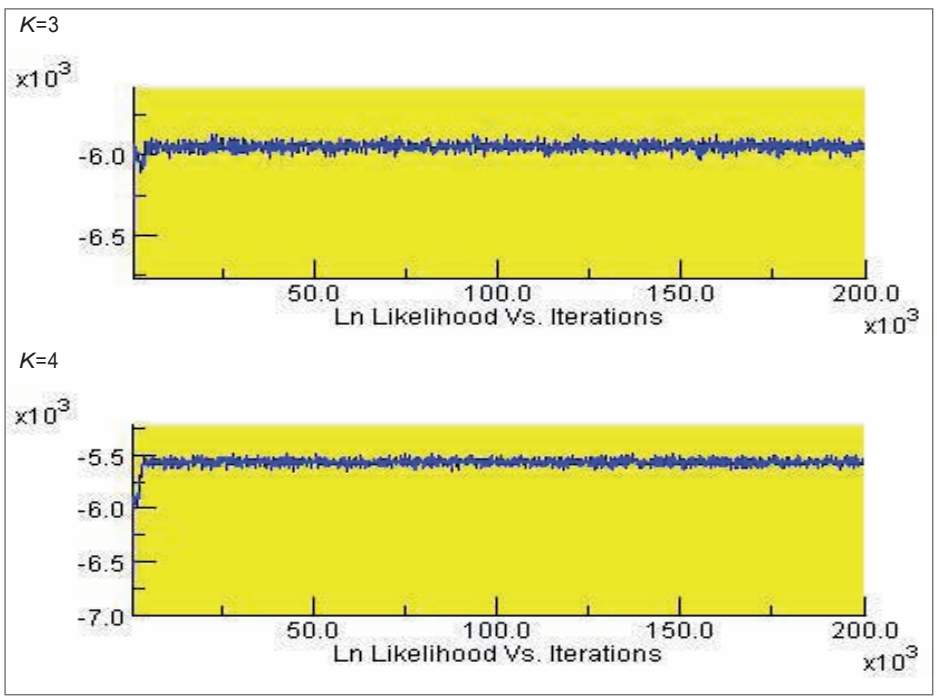


Figure 9.2. STRUCTURE analysis showing Ln Likelihood values as a function of the number of iterations in analysis, based on different values of *K* (the number of assumed groups in analysis). Analysis was carried out in the STRUCTURE 2.3.2 software package.

9.2. Identifying unusual individuals in data sets

STRUCTURE allows separate labels to be assigned to individuals in *Q* profile plots and this approach can be used to identify unusual individuals in analysis that are strongly admixtured (mixed *Q* profiles) or have completely different profiles (apparently 'misplaced' to a population). Fig. 9.3 shows the previous STRUCTURE analysis of our example *Warburgia* data set (Fig. 9.1) with individual labels assigned (*K* set at 3).

Fig. 9.3 identifies three individuals with strong admixture (not belonging mainly to a single group) as Kit07 (individual 7 in the figure; admixture of each group), Kit10 (individual 10; admixture of 'green' typical of Kitale and 'red' typical of the east of the Rift Valley) and Kib10 (individual 30; admixture of 'blue' typical of Kibale and 'red' typical of the east of the Rift Valley). These three individuals were also detected as unusual in previous ordination (see Chapter 7). The correspondence of STRUCTURE with ordination means that we can be more confident about conclusions that these individuals are indeed genetically distinct.

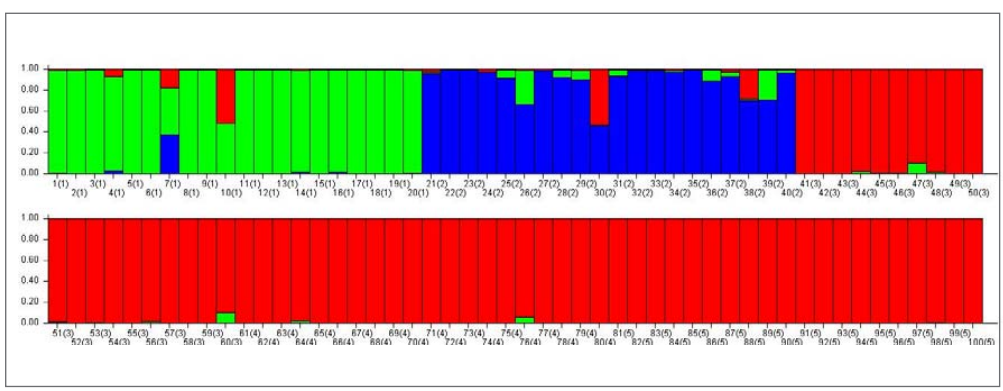


Figure 9.3. STRUCTURE analysis showing *Q* profiles (on the *y* axis, components of *Q* represented by different colours) for 100 *W. ugandensis* individuals taken from five populations, based on *K* = 3 (see Fig. 9.1) and labelling by individual (1 to 20 = Kitale, 21 to 40 = Kibale, 41 to 60 = Laikipia, 61 to 80 = Masai Mara, 81 to 100 = Lushoto). Analysis used the admixture ancestry model, the correlated allele frequencies model, and 100,000 steps during burnin, with 100,000 subsequent MCMC steps. Analysis was carried out in the STRUCTURE 2.3.2 software package.

9.3. References

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* 7: 574-578.

9.4. Suggested software

Analysis is conducted in the software package STRUCTURE (see Appendix III).

Appendix I

Mathematical formulae

Key points

We here list a range of mathematical formulae that were not included in the main body of the chapters of this guide.

Knowing how a statistic is calculated is important for knowing how results should be interpreted.

Worked examples are provided in associated Microsoft Excel workbooks on the CD-ROM that accompanies this guide.

I.1. Formulae relevant to Part 2 of the guide

I.1.1. Allele frequencies, non-Bayesian methods

Formula	References
$q_i = \sqrt{f_0}$	(the square-root method) Kraus 2000, Kremer et al. 2005
$q_i = f_0$	(the phenotypic diversity method) Kraus 2000, Kremer et al. 2005
$p_i = 1 - q_i$	(frequency of the plus-allele)
$f_0 = \frac{N_0}{N}$	(frequency of the recessive phenotypes)

Formulae are based on the following parameters:

i locus identification number
N the number of individuals with information on locus i
N_0 the number of individuals that do not show the marker for locus i

I.1.2. Allele frequencies, Bayesian methods

Formula	References
$q_i = \frac{\beta(N_0+1, N_1+1)}{\beta(N_0+0.5, N_1+1)}$	(uniform priors) Zhivotovsky 1999 (eq. 9)
$q_i = \exp\left(\ln\Gamma(N_0+1) + \ln\Gamma(N+1.5) - \ln\Gamma(N_0+0.5) - \ln\Gamma(N+2)\right)$	(uniform priors, alternative calculation method) Zhivotovsky 1999 (eq. 10)
$q_i = \frac{\beta(N_0+a+0.5, N_1+b)}{\beta(N_0+a, N_1+b)}$	(non-uniform priors) Zhivotovsky 1999 (eq. 5)
$q_i = \exp\left(\ln\Gamma(N_0+a+0.5) + \ln\Gamma(N+a+b) - \ln\Gamma(N_0+a) - \ln\Gamma(N+a+b+0.5)\right)$	(non-uniform priors, alternative calculation method) Zhivotovsky 1999 (eq. 10)
$a = \overline{R}\left(\frac{\overline{R}(1-\overline{R})}{\sigma_R^2} - 1\right)$	(parameter of beta or gamma distribution) Zhivotovsky 1999 (eq. 13)
$b = (1-\overline{R})\left(\frac{\overline{R}(1-\overline{R})}{\sigma_R^2} - 1\right)$	(parameter of beta or gamma distribution) Zhivotovsky 1999 (eq. 13)
$\overline{R} = \sum_{i=1}^{L} \frac{N_i}{N_L} \frac{N_{i0}}{N_i}$	(average frequency of recessive phenotype over loci) Zhivotovsky 1999 (eq. 12 and note 4)
$\sigma_R^2 = \left(\sum_{i=1}^{L} \frac{N_i}{N_L}\left(\frac{N_{i0}}{N_i}\right)^2\right) - \overline{R}^2$	(variance in frequencies of recessive phenotype over loci) Zhivotovsky 1999 (eq. 12 and note 4)
$N_L = \sum_{i=1}^{L} N_i$	

Formulae are based on the following parameters:

i	locus identification number
N_0	the number of individuals that do not show the marker for locus i
N_1	the number of individuals that show the marker for locus i
N	the total number of individuals with information for locus i
N_i	the number of individuals with information on locus i
N_{i0}	the number of individuals that do not show the marker for locus i
L	the number of loci

I.1.3. Diversity statistics

Formula	References
$H_e = 1 - p_i^2 - q_i^2$	(Nei diversity = expected heterozygosity) Nei 1978, Bonin et al. 2007
$UH_e = \frac{2N}{2N-1}\left(1 - p_i^2 - q_i^2\right)$	(unbiased Nei diversity) Nei 1978, Bonin et al. 2007
$\hat{H}_{j(i)} = 2q_i(1-q_i) + 2\frac{f_1}{4N}$	(unbiased Nei diversity, alternative calculation method) Lynch and Milligan 1994 (eqs. 2b, 4a)
$N_e = \frac{1}{p_i^2 + q_i^2}$	(number of effective alleles) Peakall and Smouse 2006 (GenAlEx results)
$I = -\left(p_i \ln(p_i) + q_i \ln(q_i)\right)$	(Shannon diversity index) Bonin et al. 2007

Formulae are based on the following parameters:

i	locus identification number
p_i	the frequency of the plus-allele of locus *i*
q_i	the frequency of the null-allele of locus *i*
N	the number of individuals
j	population number
f_1	the frequency of the dominant phenotype

I.1.4. Summing diversity across loci

Formula	Reference
$H_e = \frac{\sum_{i=1}^{L}(1 - p_i^2 - q_i^2)}{L}$	(average Nei diversity) Nei 1978, Bonin et al. 2007
$UH_e = \frac{2N}{2N-1}\frac{\sum_{i=1}^{L}(1 - p_i^2 - q_i^2)}{L}$	(average unbiased Nei diversity) Nei 1978, Bonin et al. 2007
$\hat{H}_j = \frac{\sum_{i=1}^{L}\hat{H}_{j(i)}}{L}$	(average unbiased Nei diversity, alternative calculation method) Lynch and Milligan 1994 (eq. 5)
$\hat{H}_{j(i)} = 2q_i(1 - q_i) + 2\frac{f_1}{4N}$	(unbiased Nei diversity, alternative calculation method) Lynch and Milligan 1994 (eqs. 2b, 4a)

Formulae are based on the following parameters:

i	locus identification number
p_i	the frequency of the plus-allele for locus *i*
q_i	the frequency of the null-allele for locus *i*
N	the number of individuals
L	the number of loci
j	population number
f_1	the frequency of the dominant phenotype for locus *i*

I.1.5. Genetic distances between populations

Formula	Reference
$Nei_s = -\ln \frac{\sum_{i=1}^{L} p_{i1}p_{i2} + q_{i1}q_{i2}}{\sqrt{\sum_{i=1}^{L}\left(p_{i1}^2 q_{i1}^2\right)\sum_{i=1}^{L}\left(p_{i2}^2 q_{i2}^2\right)}}$	(Standard Nei) Nei 1978 (eq. 4), Reif et al. 2005
$Nei_{us} = -\ln \frac{\sum_{i=1}^{L} p_{i1}p_{i2} + q_{i1}q_{i2}}{\sqrt{\sum_{i=1}^{L}\left(\frac{2N_1}{2N_1-1} p_{i1}^2 q_{i1}^2\right)\sum_{i=1}^{L}\left(\frac{2N_2}{2N_2-1} p_{i2}^2 q_{i2}^2\right)}}$	(Unbiased estimation of standard Nei) Nei 1978 (eq. 6)
$Nei_m = \frac{\sum_{i=1}^{L}\left(p_{i1}^2 q_{i1}^2\right)}{2} + \frac{\sum_{i=1}^{L}\left(p_{i2}^2 q_{i2}^2\right)}{2} - \sum_{i=1}^{L}\left(p_{i1}p_{i2} + q_{i1}q_{i2}\right)$	(Minimum Nei) Nei 1978
$Nei_{um} = \frac{\sum_{i=1}^{L}\left(\frac{2N_1}{2N_1-1} p_{i1}^2 q_{i1}^2\right)}{2} + \frac{\sum_{i=1}^{L}\left(\frac{2N_2}{2N_2-1} p_{i2}^2 q_{i2}^2\right)}{2} - \sum_{i=1}^{L}\left(p_{i1}p_{i2} + q_{i1}q_{i2}\right)$	(Unbiased estimation of minimum Nei) Nei 1978 (eq. 12)
$TakezakiNei = \frac{2}{\Pi L}\sum_{i=1}^{L}\sqrt{2\left(1 - \sqrt{p_{i1}p_{i2}} - \sqrt{q_{i1}q_{i2}}\right)}$	Takezaki and Nei 1996
$Rogers_o = \frac{\sqrt{\sum_{i=1}^{L}\left(\frac{(p_{i1}-p_{i2})^2 + (q_{i1}-q_{i2})^2}{2}\right)}}{L}$	(Original Rogers) Rogers 1972, Reif et al. 2005
$Rogers_m = \frac{\sqrt{(p_{i1}-p_{i2})^2 + (q_{i1}-q_{i2})^2}}{\sqrt{2L}}$	(Modified Rogers) Wright 1978, Reif et al. 2005

Formulae are based on the following parameters:

i	locus identification number
L	the number of loci
p_{i1}	the frequency of the plus-allele for locus i in the first population
q_{i1}	the frequency of the null-allele for locus i in the first population
p_{i2}	the frequency of the plus-allele for locus i in the second population
q_{i2}	the frequency of the null-allele for locus i in the second population
N_1	the number of individuals in the first population
N_2	the number of individuals in the second population

I.2. Formulae relevant to Part 3 of the guide

Formula	References
$Jaccard.Dist = \frac{N_{10} + N_{01}}{N_{11} + N_{10} + N_{01}}$	Jaccard 1908, Kosman and Leonard 2005, Reif et al. 2005, Bonin et al. 2007
$Dice.Dist = \frac{N_{10} + N_{01}}{2N_{11} + N_{10} + N_{01}}$	Dice 1945, Sørenson 1948 , Kosman and Leonard 2005, Reif et al. 2005, Bonin et al. 2007
$Simple.Match.Dist = \frac{N_{10} + N_{01}}{N_{11} + N_{10} + N_{01} + N_{00}}$	Sokal and Michener 1958, Kosman and Leonard 2005, Reif et al. 2005, Bonin et al. 2007
$Euclidean.Dist = \sqrt{N_{10} + N_{01}}$	Legendre and Legendre 1998
$Squared.Euclidean.Dist = N_{10} + N_{01}$	Legendre and Legendre 1998

Formulae are based on the following parameters:

N_{11} the number of loci with shared presences for both individuals
N_{10} the number of loci with presence for the first individual and absence for the second individual
N_{01} the number of loci with absence for the first individual and presence for the second individual
N_{00} the number of loci with shared absences for both individuals

I.3. Formulae relevant to Part 4 of the guide

I.3.1. Analysis of molecular variance (AMOVA)

Formula	Reference
$\Phi = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2}$	(*F*-statistic) Excoffier et al. 1992
$\sigma_a^2 = \frac{MSAP - MSWP}{n}$	(the estimated variance for the among population level)
$n = \frac{N - \sum_{i=1}^{P} \frac{N_i^2}{N}}{P-1}$	
$\sigma_w^2 = MSWP = \frac{SSWP}{N-P}$	(the mean square and estimated variance for the within population level)
$MSAP = \frac{SSAP}{P-1}$	(the mean square for the among population level)
$SSAP = SSTOT - SSWP$	(the total sum of squared distances among populations)
$SSTOT = \sum_{i=1}^{P}\sum_{j=1}^{P}\sum_{a=1}^{N_i}\sum_{b=1}^{N_j} D_{ia-jb}^2$	(the total sum of squared distances)
$SSWP = \sum_{i=1}^{P} SSD(P_i)$	(the sum of squared distances within all populations)
$SSD(P_i) = \sum_{a=1}^{N_i}\sum_{b=1}^{N_i} D_{ia-ib}^2$	(the sum of squared distances within population *X*)

Formulae are based on the following parameters:

$D_{ia\text{-}jb}$ the Euclidean distance between individual *a* of population *i* and individual *b* of population *j*
N_i the number of individuals in population i
N the number of individuals in all populations
P the number of populations

I.3.2. *G*-statistics

Formula	Reference
$G_{ST} = \frac{H_T - H_W}{H_T}$	(*F*-statistic) Nei 1973
$H_W = \frac{\sum_{j=1}^{n} H_j}{n}$	(Average diversity of the populations)
$H_j = \frac{\sum_{i=1} (1 - p_{ij}^2 - q_{ij}^2)}{L}$	(See formula for H_e in section 2.2)
$p_{ia} = \frac{\sum_{j=1}^{n} p_{ij}}{n}$	(Average frequency of the plus-allele of the populations at the locus)
$q_{ia} = \frac{\sum_{j=1}^{n} q_{ij}}{n}$	(Average frequency of the null-allele of the populations at the locus)
$H_T = \frac{\sum_{i=1}^{L} (1 - p_{ia}^2 - q_{ia}^2)}{L}$	(See formula for H_e in section 2.2)

Formulae are based on the following parameters:

j	population identification number
n	number of populations
i	locus identification number
L	the number of loci
p_{ij}	the frequency of the plus-allele for locus i in population j
q_{i1}	the frequency of the null-allele for locus i in population j

I.3.3. θ-statistics

Formula	Reference
$\hat{\theta} = \frac{\sum_{i=1}^{L} \theta_{in}}{\sum_{i=1}^{L} \theta_{id}}$	(*F*-statistic) Weir and Cockerham 1984 (top of page 1363) (this is the formula for random mating within populations)
$\theta_{in} = s_i^2 - \frac{1}{2\overline{N}-1}\left[p_i(1-p_i) - \frac{n-1}{n} s_i^2 \right]$	(nominator part of the formula on page 1363 of the Weir and Cockerham article)
$\theta_{id} = \left[1 - \frac{2\overline{N}C^2}{(2\overline{N}-1)n} \right] p_i(1-p_i) + \left[1 + \frac{2\overline{N}(n-1)C^2}{(2\overline{N}-1)n} \right] \frac{s_i^2}{n}$	(denominator part of the formula on page 1363 of the Weir and Cockerham article)
$C = \frac{StDev(N_j)}{\overline{N}}$	(the coefficient of variation for the number of individuals in each population)
$s_i^2 = \frac{\sum_{j=1}^{n} N_i (p_j - p_i)}{(n-1)\overline{N}}$	(the variance of the plus-allele in the populations)
$p_i = \frac{\sum_{j=1}^{n} N_j p_j}{n\overline{N}}$	(the average frequency of the plus-allele in the populations)
$\overline{N} = \frac{\sum_{j=1}^{n} N_j}{n}$	(the average number of individuals in a population)

Formulae are based on the following parameters:

i	locus identification number
L	the number of loci
j	population identification number
n	the number of populations
N_j	the number of individuals in population *j*
p_{ij}	the frequency of the plus-allele for locus *i* in population *j*

I.4. References

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology* 16: 3737-3758.

Dice LR (1945) Measures of the amount of ecological association between species. *Ecology* 26: 297-302.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479-491.

Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise Sci. Nat.* 44: 223-270.

Kosman E, Leonard KJ (2005) Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploidy species. *Molecular Ecology* 14: 415-424.

Krauss SL (2000) Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. *Molecular Ecology* 9: 1241-1245.

Kremer A, Caron H, Cavers S, Colpaert N, Gheysen G, Gribel R, Lemes M, Lowe AJ, Margis R, Navarro C, Salgueiro F (2005) Monitoring genetic diversity in tropical trees with multilocus dominant markers. *Heredity* 95: 274-280.

Legendre P, Legendre L (1998) Numerical Ecology. Developments in Ecological Modelling 20. Elsevier, Amsterdam, The Netherlands, 853 pp.

Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology* 3: 91-99.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* 70: 3321-3323.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583-590.

Peakall R, Smouse PE (2006) GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288-295.

Reif JC, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science* 45: 1-7.

Rogers JS (1972) Measures of genetic similarity and genetic distance. In: Studies in Genetics VII. University of Texas, Austin, USA, pp. 145-154.

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 38: 1409-1438.

Sørenson T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biol. Skr.* 5: 1-34.

Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* 144: 389-399.

Weir BS, Cockerham CC (1984) Estimation of *F*-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.

Wright S (1978) Evolution and the Genetics of Populations. Volume 4: Variability Within and Among Natural Populations. University of Chicago Press, Chicago, USA.

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* 8: 907-913.

Appendix II

Installing software and formatting data

Key points

We here describe how to download and install nine different data analysis packages that are available for free on the internet. Also described is how to format data for analysis in each of them.

Formatted input files for different software packages based on our *W. ugandensis* example data set are included in folders on the CD-ROM. Typical output files are also included.

How to carry out analyses in a subset of the packages presented here is described in Appendix III.

Software packages are continually being updated. We recommend that users always install the latest version available (using any new installation methods). Check the guides of new software versions for any new features and procedures for analysis.

II.1. General formatting

For basic information on formatting, please refer to Chapter 1 of this guide. Table II.1 provides again the information given there for the first 45 individuals tested and the first five loci scored for our example *Warburgia ugandensis* data set. This spreadsheet can be used as the initial input file to prepare data for a range of software packages, as illustrated in Fig. II.1. Our basic data spreadsheet can be found as *warburgiabase.xls* or *warburgiabase.txt* on the CD-ROM accompanying this guide.

Table II.1. A subset of AFLP data (45 individuals and 5 loci) collected for *Warburgia ugandensis,* showing the appropriate format for inputting results into a spreadsheet.

Individual	Population	Region	Locus001	Locus002	Locus003	Locus004	Locus005
Kit01	Kitale	west	0	0	0	0	0
Kit02	Kitale	west	0	1	0	1	0
Kit03	Kitale	west	0	0	0	1	0
Kit04	Kitale	west	0	0	0	1	0
Kit05	Kitale	west	0	0	0	1	0
Kit06	Kitale	west	0	0	0	1	0
Kit07	Kitale	west	0	1	0	1	0
Kit08	Kitale	west	0	0	0	1	0
Kit09	Kitale	west	0	0	0	0	0
Kit10	Kitale	west	0	0	0	1	0
Kit11	Kitale	west	0	0	0	1	0
Kit12	Kitale	west	0	0	0	1	0
Kit13	Kitale	west	0	0	0	0	0
Kit14	Kitale	west	0	0	0	0	0
Kit15	Kitale	west	0	0	0	0	0
Kit16	Kitale	west	0	0	0	0	0
Kit17	Kitale	west	0	0	0	0	0
Kit18	Kitale	west	0	0	0	0	0
Kit19	Kitale	west	0	0	0	0	0
Kit20	Kitale	west	0	0	0	0	0
Kib01	Kibale	west	0	0	0	0	0
Kib02	Kibale	west	0	0	0	0	0
Kib03	Kibale	west	0	0	0	0	0
Kib04	Kibale	west	0	0	0	1	0
Kib05	Kibale	west	0	0	0	0	0
Kib06	Kibale	west	0	0	0	0	0
Kib07	Kibale	west	0	1	0	1	0
Kib08	Kibale	west	0	0	0	1	0
Kib09	Kibale	west	0	0	0	1	0
Kib10	Kibale	west	0	1	0	1	0
Kib11	Kibale	west	0	0	0	0	0
Kib12	Kibale	west	0	0	0	0	0
Kib13	Kibale	west	0	0	0	0	0
Kib14	Kibale	west	0	0	0	1	0
Kib15	Kibale	west	0	0	0	0	0
Kib16	Kibale	west	0	0	0	0	0
Kib17	Kibale	west	0	1	0	1	0
Kib18	Kibale	west	0	0	1	1	0
Kib19	Kibale	west	1	0	0	0	0
Kib20	Kibale	west	1	0	0	1	0
Lai01	Laikipia	east	0	1	0	1	0
Lai02	Laikipia	east	0	1	0	1	0
Lai03	Laikipia	east	0	0	0	0	0
Lai04	Laikipia	east	0	0	0	0	0
Lai05	Laikipia	east	0	0	0	1	0

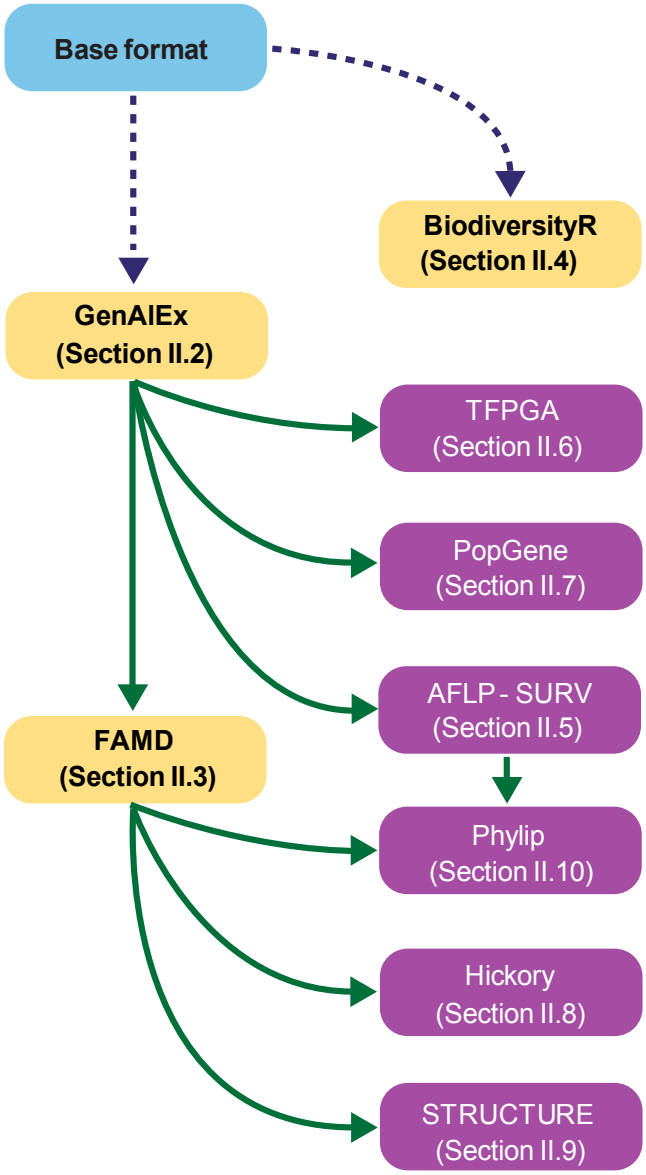


Figure II.1. Suggested pathways for preparing input data files for various software packages. Green arrows indicate where one software package can prepare data in the right format for input into another package. For two packages, GenAlEx and BiodiversityR, start with the base worksheet format.

II.2. GenAlEx (Genetic Analysis in Excel) (version 6.2)

II.2.1. Installation

GenAlEx can be obtained from the following website:
http://www.anu.edu.au/BoZo/GenAlEx/

You will obtain a zip archive named "GenAlEx 6.2.zip". Extract all the files of the zip archive to the same folder (e.g., "C:\Program Files\GenAlEx 6.2").

One of the extracted files is named "GenAlEx 6.2.xla", which is an Add-In for Microsoft Excel. Install this Add-In after launching Excel using the menu option: Tools > Add-Ins… Click on the <browse> button and point to the location of the "GenAlEx 6.2.xla" file.

The default option is that the GenAlEx start screen is shown every time Excel is started. To prevent this, use the following option: GenAlEx > Options > Generic > Hide Splash Screen on Startup.

If you are using Microsoft Excel 2007, please check the guidelines provided in the zip archive "Read Me File GenAlEx 6.1.pdf".

The suggested citation for the software is: Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6: 288-295.

II.2.2. Preparing input data

The data set must be sorted by regions and then by populations. Regions cannot be mixed down the order of rows – that is, all of the populations from the first region must come first, then populations from the second region, etc.

Follow these steps to prepare input data.

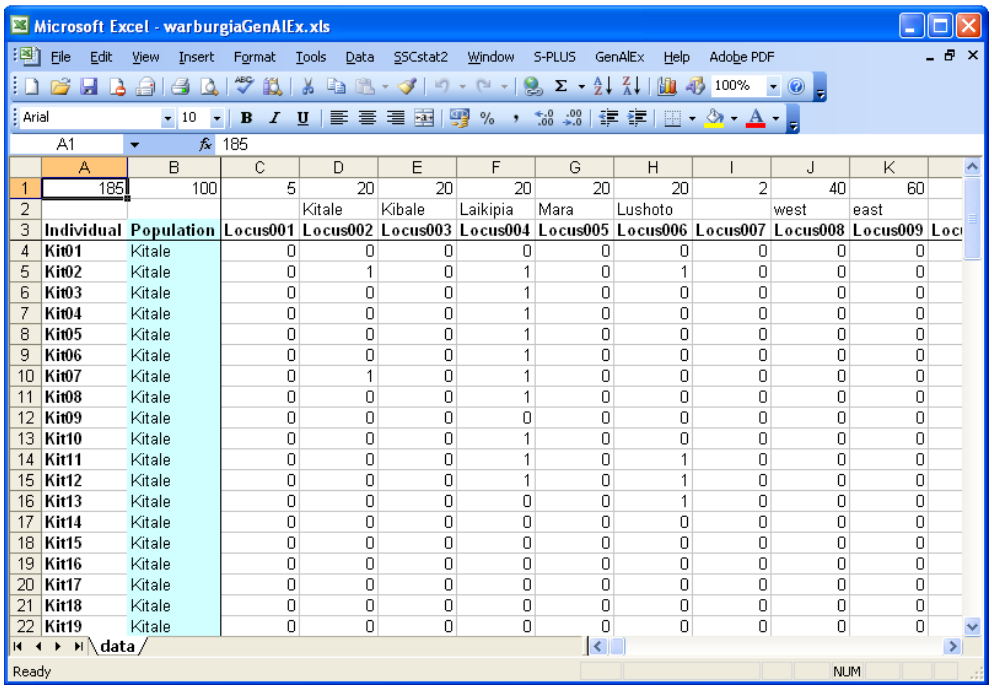
- Make a copy of the base data set ("*warburgiabase.xls*"), such as ("*warburgiaGenAlEx.xls*")
- Open the copied data set in Microsoft Excel.
- Delete the third column (column C with "Rift" in the top cell).
- Insert two rows at the top of the sheet.
- Select the menu option of GenAlEx > Parameters > Pops from Col2.
- Complete information about the number of loci using the menu option GenAlEx > No. Loci.
- Complete information about regional structure (cells I1, J1-2 and K1-2) as shown in Fig. II.2.
- Save this data set.

The *Warburgia* spreadsheet does not contain missing data. If this were the case, then a value of "-1" should be given for missing data points.

II.2.3. Exporting data from GenAlEx

GenAlEx allows data to be exported in the correct input formats for the following software packages described here: FAMD, AFLP-SURV, TFPGA and PopGene.

To access output options, the worksheet with the raw input data should be active. Export options are accessed via the menu option: GenAlEx > Export Data.



	A	B	C	D	E	F	G	H	I	J	K
1	185	100	5	20	20	20	20	20	2	40	60
2				Kitale	Kibale	Laikipia	Mara	Lushoto		west	east
3	**Individual**	**Population**	**Locus001**	**Locus002**	**Locus003**	**Locus004**	**Locus005**	**Locus006**	**Locus007**	**Locus008**	**Locus009**
4	**Kit01**	Kitale	0	0	0	0	0	0	0	0	0
5	**Kit02**	Kitale	0	1	0	1	0	1	0	0	0
6	**Kit03**	Kitale	0	0	0	1	0	0	0	0	0
7	**Kit04**	Kitale	0	0	0	1	0	0	0	0	0
8	**Kit05**	Kitale	0	0	0	1	0	0	0	0	0
9	**Kit06**	Kitale	0	0	0	1	0	0	0	0	0
10	**Kit07**	Kitale	0	1	0	1	0	0	0	0	0
11	**Kit08**	Kitale	0	0	0	1	0	0	0	0	0
12	**Kit09**	Kitale	0	0	0	0	0	0	0	0	0
13	**Kit10**	Kitale	0	0	0	1	0	0	0	0	0
14	**Kit11**	Kitale	0	0	0	1	0	1	0	0	0
15	**Kit12**	Kitale	0	0	0	1	0	1	0	0	0
16	**Kit13**	Kitale	0	0	0	0	0	1	0	0	0
17	**Kit14**	Kitale	0	0	0	0	0	0	0	0	0
18	**Kit15**	Kitale	0	0	0	0	0	0	0	0	0
19	**Kit16**	Kitale	0	0	0	0	0	0	0	0	0
20	**Kit17**	Kitale	0	0	0	0	0	0	0	0	0
21	**Kit18**	Kitale	0	0	0	0	0	0	0	0	0
22	**Kit19**	Kitale	0	0	0	0	0	0	0	0	0

Figure II.2. The required header for an Excel worksheet containing input data for the GenAlEx software package. Cell A1 contains the number of loci (185), cell B1 the number of individuals (100), cell C1 the number of populations (5), cells D1-H1 the number of individuals in each population (20), cells D2-H2 the names of the populations in the same order as data are input (Kitale, Kibale, Laikipia, Mara and Lushoto), cell I1 the number of regions (2), cells J1-K1 the number of individuals in each region (40, 60) and cells J2-K2 the names of the regions (west, east).

II.3. FAMD (Fingerprint Analysis with Missing Data)

(version 1.23 beta)

II.3.1. Installation

FAMD can be downloaded from: http://www.famd.me.uk/famd.html

Extract the three files from the zip archive "FAMD123.zip" into the same directory (e.g., C:\Program Files\FAMD).

The program does not need to be further installed but executes directly by clicking on famd.exe

The suggested citation for the software is: Schlüter PM, Harris SA (2006) Analysis of multilocus fingerprinting data sets containing missing data. *Molecular Ecology Notes* 6: 569-572.

II.3.2. Preparing input data

Prepare input files via GenAlEx (section II.2) and give the export file an extension of ".txt" (e.g., "*warburgiaFAMD.txt*").

Load this input file into FAMD via the menu option: file > import.

Provide information on the format of the input file, as shown in Fig. II.3.

Next, provide information on the population and structure of the *Warburgia* stands. The group manager of FAMD adds information on group structure to the input data and is accessed via the menu option: DataMatrix > Group manager.

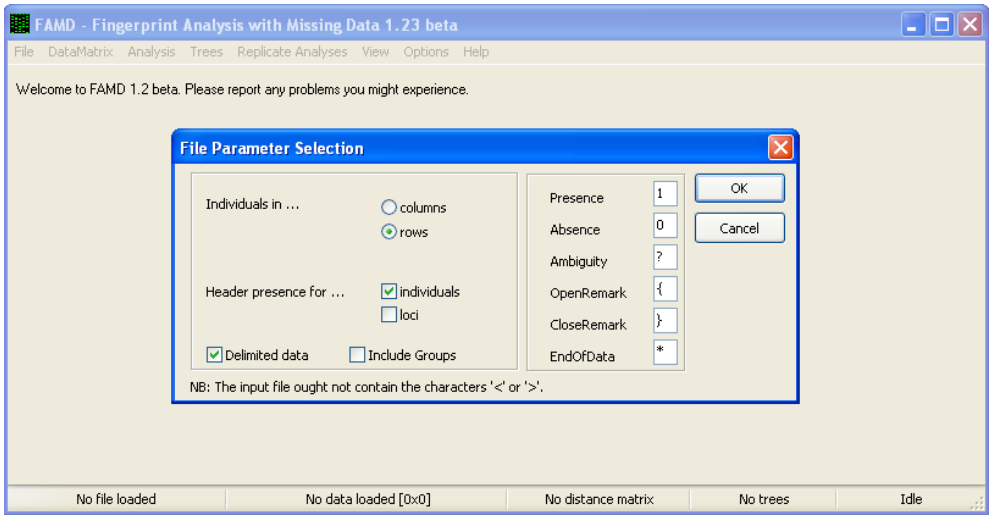


Figure II.3. Required parameters for importing the input file generated by GenAlEx into FAMD. State that individuals are provided in rows, that header presence is provided for individuals (but not for loci) and that the data is delimited.

Follow these steps:

- Select all the individuals of the same population (click on the first individual and scroll to the last individual and then SHIFT-click on it) and move these to the Data subset window using the >> button. Give the name of the population and then click <Accept subset> (Fig. II.4).
- Repeat for all populations and then click <OK>.
- As the final step, save the changed input file by the menu option: File > Save DataMatrix.

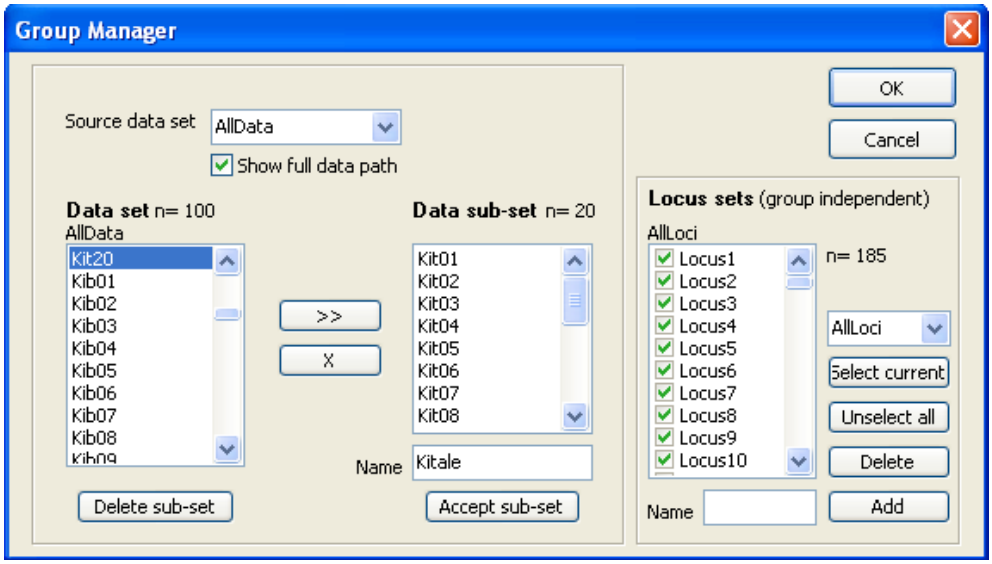


Figure II.4. Use the group manager to assign individuals to populations.

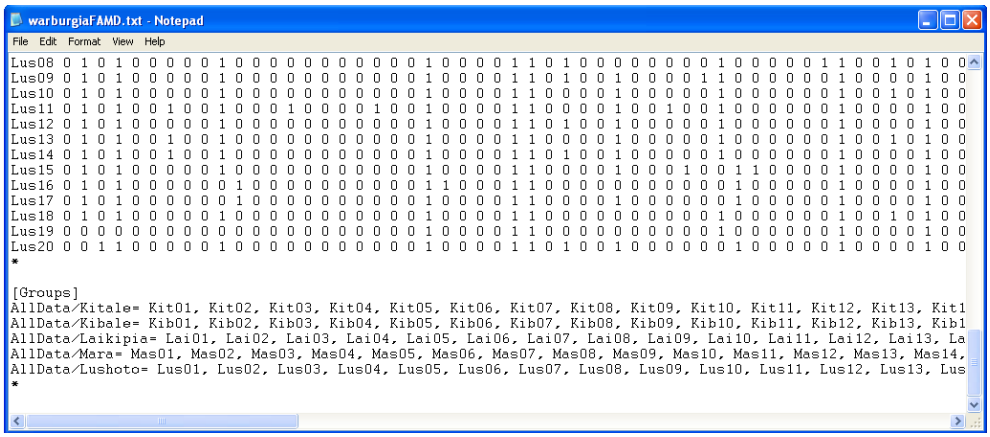


Figure II.5. After using the group manager, information on population structure is added to the input data.

The group manager will have added information on group structure at the end of the data (Fig. II.5). The next time that you import data, make sure that you specify that group structure is provided (tick "include groups", see Fig. II.3).

The *Warburgia* spreadsheet does not contain missing data. If this were the case, then a value of "?" could be used and indicated as such in the Ambiguity box (see Fig. II.3).

II.3.3. Exporting data from FAMD

FAMD allows data to be exported in the correct input formats for the following software packages described here: Hickory and STRUCTURE.

Export options are accessed via the menu option: File > Export.

II.4. BiodiversityR (version 1.4)

II.4.1. Installation

BiodiversityR is available at the following website: www.cran.r-project.org

The CD-ROM accompanying our guide contains the installation software for BiodiversityR. It can be found in the folder: Software BiodiversityR \ Installation.

BiodiversityR is a package that was developed for the R statistical environment and this needs to be installed first.

Follow these steps for installing BiodiversityR (or check the guidelines provided with the installation files on the CD-ROM):

- Download the installation file for the R statistical environment (e.g., R-2.10.0-win32.exe) from the R website (http://cran.r-project.org/bin/windows/base/), preferably from one of the mirror sites (http://cran.r-project.org/mirrors/html).
- Download all the packages that BiodiversityR builds on, from the website for 'additional packages' (http://cran.r-project.org/bin/windows/contrib/2.10/, or via the mirror sites at http://cran.r-project.org/mirrors/html): abind, akima, aplpack, BiodiversityR, car, colorspace, effects, ellipse, Hmisc, leaps, lmtest, maptree, mgcv, multcomp, mvtnorm, Rcmdr, relimp, rgl, RODBC, sp, splancs and vegan.
- Download all these packages to the same folder. Do not click on these zip files as they can only be installed from within the R system (see next step).
- Click on the installation file for R (close all other programs first).
- Follow the instructions for installation. Make sure to select the option of including the support files for tcltk (Components installation window), the option for a customised startup (Startup options installation window) and the option for the SDI (separate windows) display (Display mode window).
- Launch R.

- Select the menu option: Packages > Install packages(s) from local zip files…
- The final installation step is to select all the downloaded zip archives for the packages (use CTRL-click or CTRL-A) and click on <Open>.

Type the following commands to access BiodiversityR and its graphical user interface (note that R is case-sensitive):

```
library(BiodiversityR)
BiodiversityRGUI()
```

The suggested citation for the software is: Kindt R, Coe R (2005) Tree Diversity Analysis: A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies. The World Agroforestry Centre, Nairobi, Kenya.
URL www.worldagroforestry.org/treesandmarkets/tree_diversity_analysis.asp

BiodiversityR was initially developed as a package for community ecology and is accompanied by a manual for ecological data analysis. This manual explains various methods that are illustrated in our current guide, such as clustering and ordination.

The BiodiversityR manual is included on the CD-ROM accompanying our current guide in the folder: Software BiodiversityR \ Manual.

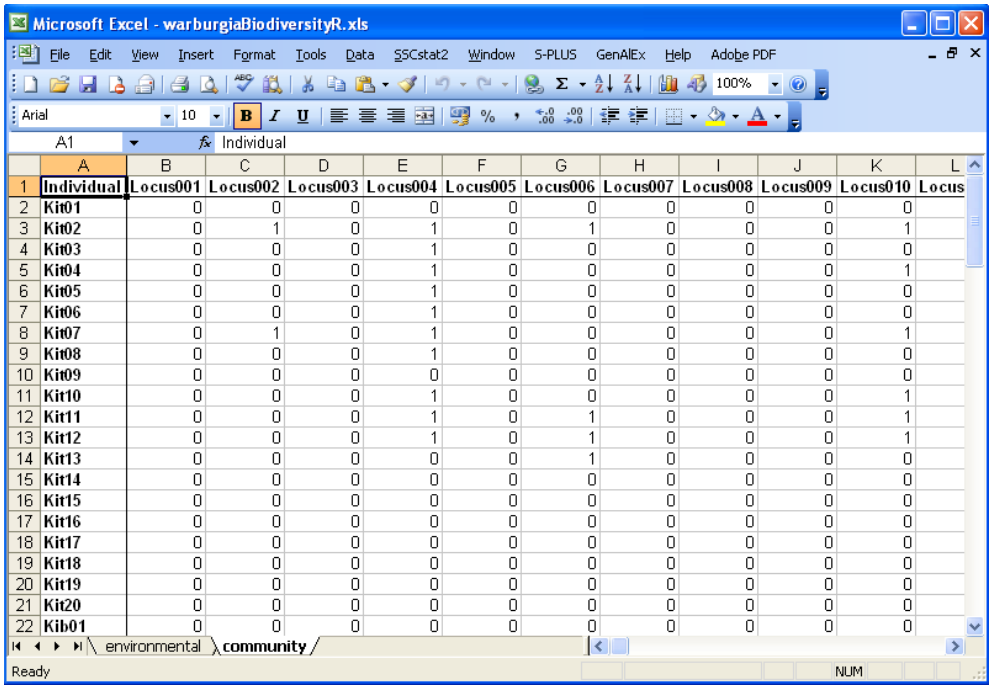
II.4.2. Preparing input data

The package uses two data sets simultaneously during analysis: a 'community' data set and an 'environmental' data set. These names refer to the origins of the package in analysing ecological data. In such analysis, the 'community' data set contains information about species composition for analysed sites, while the 'environmental' data set contains information on the environmental characteristics of sites. The package assumes that the rows of the community and environmental data sets contain information on the same sample units.

For the analysis of molecular data, marker scores are placed in the community data set and elements of 'geographical' information (population, region, etc.) are provided in the environmental data set. Unlike some other packages, it is not necessary that individuals are entered sequentially by regions and populations (but they must be in the same order for the two data sets).

Starting from the base data set of our *Warburgia* example, follow the below steps:

- Make a copy of the base data set and label it, e.g., as "*WarburgiaBiodiversityR.xls*".
- Open the copy of the data set in Microsoft Excel.
- Insert a new worksheet: Insert > Worksheet…
- Copy the information from the first three columns of the first worksheet (column A with "Individual" as the top entry, column B with "Population" as the top entry, and column C with "Region" as the top entry) to the first three columns of the second worksheet.
- Delete column B (with "Population" as the top entry) and column C (with "Region" as the top entry) from the first worksheet.
- Rename the first worksheet (this is the worksheet containing the molecular results) as "community" by double-clicking on the name of the sheet. Do not use capital letters. The final format of the worksheet should be as shown in Fig. II.6. The *Warburgia* spreadsheet does not contain missing data. If this were the case, then cells would simply have been left empty (blank).
- Rename the second worksheet (this is the worksheet containing the 'geographical' information) as "environmental" by double-clicking on the name of the sheet. Do not use capital letters. The final format of the worksheet should be as shown in Fig. II.7.
- Save the file.



Microsoft Excel - warburgiaBiodiversityR.xls

File Edit View Insert Format Tools Data SSCstat2 Window S-PLUS GenAlEx Help Adobe PDF

Arial 10 B I U 100%

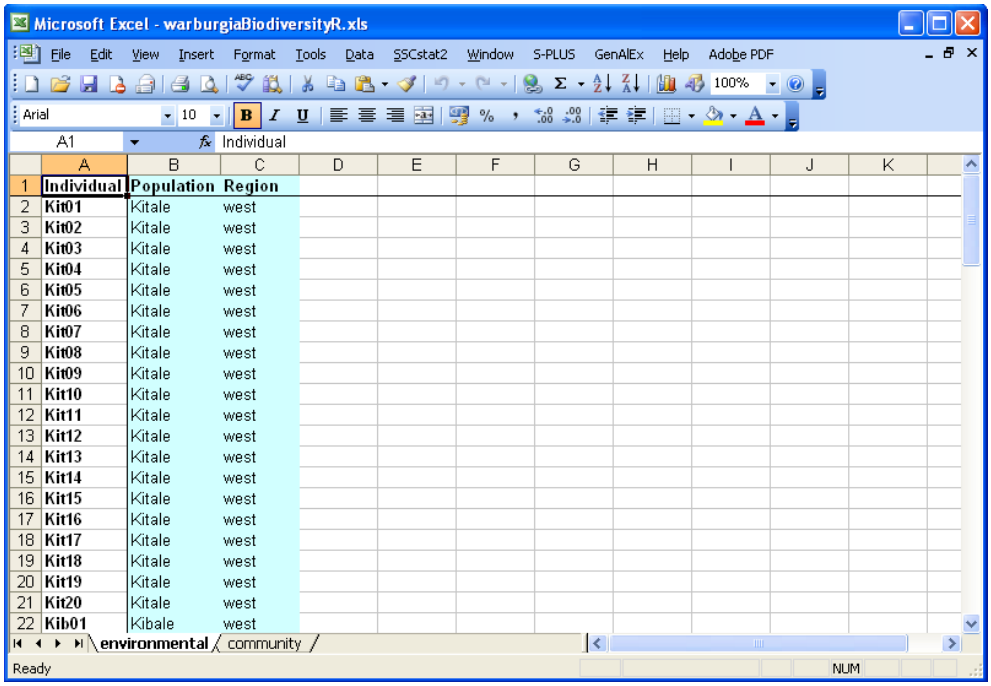
A1 Individual

	A	B	C	D	E	F	G	H	I	J	K	L
1	Individual	Locus001	Locus002	Locus003	Locus004	Locus005	Locus006	Locus007	Locus008	Locus009	Locus010	Locus
2	Kit01	0	0	0	0	0	0	0	0	0	0	
3	Kit02	0	1	0	1	0	1	0	0	0	1	
4	Kit03	0	0	0	1	0	0	0	0	0	0	
5	Kit04	0	0	0	1	0	0	0	0	0	1	
6	Kit05	0	0	0	1	0	0	0	0	0	0	
7	Kit06	0	0	0	1	0	0	0	0	0	0	
8	Kit07	0	1	0	1	0	0	0	0	0	1	
9	Kit08	0	0	0	1	0	0	0	0	0	0	
10	Kit09	0	0	0	0	0	0	0	0	0	0	
11	Kit10	0	0	0	1	0	0	0	0	0	1	
12	Kit11	0	0	0	1	0	1	0	0	0	1	
13	Kit12	0	0	0	1	0	1	0	0	0	1	
14	Kit13	0	0	0	0	0	1	0	0	0	0	
15	Kit14	0	0	0	0	0	0	0	0	0	0	
16	Kit15	0	0	0	0	0	0	0	0	0	0	
17	Kit16	0	0	0	0	0	0	0	0	0	0	
18	Kit17	0	0	0	0	0	0	0	0	0	0	
19	Kit18	0	0	0	0	0	0	0	0	0	0	
20	Kit19	0	0	0	0	0	0	0	0	0	0	
21	Kit20	0	0	0	0	0	0	0	0	0	0	
22	Kib01	0	0	0	0	0	0	0	0	0	0	

environmental community

Ready NUM

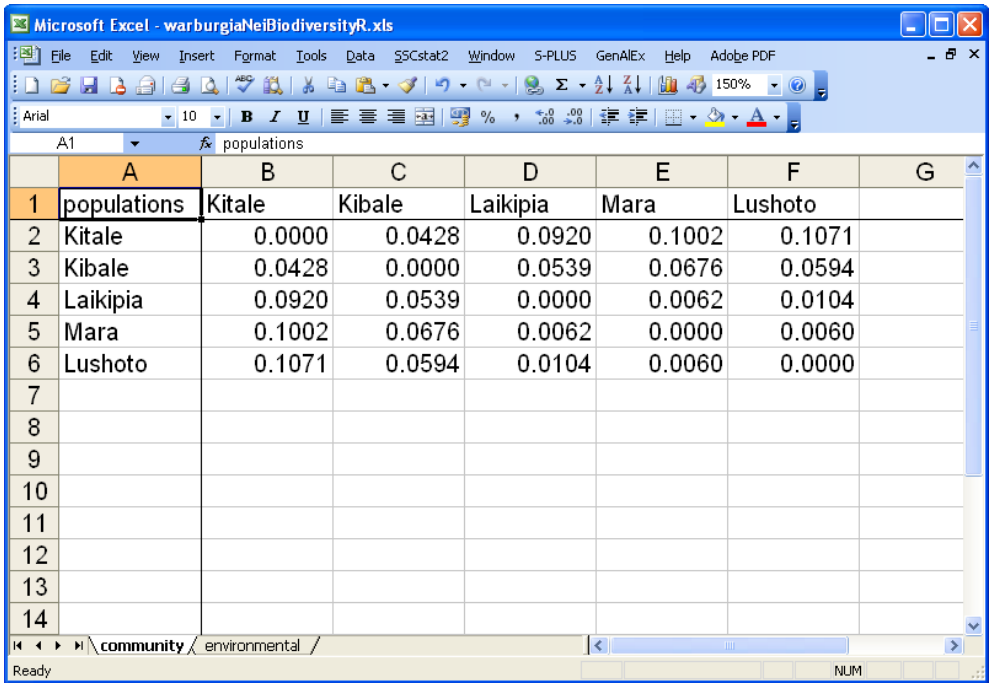
Figure II.6. Format for the "community" worksheet that contains molecular data for input into BiodiversityR.



	A	B	C
1	Individual	Population	Region
2	Kit01	Kitale	west
3	Kit02	Kitale	west
4	Kit03	Kitale	west
5	Kit04	Kitale	west
6	Kit05	Kitale	west
7	Kit06	Kitale	west
8	Kit07	Kitale	west
9	Kit08	Kitale	west
10	Kit09	Kitale	west
11	Kit10	Kitale	west
12	Kit11	Kitale	west
13	Kit12	Kitale	west
14	Kit13	Kitale	west
15	Kit14	Kitale	west
16	Kit15	Kitale	west
17	Kit16	Kitale	west
18	Kit17	Kitale	west
19	Kit18	Kitale	west
20	Kit19	Kitale	west
21	Kit20	Kitale	west
22	Kib01	Kibale	west

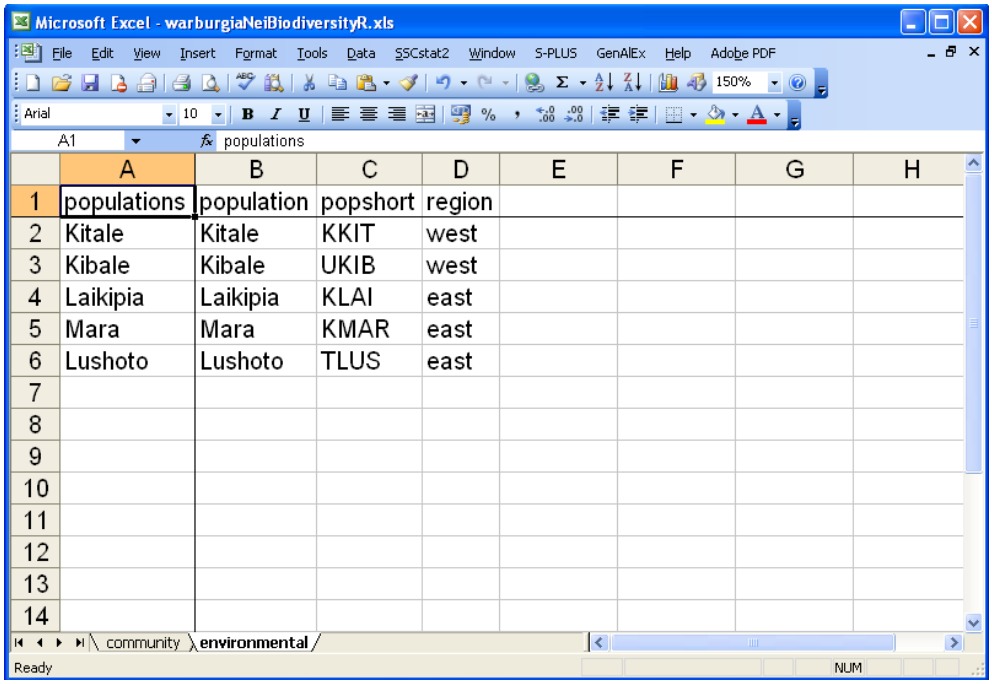
Figure II.7. Format for the "environmental" worksheet that contains 'geographical' data for input into BiodiversityR.

If importing distance matrices (pairwise differences between populations or individuals) for analysis in BiodiversityR, the format for "community" and "environmental" worksheets would be as shown in Figs. II.8 and II.9, respectively. Note that a square rather than a diagonal distance matrix is used in the "community" worksheet. Also note the same order for populations in both worksheets.



	A	B	C	D	E	F	G
1	populations	Kitale	Kibale	Laikipia	Mara	Lushoto	
2	Kitale	0.0000	0.0428	0.0920	0.1002	0.1071	
3	Kibale	0.0428	0.0000	0.0539	0.0676	0.0594	
4	Laikipia	0.0920	0.0539	0.0000	0.0062	0.0104	
5	Mara	0.1002	0.0676	0.0062	0.0000	0.0060	
6	Lushoto	0.1071	0.0594	0.0104	0.0060	0.0000	
7							
8							
9							
10							
11							
12							
13							
14							

Figure II.8. Format for the "community" worksheet that contains a genetic distance matrix for input into BiodiversityR.



	A	B	C	D	E	F	G	H
1	populations	population	popshort	region				
2	Kitale	Kitale	KKIT	west				
3	Kibale	Kibale	UKIB	west				
4	Laikipia	Laikipia	KLAI	east				
5	Mara	Mara	KMAR	east				
6	Lushoto	Lushoto	TLUS	east				
7								
8								
9								
10								
11								
12								
13								
14								

Figure II.9. Format for the "environmental" worksheet that contains 'geographical' data on populations for input into BiodiversityR.

II.5. AFLP-SURV (version 1.0)

II.5.1. Installation

AFLP-SURV can be downloaded from: http://www.ulb.ac.be/sciences/lagev/aflp-surv.html

Download the files named "AFLPsurv.exe" and "manual_AFLPsurv.pdf".

The software does not require further installation, as all features are available by clicking on "AFLPsurv.exe".

The easiest way to use the software is to copy input datasets into the same folder as where the "AFLPsurv.exe" file resides.

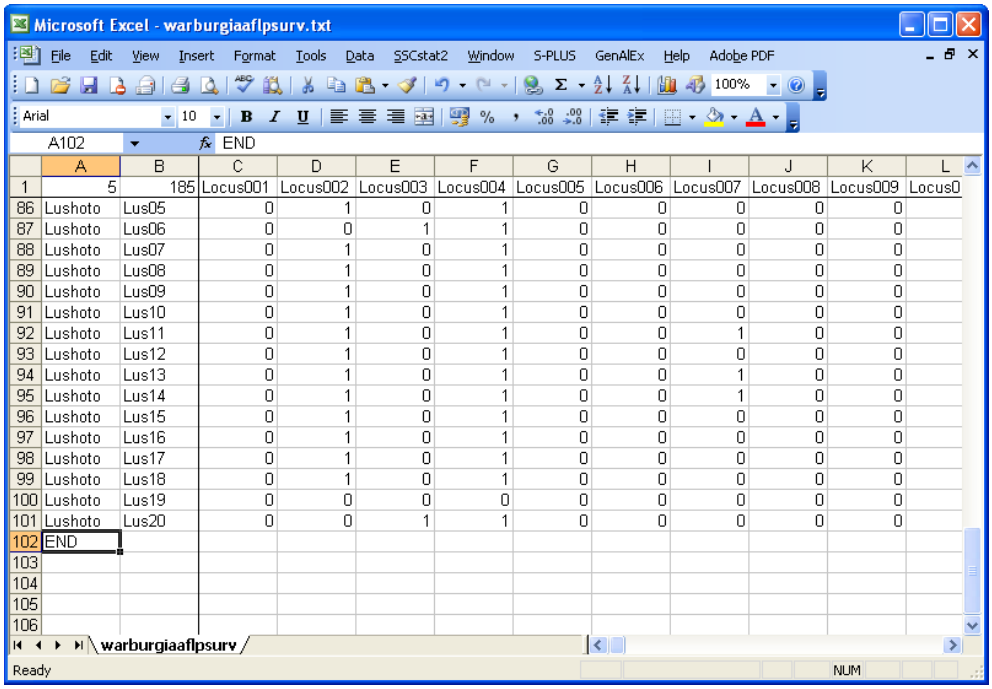
The suggested citation for the software is: Vekemans X (2002) AFLP-SURV version 1.0. Laboratoire de Génétique et Ecologie Végétale, Université Libre de Bruxelles, Belgium. URL http://www.ulb.ac.be/sciences/lagev/aflp-surv.html

II.5.2. Preparing input data

Prepare input files via GenAlEx (section II.2) and give the export file an extension of ".txt" (e.g., "*warburgiaaflpsurvdata.txt*").

Fig. II.10 provides the format of the input file.

The *Warburgia* spreadsheet does not contain missing data. If this were the case, then a value of "9" should be given for missing data points.



	A	B	C	D	E	F	G	H	I	J	K	L
1	5	185	Locus001	Locus002	Locus003	Locus004	Locus005	Locus006	Locus007	Locus008	Locus009	Locus0
86	Lushoto	Lus05	0	1	0	1	0	0	0	0	0	
87	Lushoto	Lus06	0	0	1	1	0	0	0	0	0	
88	Lushoto	Lus07	0	1	0	1	0	0	0	0	0	
89	Lushoto	Lus08	0	1	0	1	0	0	0	0	0	
90	Lushoto	Lus09	0	1	0	1	0	0	0	0	0	
91	Lushoto	Lus10	0	1	0	1	0	0	0	0	0	
92	Lushoto	Lus11	0	1	0	1	0	0	1	0	0	
93	Lushoto	Lus12	0	1	0	1	0	0	0	0	0	
94	Lushoto	Lus13	0	1	0	1	0	0	1	0	0	
95	Lushoto	Lus14	0	1	0	1	0	0	1	0	0	
96	Lushoto	Lus15	0	1	0	1	0	0	0	0	0	
97	Lushoto	Lus16	0	1	0	1	0	0	0	0	0	
98	Lushoto	Lus17	0	1	0	1	0	0	0	0	0	
99	Lushoto	Lus18	0	1	0	1	0	0	0	0	0	
100	Lushoto	Lus19	0	0	0	0	0	0	0	0	0	
101	Lushoto	Lus20	0	0	1	1	0	0	0	0	0	
102	END											
103												
104												
105												
106												

Figure II.10. Format for the input file for AFLP-SURV, showing the required header and footer. Cell A1 contains the number of populations (5), cell B1 the number of loci (185) and cell A102 indicates the end of the data ("END").

II.6. TFPGA (Tools for population genetic analyses) (version 1.3)

II.6.1. Installation

TFPGA can be downloaded from: http://www.marksgeneticsoftware.net/tfpga.htm

Download the zip archive "TFPGAPRG.ZIP" and extract all files to the same folder (e.g., C:\Program Files > TFPGA).

Click on the file "Setup.exe" and follow the given instructions.

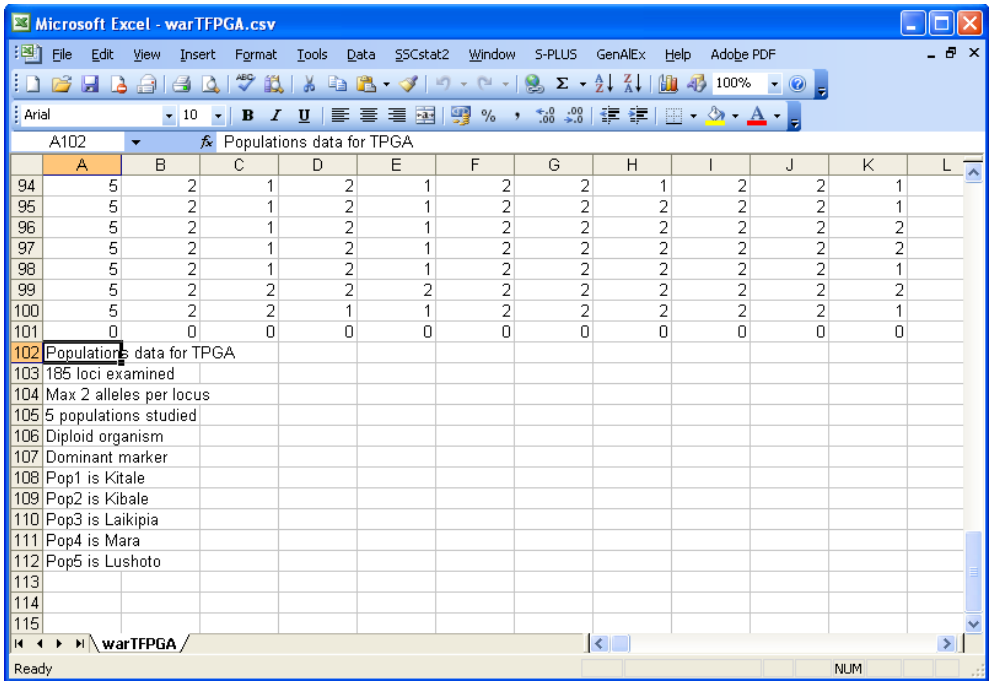
The suggested citation for the software is: Miller MP (1997) Tools for population genetic analyses (TFPGA) 1.3. A Windows program for the analysis of allozyme and molecular population genetic data. URL http://www.marksgeneticsoftware.net/tfpga.htm

II.6.2. Preparing input data

Prepare input files via GenAlEx (section II.2) and give the export file an extension of ".csv". Limit the file name to a maximum of eight more characters (the maximum that TFPGA can read (e.g., "*warTFPGA.csv*").

Data need to be entered into TFPGA in two sequential steps. The second step requires that the user provide information such as the number of loci being tested, the number of alleles per locus (2 for dominant data) and the number of populations under study. A 'reminder' of these parameters can be entered below the data constituting the 'initial' input, as shown in Fig. II.11.

The *Warburgia* spreadsheet does not contain missing data. If this were the case, then a value of "0" should be given for missing data points.



	A	B	C	D	E	F	G	H	I	J	K
94	5	2	1	2	1	2	2	1	2	2	1
95	5	2	1	2	1	2	2	2	2	2	1
96	5	2	1	2	1	2	2	2	2	2	2
97	5	2	1	2	1	2	2	2	2	2	2
98	5	2	1	2	1	2	2	2	2	2	1
99	5	2	2	2	2	2	2	2	2	2	2
100	5	2	2	1	1	2	2	2	2	2	1
101	0	0	0	0	0	0	0	0	0	0	0
102	Populations data for TPGA										
103	185 loci examined										
104	Max 2 alleles per locus										
105	5 populations studied										
106	Diploid organism										
107	Dominant marker										
108	Pop1 is Kitale										
109	Pop2 is Kibale										
110	Pop3 is Laikipia										
111	Pop4 is Mara										
112	Pop5 is Lushoto										
113											
114											
115											

Figure II.11. Input file for TFPGA. Product presence is coded as “1”, Product absence as “2”. Populations are coded as “1”, “2”, “3”, “4” or “5” (column A). A row of zeroes (row 101) indicates the end of the data. Further information on data is provided below this row.

II.7. PopGene (version 1.32)

II.7.1. Installation

PopGene can be downloaded from: http://www.ualberta.ca/~fyeh/index.htm

Click on the file "pop32.exe"and follow the given instructions.

The suggested citation for the software is: Yeh FC, Yang R-C, Boyle T (1999) PopGene version 1.32. Microsoft Windows-based freeware for population genetic analysis. URL http://www.ualberta.ca/~fyeh/index.htm

II.7.2. Preparing input data

Prepare input files via GenAlEx (section II.2) and give the export file an extension of ".txt" (e.g., "*warburgiaPopgene.txt*"). The input file for PopGene will have a format as shown in Fig. II.12.

If known, information on the level of Hardy-Weinberg disequilibrium (the inbreeding coefficient, or F_{IS} value) can be provided in the data input file immediately below the name of a population. This information needs to be inserted manually after preparing the input file through GenAlEx.

The PopGene manual gives an example of a diploid dominant marker data set.

The *Warburgia* spreadsheet does not contain missing data. If this were the case, then a value of "." should be given for missing data points.

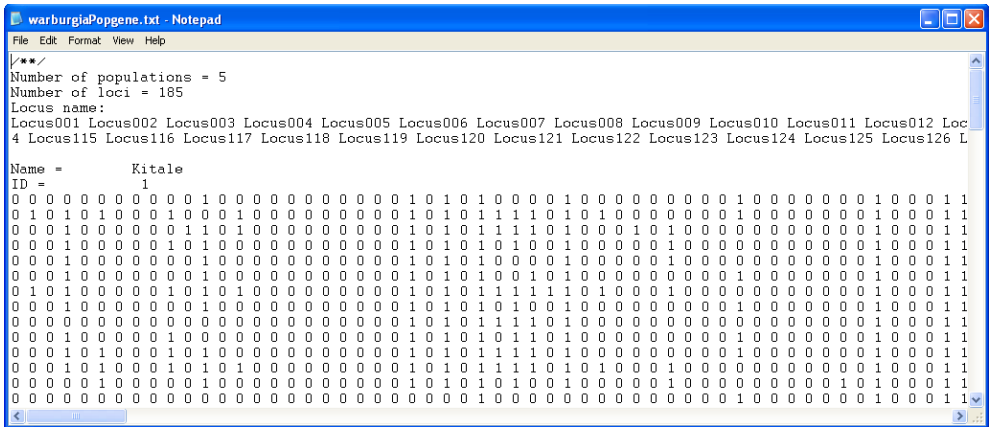


Figure II.12. Input file for PopGene.

II.8. Hickory (version 1.1)

II.8.1. Installation

Hickory can be downloaded from: http://darwin.eeb.uconn.edu/hickory/software.html

Click on the file "hickory-setup-v1.1.exe" and follow the given instructions.

The suggested citation for the software is: Holsinger KE, Lewis PO (2007) Hickory v1.1: a package for analysis of population genetic data.
URL http://darwin.eeb.uconn.edu/hickory/software.html

II.8.2. Preparing input data

We suggest that input files are prepared via FAMD (section II.3). Give the export file an extension of ".nex" (e.g., "*warburgiaHickory.nex*"). The input file for Hickory will have a format as shown in Figure II.13.

The *Warburgia* spreadsheet does not contain missing data. If this were the case, then any specified character such as "?" can be used for missing data points.

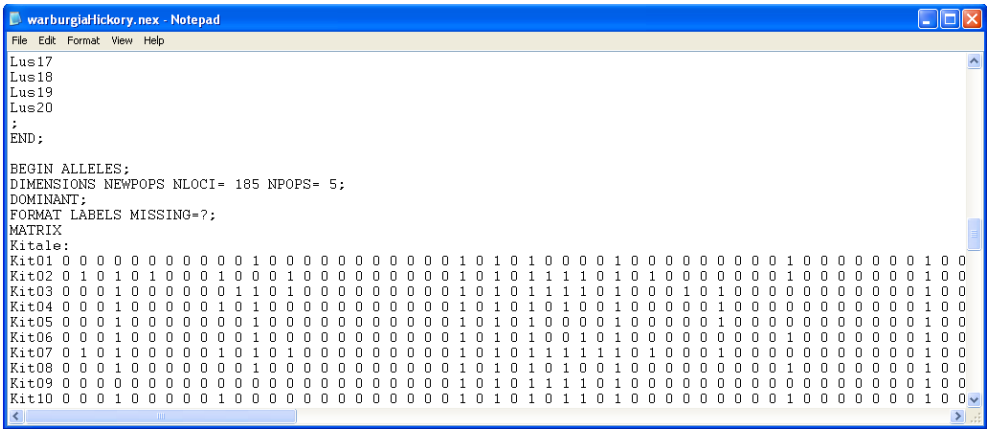


Figure II.13. Part of the input file for Hickory.

II.9. STRUCTURE (version 2.3.1)

II.9.1. Installation

STRUCTURE can be downloaded from: http://pritch.bsd.uchicago.edu/software.html

Click on the file "structure2.3.1.win.exe" and follow the given instructions.

If problems are encountered in opening the program (e.g., as can occur in Windows Vista), then start STRUCTURE from the "structure_start.bat" file.

The suggested citation for the software is: Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* 7: 574-578.

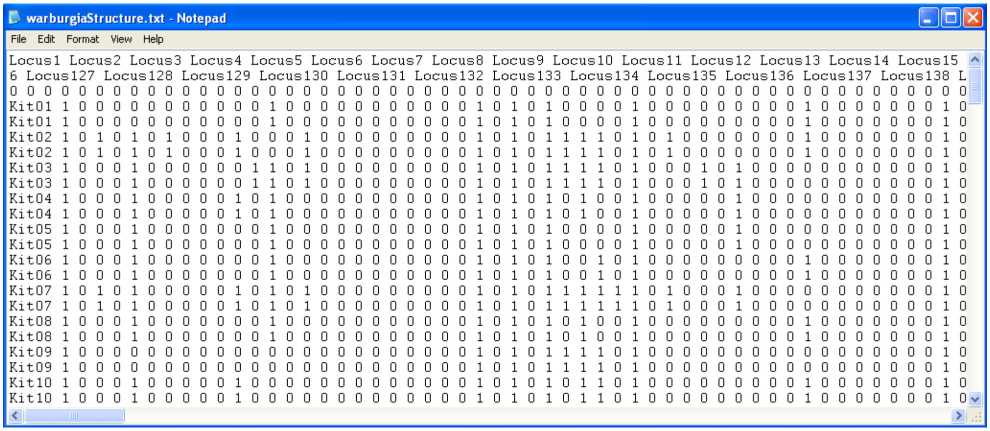
II.9.2. Preparing input data

We suggest that input files are prepared via FAMD (section II.3). Give the export file an extension of ".txt" (e.g., "*warburgiaStructure.txt*"). The input file for Structure will have a format as shown in Figure II.14.

The *Warburgia* spreadsheet does not contain missing data. If it did, then any specified integer such as "-9" can be used for missing data points.

The data set can be imported into STRUCTURE through the menu option:
File > New Project.

Then follow the steps given in Figs.II.15 and II.16.



```
warburgiaStructure.txt - Notepad
File Edit Format View Help
Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7 Locus8 Locus9 Locus10 Locus11 Locus12 Locus13 Locus14 Locus15
6 Locus127 Locus128 Locus129 Locus130 Locus131 Locus132 Locus133 Locus134 Locus135 Locus136 Locus137 Locus138 L
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Kit01 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit01 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit02 1 0 1 0 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit02 1 0 1 0 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit03 1 0 0 0 1 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit03 1 0 0 0 1 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit04 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit04 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit05 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit05 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit06 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit06 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit07 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit07 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit08 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit08 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit09 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit09 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
Kit10 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Kit10 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
```

Figure II.14. Part of the input file for Structure.

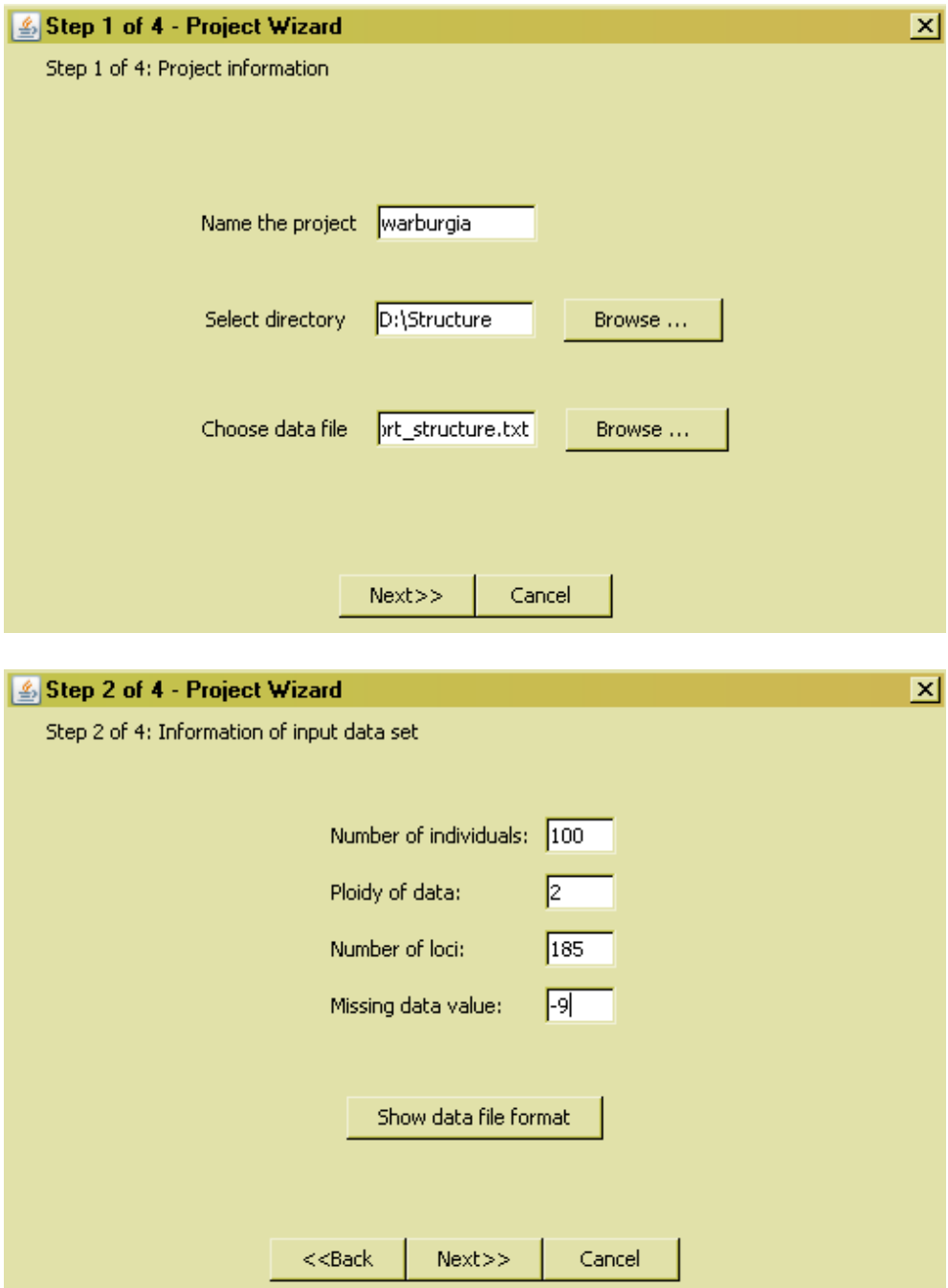


Figure II.15. Steps 1 and 2 to specify a new project in STRUCTURE.

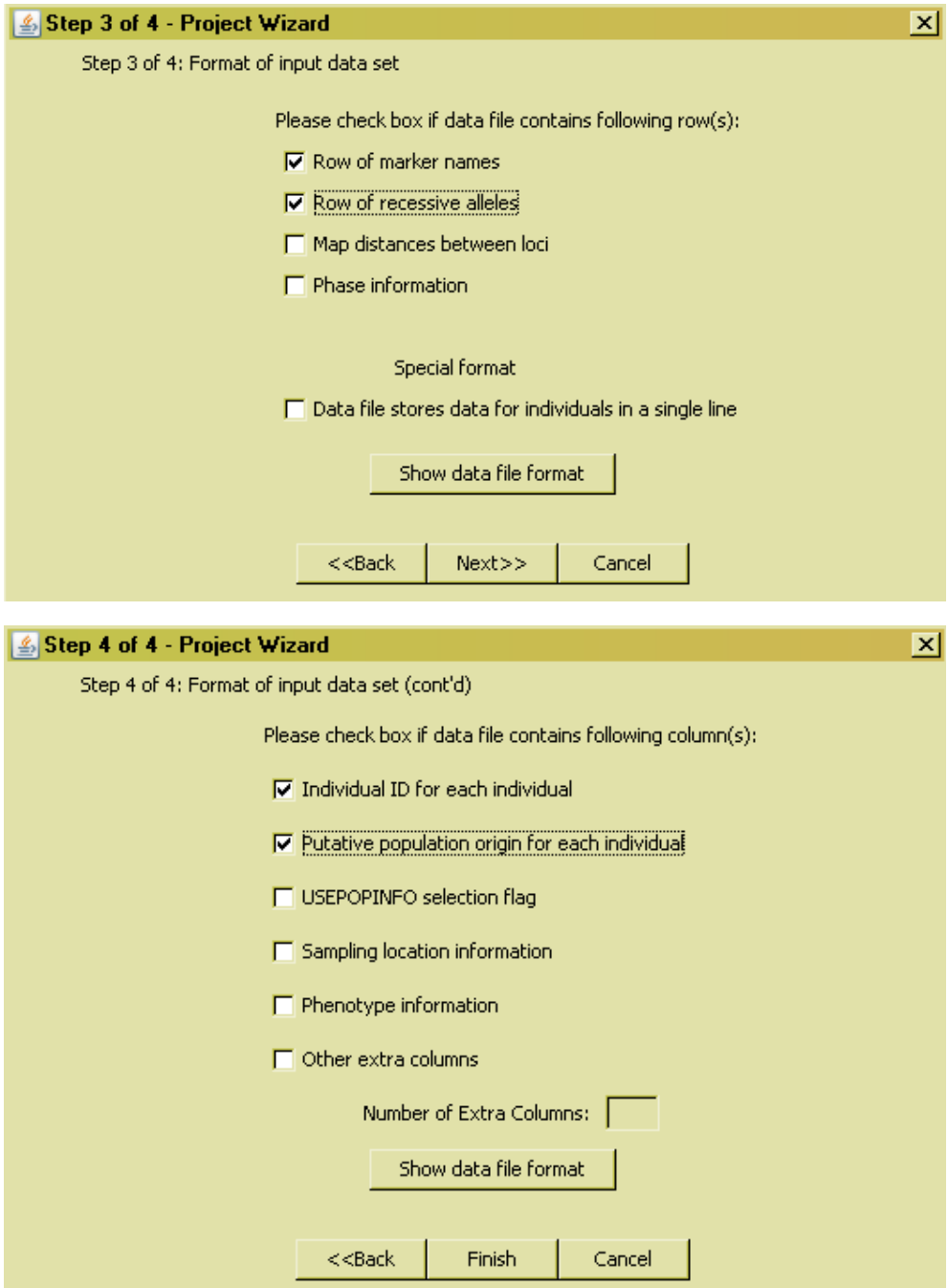


Figure II.16. Steps 3 and 4 to specify a new project in STRUCTURE.

After these steps, the screen depicted in Fig. II.17 will appear.

Save the project (e.g., as "*warburgia.spj*") by the menu option: File > Save Project.

After having imported the data for a project once, the project (data and any previous calculations) can be accessed from the menu option: File > Open Project...

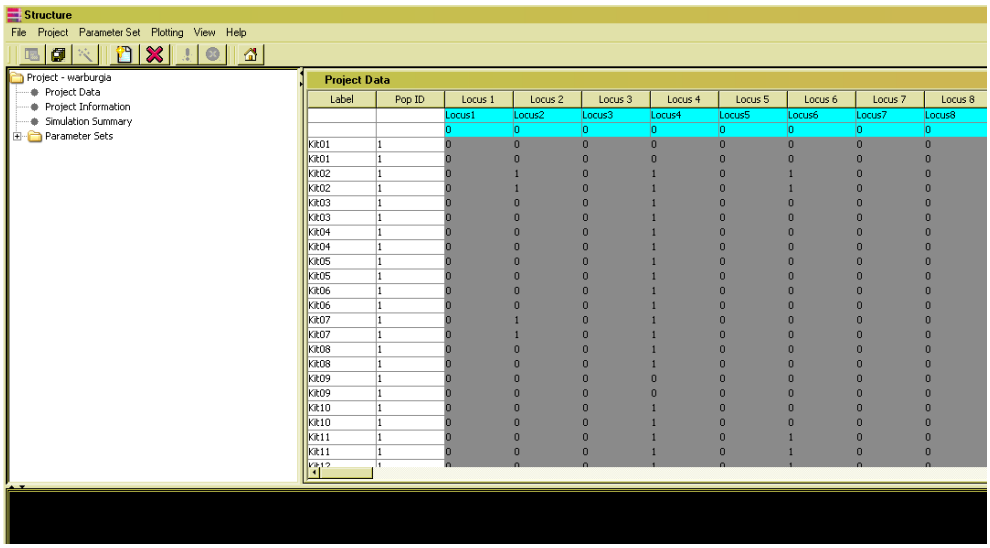


Figure II.17. Screen view after successfully importing data into STRUCTURE. Data are now ready for analysis.

II.10. PHYLIP (Phylogeny Inference Package)

(version 3.69)

II.10.1. Installation

PHYLIP consists of various programs that can be together downloaded from: http://evolution.genetics.washington.edu/phylip.html

Click on the zip archive "phylip-3.69.exe" and extract all files to the same folder (e.g., C:\Program Files\phylip3.69).

Programs are run by clicking on the appropriate icon in the 'exe' folder of PHYLIP (e.g., C:\Program Files \phylip3.69\exe\consense.exe).

The suggested citation for PHYLIP is: Felsenstein J (2009) Phylogeny Inference Package (PHYLIP). Version 3.69. URL http://evolution.genetics.washington.edu/phylip.html

II.10.2. Preparing input data

We are interested in using PHYLIP to analyse input files generated by AFLP-SURV (section II.5), which prepares appropriate input files automatically ("aflp_fst.txt", "aflp_nei.txt" and "aflp_reyn.txt").

These files can be analysed by the neighbor.exe program of PHYLIP.

FAMD (section II.3) also generates input files in a format that can be read by the neigbor.exe program. FAMD also generates tree files that can be analysed with the consense.exe program, or that can be plotted with the drawtree.exe and drawgram.exe programs.

Appendix III

Undertaking analysis in different software packages

Key points

We here give step-by-step instructions on how to analyse molecular data in particular software packages. Instructions are given following the same layout of chapters as in the main body of the guide. We give some recommendations on preferred packages but also include descriptions of methods as conducted in other software suites.

Please refer to Appendix II for detailed information on how to install the various software packages referred to here, and on how to format data for analysis in them.

The CD-ROM accompanying this guide provides a collection of input data spreadsheets formatted for different software packages and types of analysis. Also included are the corresponding results files produced by different programs.

The test analyses referred to in this guide are based on an example AFLP data set from the African medicinal tree *Warburgia ugandensis*.

Chapter 2. Measuring diversity

2.1. Analysis in AFLP-SURV 1.0

We recommend the use of AFLP-SURV as it provides for Bayesian estimates of allele frequencies, as well as providing 'classical' frequency calculations such as the 'square-root' method.

Input data needs to be in the same directory as the software. The software has no menu interface, so the user needs to type in the following after launching the program:

Input file: warburgiaaflpdata <RETURN>
Output file: <RETURN>
Subset file: <RETURN>
Choose a method of computation of allelic frequencies: 4 <RETURN>
If you assume Hardy-Weinberg genotypic proportions: <RETURN>
Enter the number of permutations for test on Fst: <RETURN>
Enter the number of bootstraps for genetic distances: <RETURN>
Press Return to close the window: <RETURN>

Various text files will have been generated in the folder where the program resides. Give these files different names (for example, "Warburgia aflpout Bayesian.txt") so that future runs of the program do not overwrite these files.

The program gives unbiased estimates for diversity based on the formulae developed by Lynch and Milligan (1994).

The AFLP-SURV software allows the inbreeding coefficient to be specified, making more advanced analyses possible.

2.2. Analysis in FAMD 1.23 beta

We recommend the use of FAMD as it provides Bayesian estimates of allele frequencies, as well as providing 'classical' estimates such as based on the 'square-root' method. FAMD also allows for different approaches to handling missing data.

First import the data by following the menu option: File > Load

State that individuals are provided in rows, header presence is available for individuals, it is delimited data and that groups should be included.

Use the following menu option to estimate allele frequencies using various methods: Analysis > Null Allele Frequencies

The software may request whether data should be appended to a previously generated results file or whether all previous results should be deleted first.

Results can be accessed via the menu option: View > Analysis File

Alternatively, open the analysis file ("analysis.txt") that was generated in the same directory as where the input file was located.

FAMD has three Bayesian estimation methods: uniform priors, non-uniform priors based on differences in frequencies among populations, and non-uniform priors based on differences in frequencies among loci.

2.3. Analysis in GenAlEx 6.2

GenAlEx does not apply Bayesian approaches to allele frequency estimation and uses the 'square-root' method only, with Hardy-Weinberg assumptions. For formal analysis we therefore prefer the use of AFLP-SURV or FAMD.

GenAlEx provides estimates of standard and unbiased *H*. These are referred to in the package as expected heterozygosity, He (and mean He), and unbiased expected heterozygosity, UHe (and mean UHe), respectively. GenAlEx also calculates the percentage of polymorphic loci in populations (referred to as %P).

Open the Excel worksheet that contains the data.

To estimate diversity for each individual locus and across loci in populations, use the following menu options:

GenAlEx > Frequency…
→ Data Format: Binary (Diploid) <OK>
→ Frequency & Heterozygosity by Pop: Yes (ticked)
→ Frequency & Heterozygosity by Locus: Yes (ticked)
→ Allelic Patterns: Yes (ticked) <OK>

Results are presented in Excel worksheets "AFP", "AFL" and "APT".

2.4. Analysis in PopGene 1.32

PopGene does not apply Bayesian approaches to allele frequency estimation and uses the 'square-root' method only. It does however allow the inbreeding coefficient to be specified for each population.

After launching the program, use the following menu option to import data:
File > Load Data > Dominant Marker Data

To estimate diversity at the locus level, use the following menu options:

Dominant > Diploid data…
→ Data Format: Variable as column
→ Hierarchical structure: Single populations
→ Estimation: gene frequency, allele number, effective allele number, polymorphic loci, gene diversity, Shannon index <OK>
→ Do you want to retain all loci for further analysis? <Yes>
→ Do you want to retain all populations for further analysis? <Yes>

2.5. Analysis in TFPGA 1.3

TFPGA does not apply Bayesian approaches to allele frequency estimation. We therefore prefer the use of AFLP-SURV or FAMD.

First import data by following the menu option: File > Open Data File

Next describe the data set. First scroll to the bottom of the input data. Choose the menu option of:

Describe Data > Populations
→ 185 # loci examined
→ 2 Max. # of alleles at a locus
→ 5 # of populations studied
→ Organism type: Diploid (ticked)
→ Marker type: Dominant
→ Diploid/Dominant Options: Square Root of the frequency of recessive genotype
<OK> <OK>

Follow these menu options to estimate diversity at the locus level:

Analyze > Descriptive statistics > Options
→ Calculate statistics for: Populations
→ Options: Calculate Allele and Heterozygote Frequencies, Calculate Per Locus Heterozygosities <OK>
Analyze > Descriptive statistics > Start Analysis
→ The results file from your last analysis is too big: <Rename>
(e.g., "Warresult.txt"; limit the name of the file to 8 characters)

Box III.1. Comparison of diversity values in BiodiversityR

In Chapter 2 we mentioned that statistical significance could be ascribed to differences in diversity estimates between populations. This box explains how this can be done using BiodiversityR and the paired t-test, using the Rcmdr ("R-commander") graphical user interface (GUI) of the package. We illustrate the approach using two of our Warburgia *example populations, Laikipia and Lushoto (the most and least diverse populations of our data set, respectively).*

Data need to be prepared in a different format from normal, whereby rows correspond to information from particular loci and columns correspond to the diversity of different populations (e.g., Nei's diversity measure).

Use the following menu options to import the data:

Data > Import data > From Excel, Access or dBase data set…
→ Enter name of data set: paireddata <click OK button>
→ Select the "Multiple comparisons Rcmdr data.xls" dataset <OK>
→ Select the "input data" data <OK>

Use the following menu options for multiple comparisons:

Statistics > Means > Paired t-Test…
→ First variable: Laikipia.UH
→ Second variable: Lushoto.UH
→ Alternative Hypothesis: Two-sided (click to select)
→ Confidence level: 0.995 <OK>

When we investigated for the difference in unbiased H between Laikipia and Lushoto, the following results were obtained:

Box III.1. continued

```
Paired t-test

data:  paireddata$Laikipia.UH and paireddata$Lushoto.UH
t = 4.1965, df = 184, p-value = 4.214e-05
alternative hypothesis: true difference in means is not equal to 0
99.5 percent confidence interval:
 0.01151555 0.05980176
sample estimates:
mean of the differences
             0.03565866
```

Since the 99.5% confidence interval does not span zero but has the limits of 0.0115 and 0.0598, we have a statistical basis for a true difference between Laikipia and Lushoto. Note that a 99.5% rather than the more common 95% confidence interval is used as 10 comparisons can theoretically be made between the five populations, so the "type-I error" of 5% is divided by 10. This is called the Bonferroni correction method and is used to choose the final desired significance level in multiple comparisons.

The same analysis of diversity for other pairs of stands (which are closer in diversity) result in ranges that span zero, meaning that there is no support statistically for estimates being different.

Chapter 3. Measuring genetic distance between populations

3.1. Analysis in AFLP-SURV 1.0

We recommend the use of AFLP-SURV as it can calculate Nei's genetic distance based on Bayesian and 'classical' estimates of allele frequencies (Bayesian estimates are better, see Chapter 2).

Input data needs to be in the same directory as the software. The software has no menu interface, so the user needs to type in the following after launching the program:

Input file: warburgiaaflpdata <RETURN>
Output file: <RETURN>
Subset file: <RETURN>
Choose a method of computation of allelic frequencies: 4 <RETURN>
If you assume Hardy-Weinberg genotypic proportions: <RETURN>
Enter the number of permutations for test on Fst: <RETURN>
Enter the number of bootstraps for genetic distances: <RETURN>
Press Return to close the window: <RETURN>

Various text files will have been generated in the folder where the program resides. Give these files different names so that future runs of the program do not overwrite these files.

The program provides the Nei genetic distance between populations based on the formulae developed by Lynch and Milligan (1994).

3.2. Analysis in FAMD 1.23 beta

Although FAMD provides Bayesian estimates of allele frequencies it does not calculate matrices for the commonly used Nei genetic distance (it uses chord distance).

First import data by the menu option: File > Load

State that individuals are provided in rows, header presence is available for individuals, it is delimited data and that groups should be included.

Use the following menu option to estimate the chord distance between populations: Analysis > Population Distance

The program will ask you to specify the method of estimating allele frequencies. The software will request whether data needs to be appended to a previously generated results file or whether all previous results should be deleted first.

Results can be accessed via menu option: View > Analysis File

Alternatively, open the analysis file ("analysis.txt") that was generated in the same directory as where the input file is located.

Another alternative for viewing results is available via menu option: File > Export > Phylip Distance Matrix

After calculating the distance matrix, the data needs to be imported again by the menu option: File > Load or via menu option: DataMatrix > Restore Original Matrix

3.3. Analysis in GenAlEx 6.2

GenAlEx does not apply Bayesian approaches to allele frequency estimation and uses the 'square-root' method only, with Hardy-Weinberg assumptions. It does not therefore provide the best estimates for genetic distances. For formal analysis we therefore prefer the use of AFLP-SURV.

GenAlEx provides estimates of standard and unbiased Nei genetic distances (referred to in the package as Nei distance and Nei unbiased distance, respectively).

Open the Excel worksheet that contains the data.

To calculate genetic distance matrices between populations using Nei's measures, use the following menu options:

GenAlEx > Frequency…
→ Data Format: Binary (Diploid) <OK>
→ Multiple pop options: Nei distance, Nei unbiased distance <OK>

Results are presented in Excel worksheets "NeiP" (for standard distances) and "UNeiP" (for unbiased distances).

3.4. Analysis in PopGene 1.32

PopGene does not apply Bayesian approaches to allele frequency estimation and uses the 'square-root' method only (although it does allow the inbreeding coefficient to be specified for each population). It therefore does not provide the best estimates for genetic distances and for formal analysis we prefer the use of AFLP-SURV.

After launching the program, use the following menu option to import data:
File > Load Data > Dominant Marker Data

To calculate the biased and unbiased standard Nei distances, use the following options:

Dominant > Diploid data…
→ Data format: Variable as column
→ Assumption: Hardy-Weinberg equilibrium
→ Hierarchical structure: Multiple populations
→ Estimation: Genetic distance <OK>
→ Do you want to retain all loci for further analysis? <Yes>
→ Do you want to retain all populations for further analysis? <Yes>

3.5. Analysis in TFPGA 1.3

TFPGA does not apply Bayesian approaches to allele frequency estimation and therefore does not provide the best estimates of genetic distance. The software does however provide a wide range of distance measures, including the standard Nei distance (with unbiased version), the minimum Nei distance (with unbiased version), and the Reynolds and Rogers (standard and modified versions) distances.

First import data by following the menu option: File > Open Data File

Next describe the data set. First scroll to the bottom of the input data. Choose the menu option of:

Describe Data > Populations
→ 185 # loci examined
→ 2 Max. # of alleles at a locus
→ 5 # of populations studied
→ Organism type: Diploid (ticked)
→ Marker type: Dominant
→ Diploid/Dominant Options: Square Root of the frequency of recessive genotype
<OK> <OK>

The following menu options estimate genetic distances:

Analyze > Genetic distance > Options
→ Select distance measure: Nei (1972, 1978)
→ Show distances in matrix format: Yes (ticked) <OK>
Analyze > Genetic distance > Start Analysis

Box III.2. Comparison of distance matrices in BiodiversityR

As we explained in Chapter 3 (Box 3.1), methods are available to graphically compare the genetic distances provided by different measures to see how closely they correspond. Here we describe how the BiodiversityR package can be used for this purpose. For the analysis shown below, Nei distances were obtained from AFLP-SURV and coancestry distances from TFPGA.

First import the data set of Nei distances through the following menu option:

BiodiversityR > Community dataset > Import datasets from Excel…
→ Enter name for community data set: NeiDistance
→ Enter name for environmental data set: PopEnv
→ Enter name for variable with sites: populations <OK>
→ Select the "warburgiaNeiBiodiversityR.xls" dataset

Next import the data set of coancestry distances through the following menu option:

BiodiversityR > Community dataset > Import datasets from Excel…
→ Enter name for community data set: CoancestryDistance
→ Enter name for environmental data set: PopEnv
→ Enter name for variable with sites: populations <OK>
→ Select the "warburgiaCoancestryBiodiversityR.xls" dataset

Plotting one distance matrix against another is not a standard option within the BiodiversityR graphical user interface. Therefore, copy and submit the following commands in the Rcmdr window (or, alternatively, in the R GUI window) to obtain a graph:

```
Nei <- as.dist(NeiDistance)
Coancestry <- as.dist(CoancestryDistance)
plot(Nei,Coancestry,xlab="standard Nei distance",ylab="Coancestry
      distance (Reynolds et al.)",cex=2)
```

Be careful not to change the case of any of the commands shown above, as BiodiversityR is case-sensitive.

Chapter 4. Visualising genetic distances by cluster analysis

4.1. Analysis in AFLP-SURV 1.0 with PHYLIP 3.69

The combination of AFLP-SURV with PHYLIP allows estimation of allele frequencies by Bayesian methods, provides distance matrices based on Nei's measure, and allows bootstrap analysis. We recommend use for bootstrap analysis, but suggest that BiodiversityR is otherwise better for carrying out clustering.

Input data needs to be in the same directory as the software. The software has no menu interface, so the user needs to type in the following after launching the program:

Input file: warburgiaaflpdata <RETURN>
Output file: <RETURN>
Subset file: <RETURN>
Choose a method of computation of allelic frequencies: 4 <RETURN>
If you assume Hardy-Weinberg genotypic proportions: <RETURN>
Enter the number of permutations for test on Fst: <RETURN>
Enter the number of bootstraps for genetic distances: 10000 <RETURN>
Press Return to close the window: <RETURN>

In the folder where the AFLP-SURV software resides, three files will have been created: "aflp_nei.txt", "aflp_fst.txt" and "aflp_reyn.txt". These files need to be copied into the *exe* folder of the PHYLIP package. It is also useful to change the names of files to indicate more specifically what they contain, since AFLP-SURV always uses the same names to export distance matrices. For example, use "aflpBayesian_Nei.txt" for results based on Bayesian estimations of allele frequencies.

From the *exe* folder of PHYLIP, click on the program *Neighbor.exe*. Next provide the following parameters:

- Please enter a new file name: aflpBayesian_nei.txt <ENTER>
- (Except if using the software for the first time) The file "outfile" that you wanted already exists: F <ENTER>
- (Except if using the software for the first time) Please enter a new file name: BayesianNei_result.txt <ENTER>
- N <ENTER> (to get UPGMA)
- M <ENTER> (to analyse multiple data sets)
- How many data sets: 10000 <ENTER>
- Random number seed: 5 <ENTER>
- J <ENTER> (to get the input order of species)
- Y <ENTER> (to accept settings)
- (Except if using the software for the first time) The file "outtree" already exists: F <ENTER>
- (Except if using the software for the first time) Please enter a new file name: NeighbourBayesianNei_tree.txt

The results needed for the next step will now be available in the file "BayesianNeiNeighbourtree.txt". Bootstrap analysis is completed by the *consense.exe* program. Click on this program and provide the following parameters:

- Please enter a new file name: BayesianNeiNeighbourtree.txt <ENTER>
- The file "outfile" already exists: F <ENTER>
- Please enter a new file name: ConsensusBayesianNei_result.txt <ENTER>
- R <ENTER> (To treat the trees as being rooted)
- Settings for this run:Y <ENTER>
- The file "outtree" already exists: F <ENTER>
- Please enter a new file name: ConsensusBayesianNei_tree.txt <ENTER>

Results can be viewed by opening the file "neicons.txt" in the *exe* folder of PHYLIP (see Box III.3)

Box III.3. Interpreting the output from the PHYLIP Consensus tree program

The Consensus tree program provides the following output:

```
Consensus tree program, version 3.69

Species in order:

  1. Kitale
  2. Kibale
  3. Laikipia
  4. Mara
  5. Lushoto


Sets included in the consensus tree

Set (species in order)     How many times out of 10000.00

..***                      10000.00
**...                      9178.00
...**                      5239.00


Sets NOT included in consensus tree:

Set (species in order)     How many times out of 10000.00

..**.                      4745.00
.****                      822.00
..*.*                      16.00
```

Box III.3. continued

The output first gives the identification number and names of the different populations (the software assumes that input data were for different species, but the fact that we investigate populations does not change the interpretation of results).

Next, information on the "consensus tree" is provided. This contains the clustering structure that was most frequently encountered for the 10,000 bootstrapped data sets. Information on the clusters ("sets") of the consensus tree are provided in a somewhat cryptic matrix format whereby the rows correspond to a specific clustering level and the columns to a population; a star at the position of the population indicates that the population was included in the cluster formed at that clustering level. The first cluster of the consensus tree included Laikipia ("species 3"), Mara ("species 4") and Lushoto ("species 5"). This cluster was formed for 10,000 out of 10,000 bootstrapped data sets.

The bootstrap results further show that we can be reasonably confident about the clustering of Kibale ("species 1") and Kitale ("species 2") as 9178 random data sets (91.8% of bootstraps) formed a cluster that included only these two populations. The results suggest that we should be less confident about the clustering of Mara ("species 4") and Lushoto ("species 5"), since in only 52.4% of random data sets was a cluster formed that included only these two populations.

Since Mara, Lushoto and Laikipia clustered together in each of the 10,000 bootstraps, there is no statistical evidence to conclude that Mara is more similar to Lushoto than to Laikipia (in 47.4% of random data sets, Mara clustered with Laikipia first).

4.2. Analysis in BiodiversityR

Except for the case of bootstrapping (see use of AFLP-SURV with PHYLIP), we recommend the use of BiodiversityR for clustering since it offers a wide range of algorithms and graphical outputs.

Genetic distances between populations need to be calculated in another software package such as AFLP-SURV and then imported into BiodiversityR.

Use the following menu options to import the distance matrix:

BiodiversityR > Community dataset > Import datasets from Excel…
→ Enter name for community data set: neidist
→ Enter name for environmental data set: warpop
→ Enter name for variable with sites: populations <OK>
→ Select the "WarburgiaNeiBiodiversityR.xls" dataset

Use the following menu options to calculate and plot agglomerative clustering results:

Biodiversity > Analysis of ecological distance > Clustering…
→ Cluster method: hclust or agnes
→ as.dist(Community): yes (ticked)
→ cophenetic correlation: yes (ticked; this is an advanced option)
→ Cluster options: average, single, complete or ward <OK>
→ Plot options: dendrogram1 <Plot>
→ Plot options: cophenetic <Plot> (this is an advanced option, see Box III.4)

Use the following menu options to calculate and plot divisive clustering results:

Biodiversity > Analysis of ecological distance > Clustering…
→ Cluster method: agnes
→ as.dist(Community): yes (ticked)
→ cophenetic correlation: yes (ticked; this is an advanced option) <OK>
→ Plot options: dendrogram1 <Plot>
→ Plot options: cophenetic (this is an advanced option) <Plot>

Box III.4. Interpreting cophenetic distances

Table 3.2 and Fig. 4.1 in Part 2 of this guide are both repeated below for the purposes of this exercise:

Table 3.2 Unbiased Nei distance (Lynch and Milligan 1994) between five populations of *Warburgia ugandensis*. These results were obtained with the AFLP-SURV package with the Bayesian estimation method with non-uniform priors to estimate allele frequencies.

Nei's unbiased genetic distance	Kitale	Kibale	Laikipia	Mara
Kibale	0.0428			
Laikipia	0.0920	0.0539		
Mara	0.1002	0.0676	0.0062	
Lushoto	0.1071	0.0594	0.0104	0.0060

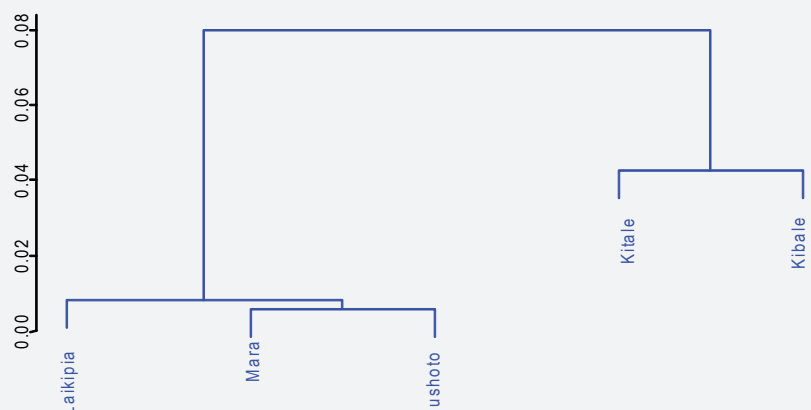


Figure 4.1. A phenogram showing cluster analysis of five *Warburgia* populations, based on the UPGMA clustering method and Nei's unbiased genetic distances (Table 3.2). The vertical axis shows the genetic distance at which populations cluster. The figure was created with the BiodiversityR package.

As we saw in Chapter 4, phenograms join clusters based on a summary statistic (such as the *average* for the UPGMA method). As a result, a cluster diagram only provides a summary of pairwise distances. For example, the phenogram of Fig. 4.1 suggests that the distance of Laikipia to the Mara or Lushoto populations is about

Box III.4. continued

0.01, but does not indicate what the actual distances between populations Laikipia and Mara, or Laikipia and Lushoto, are. The **cophenetic distances** between populations, which are defined as the pairwise distances that are suggested by the phenogram, are shown in Table III.1.

Table III.1 Cophenetic distances corresponding to the phenogram depicted in Fig. 4.1.

Cophenetic distance	Kitale	Kibale	Laikipia	Mara
Kibale	0.0428			
Laikipia	0.0800	0.0800		
Mara	0.0800	0.0800	0.0083	
Lushoto	0.0800	0.0800	0.0083	0.0060

It is now possible to plot the cophenetic distances against the original distances (the distances from the distance matrix) as in Fig. III.1.

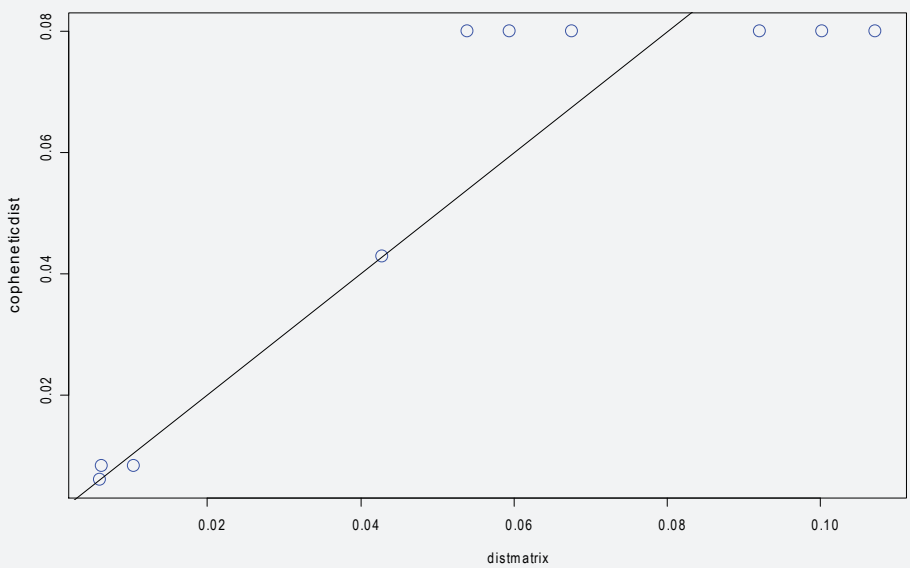


Figure III.1 Plot of pairwise genetic distances between populations (as shown in Table 3.2) on the horizontal axis against pairwise cophenetic distances between populations as suggested by the cluster algorithm on the vertical axis (as shown in Table III.1). The line shows where values on the horizontal axis are equal to values on the vertical axis (when genetic distance = cophenetic distance). The figure was created with the BiodiversityR package.

Box III.4. continued

Fig. III.1 is a graphical representation of what happens during the clustering process. Checking against the original distance matrix, we can infer that the lower left circle in the figure shows the merger of the Mara and Lushoto populations (at a distance of 0.0060). As the cluster was formed between these two populations only, the average is the same as the actual distance, and the circle is drawn on top of the identity line. The two circles directly above the first circle show the merger of the cluster of [Mara+Lushoto] with the Laikipia population, at a distance of 0.0083. The cluster analysis summarises the distances between Mara and Laikipia (0.0062) and Lushoto and Laikipia (0.0104) by the average of both these distances. The next circle we observe when scanning the diagram from bottom to top shows the merger of Kitale and Kibale at a distance of 0.0428, the actual distance between these two populations. The remaining six circles at the top of the diagram show the merger of the two clusters ([Mara, Laikipia and Lushoto] and [Kitale and Kibale]) at a cophenetic distance of 0.0800. This cophenetic distance now summarizes the 6 pairwise differences between Kitale and Laikipia (0.0920), Kitale and Mara (0.1002), Kitale and Lushoto (0.1071), Kibale and Laikipia (0.0539), Kibale and Mara (0.0676) and Kibale and Lushoto (0.0594).

The more circles on a horizontal belt in a graph such as Fig. III.1, the greater the number of pairwise distances that were summarised in the corresponding cophenetic distance matrix. The shorter the horizontal belts are, the closer are the actual genetic distances between populations to the cophenetic distances, and therefore the better the summary the cluster diagram represents.

4.3. Analysis in FAMD 1.23 beta

Although FAMD provides Bayesian estimates of allele frequencies it does not calculate distance matrices for the commonly used Nei measure and so we do not favour its use for clustering. It does however allow bootstrapping.

First import data by the menu option: File > Load

State that individuals are provided in rows, header presence is available for individuals, it is delimited data and that groups should be included.

To undertake bootstrap analysis, first select the number of bootstraps (parameter BS) from the menu via: Options > Bootstrapping and Replicate

Next generate all bootstrapped cluster results via the menu option: Replicate Analyses > Bootstrap Population Tree

Finally, analyse consensus among bootstrapped trees via the menu option:
Trees > Majority Rule Consensus or Trees > Strict Consensus

Bootstrap results can be accessed via the menu option: View > Analysis File

Alternatively, open the analysis file ("analysis.txt") that was generated in the same directory as where the input file is located.

An output file in the Newick standard (default: "constree.ph") will also be available in the same directory as the input file. This file can be read by PHYLIP to generate dendrograms with the drawgram or drawtree programs.

4.4. Analysis in TFPGA 1.3

TFPGA does not apply Bayesian approaches to allele frequency estimation and therefore does not provide the best estimates of genetic distance upon which cluster analysis is then based. The software does however allow for clustering for a wide range of distance measures.

First import data by following the menu option: File > Open Data File

Next describe the data set. First scroll to the bottom of the input data. Choose the menu option of:

Describe Data > Populations
→ 185 # loci examined
→ 2 Max. # of alleles at a locus
→ 5 # of populations studied
→ Organism type: Diploid (ticked)
→ Marker type: Dominant
→ Diploid/Dominant Options: Square Root of the frequency of recessive genotype
<OK> <OK>

The following menu options provide for average linkage clustering between populations and carry out a bootstrap analysis:

Analyze > UPGMA > Options
→ Base tree on: Populations
→ Distance measure: Nei (1978)
→ Bootstrap over loci: Yes (ticked)
→ # of permutations: 10000 <OK>
Analyze > Descriptive statistics > Start Analysis

4.5. Analysis in PopGene 1.32

PopGene does not apply Bayesian approaches to allele frequency estimation and uses the 'square-root' method only (although it does allow the inbreeding coefficient to be specified for each population). It therefore does not provide the best estimates for the genetic distances that are used for clustering. In addition, graphical outputs do not provide clustering distances.

After launching the program, use the following menu option to import data:
File > Load Data > Dominant Marker Data

The following menu options provide average linkage results for standard and unbiased Nei's genetic distances (referred to in the package as Nei's original measure and Nei's unbiased measure, respectively):

Dominant > Diploid data…
→ Data Format: Variable as column
→ Assumption: Hardy-Weinberg equilibrium
→ Hierarchical structure: Multiple populations
→ Estimation: dendrogram <OK>
→ Do you want to retain all loci for further analysis? <Yes>
→ Do you want to retain all populations for further analysis? <Yes>

Output files are generated in the same folder that contained input data. They can be inserted directly into word processors packages such as Microsoft Word.

Chapter 5. Visualising genetic distances by ordination

5.1. Analysis in BiodiversityR

We recommend use of the BiodiversityR package as it allows cluster information to be superimposed onto ordination diagrams and has other features that provide added value in ordination analysis.

Genetic distances between populations need to be calculated first by another package.

Use the following menu options to import the distance matrix:

BiodiversityR > Community dataset > Import datasets from Excel…
→ Enter name for community data set: warneidist
→ Enter name for environmental data set: warpop
→ Enter name for variable with sites: populations <OK>
→ Select the "WarburgiaNeiBiodiversityR.xls" dataset

Use the following options to calculate and plot principal coordinate analysis results:

Biodiversity > Analysis of ecological distance > Unconstrained ordination…
→ Ordination method: PCoA
→ PCoA/NMS axes: 4
→ as.dist(Community): yes (ticked) <OK>
→ Plot method: ordiplot empty <Plot>
→ Plot method: points sites <Plot>
→ cex: 2 (this is a way of changing the size of plotting)
→ Plot method: identify sites <Plot; click next to symbols>
→ Plot method: origin axes <Plot>
→ Plot options: ordicluster <Plot>
→ Plot options: distance displayed <Plot>

5.2. Analysis in GenAlEx 6.2

GenAlEx does not apply Bayesian approaches to allele frequency estimation. It therefore does not provide the best estimates for the genetic distances that are used for ordination. For formal analysis we therefore prefer the use of the more powerful BiodiversityR.

GenAlEx undertakes PCoA but not other methods of ordination. It can undertake analysis based on estimates of standard and unbiased Nei genetic distances (referred to in the package as Nei distance and Nei unbiased distance, respectively). In the example below, we use the unbiased distance option.

Open the Excel worksheet that contains the data.

To undertake PCoA, use the following menu options:

GenAlEx > Frequency…
→ Data format: One column/Locus: Binary (Diploid) <OK>
→ Multiple Pop options: Nei unbiased distance (ticked) <OK>

Continue from the worksheet named "UNeiP" with the following menu options:

GenAlEx > PCA…
→ Input Data Type: Tri Distance Matrix (ticked)
→ PCA Method: Distance - Not standardized <OK>

5.3. Analysis in FAMD 1.23 beta

Although FAMD provides Bayesian estimates of allele frequencies, it does not calculate matrices for the commonly used Nei genetic distance and so we do not favour its use for ordination.

First import data by the menu option: File > Load

Import the data without information on regional structure.

State that individuals are provided in rows, header presence is available for individuals, data is delimited and that groups should be included.

Use the following menu option to estimate the chord distance between populations: Analysis > Population Distance

The program will ask you to specify the method of estimating allele frequencies. The software will request whether data needs to be appended to a previously generated results file or whether all previous results should be deleted first.

After calculating the distance matrix, principal coordinate analysis results can be obtained via menu option: Trees > Principal Coordinates Analysis

Chapter 6. Measuring genetic distance between individuals

6.1. Analysis in BiodiversityR

We recommend use of BiodiversityR because it offers a wide range of genetic distance measures to compare pairs of individuals, and it provides a number of options for subsequent ordination analysis.

Using the following menu options, import the *Warburgia* data:

BiodiversityR > Community dataset > Import datasets from Excel…
→ Enter name for community data set: warcom
→ Enter name for environmental data set: warenv
→ Enter name for variable with sites: Individual <OK>
→ Select the "WarburgiaBiodiversityR.xls" dataset

Using the following menu options, calculate genetic distances between pairs of individuals:

→ Save data as: jaccard.matrix
→ Distance: Jaccard (or use other measures)
→ Make community dataset: no (not ticked) <OK>

If 'yes' had been selected for the last option above, then the distance matrix will also be interpreted as a 'community' data set. This then allows clustering or ordination from the distance matrix directly. However, it is not necessary to follow this approach to obtain results.

When selecting the genetic distance measure in BiodiversityR, users should be aware of the follow synonyms:

- Gower distance (sensu BiodiversityR) = simple mismatching distance
- Bray distance (sensu BiodiversityR) = Dice distance
- Manhattan distance (sensu BiodiversityR) = squared Euclidean distance

Plotting different pairwise distance measures against each other (as in Chapter 6, Fig. 6.1) is not an option in the BiodiversityR graphical user interface. To do so, copy and submit the following commands in the Rcmdr window (or, alternatively, in the R GUI window):

```
jaccard.matrix <- vegdist(warcom,method='jaccard', na.rm=T)
simplematch.matrix <- vegdist(warcom,method='gower', na.rm=T)
dice.matrix <- vegdist(warcom,method='bray', na.rm=T)
plot(simplematch.matrix,jaccard.matrix,xlab="simple matching
      distance",ylab="Jaccard distance",cex=2)
plot(simplematch.matrix,dice.matrix,xlab="simple matching
      distance",ylab="Dice distance",cex=2)
plot(jaccard.matrix,dice.matrix,xlab="Jaccard
      distance",ylab="Dice distance",cex=2)
```

Be careful not to change the case of any of the above commands, as BiodiversityR is case-sensitive.

6.2. Analysis in FAMD 1.23 beta

FAMD offers an even wider range of genetic distance measures than BiodiversityR and has several methods for dealing with missing data. However, the package offers fewer options for subsequent analysis of distance matrices.

First import data by the menu option: File > Load

State that individuals are provided in rows, header presence is available for individuals, data is delimited and that groups should be included.

Use the following menu option to select the genetic distance measure:
Options > (Dis)Similarity Coefficients

Next use the following menu option to obtain the distance matrix:
Analysis > Standard Similarity.

If your data set contains missing values, the option of Standard Similarity ignores them, while the option of Minimum Similarity estimates the theoretical maximum distance (different allelic states assumed for missing comparisons), the option of Maximum Similarity estimates the theoretical minimum distance (the same allelic states assumed for missing comparisons), and the option of Average Similarity is based on randomly replacing missing values.

Results can be accessed via the menu option: View > Analysis File

Alternatively, open the file "analysis.txt" that was generated in the same directory as the input file.

6.3. Analysis in GenAlEx 6.2

GenAlEx calculates squared Euclidean distances among individuals. For formal analysis we prefer to use other coefficients (e.g., simple mismatching, Dice, Jaccard) such as provided for by BiodiversityR or FAMD.

Open the Excel worksheet that contains the data.

To calculate a genetic distance matrix between individuals, use the following menu options:

GenAlEx > Distance > Genetic…
→ Distance Calculation: Binary (Diploid)
→ Label matrix: Sample <OK>

Chapter 7. Visualising genetic distances by ordination

7.1. Analysis in BiodiversityR

We recommend use of BiodiversityR because it offers a wide range of genetic distance measures to compare pairs of individuals and it provides for a large number of options in ordination.

Using the following menu options, import the *Warburgia* data:

BiodiversityR > Community dataset > Import datasets from Excel…
→ Enter name for community data set: warcom
→ Enter name for environmental data set: warenv
→ Enter name for variable with sites: Individual <OK>
→ Select the "WarburgiaBiodiversityR.xls" dataset

Using the following menu options, calculate principal coordinate analysis results:

BiodiversityR > Analysis of ecological distance > Unconstrained ordination…
→ Ordination method: PCoA (PCoA [Caillez] allows for an alternative analysis)
→ Distance: Jaccard (or another measure)
→ model summary: Yes (ticked) <OK>
→ Plot method: ordiplot empty <Plot>
→ Plot method: ordisymbol
→ Plot variable: Population [Factor] <Plot>
→ Plot method: identify sites <Plot; click next to symbols>
→ Plot method: ordispider <Plot>

Using the following menu options, undertake a constrained analysis of principal coordinates (this is an advanced analysis method that provides results that are similar to those provided by an analysis of molecular variance [AMOVA, Chapter 8]; these options will also generate results based on a non-parametric multivariate analysis of variance):

Biodiversity > Analysis of ecological distance > Constrained ordination…
→ Ordination method: capscale
→ Distance: Jaccard (or another measure)
→ model summary: Yes (ticked)
→ permutations: 999
→ Explanatory: Region + Population <OK> (choose by double-clicking)
→ Scaling: 2
→ Plot method: ordiplot empty <Plot>
→ Plot variable: population
→ Plot method: ordispider <Plot>

7.2. Analysis in FAMD 1.23 beta

FAMD offers an even wider range of genetic distance measures than BiodiversityR and has several methods for dealing with missing data. It is however more limited than BiodiversityR in ordination options.

First import data by the menu option: File > Load

State that individuals are provided in rows, header presence is available for individuals, data is delimited and that groups should be included.

Use the following menu option to select the genetic distance measure:
Options > (Dis)Similarity Coefficients

Next use the following menu option to obtain the distance matrix:
Analysis > Standard Similarity (see previous section [FAMD analysis for Chapter 6] for more information)

Use the following menu option to obtain the results of principal coordinate analysis:
Trees > Principal Coordinates Analysis

Results can be accessed via the menu option: View > Analysis File

Alternatively, open the file "analysis.txt" that was generated in the same directory as the input file.

7.3. Analysis in GenAlEx 6.2

Ordination in GenAlEx is based on Euclidean distances among individuals and for formal analysis we prefer to use other coefficients (e.g., simple mismatching, Dice, Jaccard) such as provided for by BiodiversityR or FAMD. GenAlEx undertakes PCoA but not other methods of ordination.

Open the Excel worksheet that contains the data.

To first calculate a genetic distance matrix between individuals, use the following menu options:

GenAlEx > Distance > Genetic…
→ Distance Calculation: Binary (Diploid)
→ Label matrix: Sample <OK>

Continue from the worksheet named "GD" (which represents squared Euclidean distances, although ordination is actually based on Euclidean distances) with the following menu options:

GenAlEx > PCA…
→ Input Data Type: Tri Distance Matrix (ticked)
→ PCA Method: Covariance – Not Standardized <OK>

Chapter 8. Analysis of molecular variance (AMOVA)

8.1. Analysis in GenAlEx 6.2

We recommend using GenAlEx for the nested AMOVA approach. We suggest that analysis be undertaken with (e.g., populations within regions within total) and without (e.g., populations within total) nesting and that both sets of results are reported. Analysis is based on squared Euclidean distances.

Open the Excel worksheet that contains the data.

To undertake AMOVA, use the following menu options:

GenAlEx > AMOVA…
→ Input data type: raw data <OK>
→ Distance Calculation: Binary (Diploid)
→ Output: none <OK>
→ Total dataset options: permutations = 9999
→ Pairwise population options: Output Pairwise PhiPT Matrix
→ Pairwise population options: permutations = 9999 <OK>

Results are presented in Excel worksheets "PhiPT" and "PhiPTP". Included is a visual representation of the percentage of variation among populations as a pie chart.

8.2. Analysis in FAMD 1.23 beta

FAMD provides AMOVA statistics for a wider range of genetic distance measures than GenAlEx but it does not provide tests for statistical significance.

First import data by the menu option: File > Load

State that individuals are provided in rows, header presence is available for individuals, data is delimited and that groups should be included.

To obtain nested AMOVA results, you need to first assign individuals to regional levels and subsequently assign individuals to population levels that are nested within the regional levels.

Use the following menu option: Analysis > AMOVA

Results can be accessed via the menu option: View > Analysis File

8.3. Analysis in Hickory 1.1

Hickory uses a Bayesian estimation method specifically developed for dominant markers to calculate F-*statistics. Among the different options that the software provides, we recommend the "f free model" and the "theta-II" parameter.*

First import data by the following menu option: Data > Load from file...

Select one of the following menu options to start analysis: Analyses > Run full model; Analyses > Run f=0 model; or Analyses > Run f free model

8.4. Analysis in PopGene 1.32

PopGene applies a version of F*-statistics ("*G*-statistics", see Appendix I) that assumes all populations were sampled rather than treating populations as random samples from a larger set of populations. This is not a realistic assumption.*

After launching the program, use the following menu option to import data:
File > Load Data > Dominant Marker Data

To estimate *G*-statistics, use the following menu options:

Dominant > Diploid data…
→ Data Format: Variable as column
→ Assumption: Hardy-Weinberg equilibrium
→ Hierarchical structure: Single populations, Multiple populations
→ Estimation: F-statistics <OK>
→ Do you want to retain all loci for further analysis? <Yes>
→ Do you want to retain all populations for further analysis? <Yes>

8.5. Analysis in AFLP-SURV 1.0

AFLP-SURV applies a version of F*-statistics ("*G*-statistics") that assumes all populations were sampled rather than treating populations as random samples from a larger set of populations. This is not a realistic assumption.*

Input data needs to be in the same directory as the software. The software has no menu interface, so the user needs to type in the following after launching the program:

Input file: warburgiaaflpdata <RETURN>
Output file: <RETURN>
Subset file: <RETURN>
Choose a method of computation of allelic frequencies: 4 <RETURN>
If you assume Hardy-Weinberg genotypic proportions: <RETURN>
Enter the number of permutations for test on Fst: 10000 <RETURN>
Enter the number of bootstraps for genetic distances: <RETURN>
Press Return to close the window: <RETURN>

Various text files will have been generated in the folder where the program resides. Give these files different names so that future runs of the program do not overwrite them.

8.6. Analysis in TFPGA 1.3

As in Hickory and GenAlEx, TFPGA calculates F*-statistics that treat populations as random samples from a larger set of populations. TFPGA does not base calculations on Bayesian estimates of allele frequencies as in Hickory.*

First import data by following the menu option: File > Open Data File

Next describe the data set. First scroll to the bottom of the input data. Choose the menu option of:

Describe Data > Populations
→ 185 # loci examined
→ 2 Max. # of alleles at a locus
→ 5 # of populations studied
→ Organism type: Diploid (ticked)
→ Marker type: Dominant
→ Diploid/Dominant Options: Square Root of the frequency of recessive genotype
<OK> <OK>

Follow these menu options to estimate F_{ST} (θ) values:

Analyze > F-statistics > Options
→ Show results for each locus: Yes (ticked)
→ Bootstrap over loci: 1000 reps, 95% Confidence Interval <OK>
Analyze > F-statistics > Start Analysis

Box III.5. CA comparison of different methods of obtaining *F*-statistics

Table III.2 provides results of *F*-statistics obtained by different approaches (see Chapter 8 and Appendix I). *F*-statistics should be interpreted as the ratio of diversity among populations / total diversity, whereby total diversity = diversity within populations + diversity among populations.

Table III.2 shows that the method of estimating *F*-statistics and whether or not nesting is considered sometimes influences the results obtained. We therefore recommend presenting these differences during formal reporting, especially if results relate to treating populations as random samples from a larger set of populations, as is the case for *Theta* and *Phi*.

Table III.2. Results for different methods of *F*-statistics provided by different software packages.

Software	*F*-statistic	Type of analysis	Result	Percentage diversity among populations
GenAlEx	Phi	nested	PhiPT = 0.4350	43.5%
		not nested	PhiPT = 0.3489	34.9%
TFPGA	Theta	Diploid, nested	thetaS = 0.3751	37.5%
		Diploid, not nested	theta = 0.2973	29.8%
		Haploid, nested	thetaS = 0.4350	43.5%
		Haploid, not nested	theta = 0.3489	34.9%
Hickory	Bayesian Theta	Not nested, free f	theta-II = 0.3317	33.2%
		Not nested, full model	theta-II = 0.3506	35.1%
		Not nested, f=0	theta-II = 0.2633	26.3%

Box III.5. continued

Although the algorithms for obtaining *F*-statistics via AMOVA (as in GenAlEx) or via Theta-statistics based on allele frequencies for haploid organisms (as in TFPGA) are different (see Appendix I), both these approaches result in exactly the same estimate of 43.5% of diversity among populations in nested analysis and 34.9% in non-nested analysis.

Allele frequencies for haploid organisms are the same as those in completely inbred diploid organisms, and AMOVA based on Euclidean distances therefore corresponds to treating individuals as completely inbred.

Chapter 9. STRUCTURE analysis

9.1. Analysis in Structure 2.3.2

The number of steps to use during 'burnin' and afterwards (MCMC steps) is set by the user: start with a relatively low number to get a feel for the analysis, as large numbers of steps can take a long time to run. In the below example, we set $K = 2$, but analysis should be run at a range of values and results compared (starting with $K = 2$ and increasing in steps of one).

After launching the program, import data using the menu option: File > Open Project...

Next, specify the steps to be used in analysis in the menu option:
Parameter Set > New...

Run length TAB:
Length of Burnin Period: 10000 (more steps will likely be needed for 'formal' analysis, perhaps up to 100,000)
Number of MCMC Reps after Burnin: 10000 <OK> (more steps will likely be needed for 'formal' analysis, perhaps up to 100,000)

(All other parameters under 'Parameter set' can be kept as default options)

Please name the new parameter set: 10000 10000 (this example of naming is based on the number of replications at different steps, but any unique identifier can be used to describe the parameter set) <OK>

Parameter set > Run
Set number of populations assumed (this is *K*): 2

Then click on the results file with the appropriate parameter set and value of *K*. To see bar plots of results (where different options are available for how to order individuals in graphical representations), use the following commands: Bar plot > Show

To view the LN P(D) value for a run, use the following menu option:
View > Simulation Summary

Technical Manual Series

TM 1: Linking Research to Extension for Watershed Management: The Nyando Experience.

TM 2: Bonnes pratiques de culture en pépinière forestière : Directives pratiques pour les pépinières communautaires.

TM 3: Bonnes pratiques de culture en pépinière forestière : Directives pratiques pour les pépinières de recherche.

TM 4:¡Plantemos madera! Manual sobre el establecimiento, manejo y aprovechamiento de plantaciones maderables para productores de la Amazonía peruana.

TM 5: Rainwater Harvesting Innovations in Response to Water Scarcity: The Lare Experience.

TM 6: Useful Trees and Shrubs of Ethiopia : Identification, Propagation and Management for 17 Agroclimatic Zones.

TM 7: Mapping the Potential of Rainwater Harvesting Technologies in Africa: A GIS Overview on Development Domains for the Continent and Nine Selected Countries.

TM 8: Green Water Management Handbook: Rainwater Harvesting for Agricultural Production and Ecological Sustainability.

TM 9: Molecular Markers for Tropical Trees, A Practical Guide to Principles and Procedures.

TM 10: La culture du jujubier: un manuel pour l'horticulteur sahélien.

TM 11: Carbon Guide for Smallholders.

TM 12: Semillas de Especies Arbóreas para los Agricultores.

TM 13: Molecular Markers for Tropical Trees: Statistical Analysis of Dominant Data.

In the last decade, there has been an enormous increase worldwide in the use of molecular marker methods to assess genetic variation in trees. These approaches can provide significant insights into the defining features of different taxa and this information may be used to define appropriate management strategies for species.

However, a survey of the literature indicates that the implementation of practical, more optimal management strategies based on results from molecular marker research is very limited to date for tropical trees. In order to explore why this is the case, the World Agroforestry Centre (ICRAF) undertook a survey of molecular laboratories in low-income countries in the tropics. This survey looked at the kinds of molecular marker studies that were being carried out on tree species, and the problems faced by scientists in this research.

One of the constraints that ICRAF's survey identified for the proper application of molecular markers is the effective handling and analysis of data sets once they have been generated. This guide has been designed to address this need for data obtained using dominant marker techniques. It has been created especially for students (MSc, PhD) and other researchers in developing countries that find themselves isolated from their peers and – when faced with an apparently bewildering array of options – find it difficult to settle on appropriate methods for analysis.

Most benefit will be obtained from this guide if it is used together with the companion volume on practical protocols for molecular methods (ICRAF Technical Manual no. 9), and so we recommend that scientists read both.

ISBN: 978-92-9059-262-4

FOREST & LANDSCAPE

FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

World Agroforestry Centre
TRANSFORMING LIVES AND LANDSCAPES

United Nations Avenue, Gigiri, P. O. Box 30677-00100, Nairobi, Kenya
Tel: (+254 20) 722 4000, Fax: (+254 20) 722 4001, Email: icraf@cgiar.org
www.worldagroforestry.org