

PowerMarker V3.0 Manual

URL: <http://www.powermarker.net>
E-mail: powermarker@hotmail.com

Jack Liu

TABLE OF CONTENTS

Chapter 1: Introduction	4
1.1 What is PowerMarker?	4
1.2 Data types handled by PowerMarker	4
1.3 Methods implemented in PowerMarker	4
1.4 System requirements	5
1.5 Installing and uninstalling PowerMarker	5
1.6 How to cite PowerMarker	5
1.7 Acknowledgements	5
1.8 Upgrade PowerMarker	5
Chapter 2: Tutorial	6
2.1: Creating a project	6
2.2: Importing a dataset	7
2.3: Choosing a subset from the dataset	9
2.4: Producing a table of summary statistics	9
Chapter 3: Data manipulation	11
3.1 Data import	11
3.1.1 Import dataset	11
3.1.2 Import frequency data	12
3.1.3 Import distance data	13
3.1.4 Import tree data	13
3.1.5 Import table data	13
3.1.6 Import text data	14
3.2 Dataset manipulation	14
3.2.1 Export dataset	14
3.2.2 Choose subset	14
3.2.3 Partition dataset	14
3.2.4 Attach/detach marker information	14
3.2.5 Attach/detach assay information	14
Chapter 4: Data analysis	15
4.1 Summary statistics	15
4.1.1 Basic statistics	15
4.1.2 Allele and genotype frequencies	16
4.1.3 Haplotype frequencies	16
4.1.4 Hardy-Weinberg disequilibrium	18
4.1.5 Two-Locus Linkage disequilibrium	19
4.1.6 Multi-Locus linkage disequilibrium	20
4.2 Design	20
4.2.1 Line selection	20
4.3 Structure analysis	20
4.3.1 Population differentiation test	20
4.3.2 Classical F-statistics	20
4.3.3 Population specific F-statistics	21
4.3.4 Coancestry matrix	21
4.4 Phylogenetic analysis	21

- 4.4.1 Compute frequency 21
- 4.4.2 Frequency-based distance 21
- 4.4.3 Tree reconstruction 21
- 4.4.4 Bootstrap 21
- 4.5 Association analysis..... 22
 - 4.5.1 Single-Locus case control test 22
 - 4.5.2 Single-locus F-test 22
 - 4.5.3 Haplotype trend regression 22
- 4.6 Tools 23
 - 4.6.1 SNP simulation 23
 - 4.6.2 Mantel test..... 23
 - 4.6.3 Contingency table exact p-value 23
- APPENDIX A: Estimating population specific F-statistics 25
- APPENDIX B: List of frequency-based distances 29
- REFERENCES 32

Chapter 1: Introduction

1.1 What is PowerMarker?

With genetic markers becoming basic tools for geneticists, the need for reliable computer software to perform statistical analysis of marker data has grown. One of the main reasons that we have developed the PowerMarker package is to satisfy this need for elegant, but simple, reusable solutions. PowerMarker delivers a data-driven, integrated analysis environment (IAE) for marker data. The IAE integrates the data management, analysis and visualization in a user-friendly graphic user interface. It accelerates the analysis lifecycle, and enables users to maintain data integrity throughout the process.

1.2 Data types handled by PowerMarker

PowerMarker handles a variety of marker data, either haplotypic or diplotypic. The gametic phase for the diplotypic data can be known or unknown. Examples of marker data include microsatellite data, single nucleotide polymorphism (SNP) data and RFLP data.

There are six different types of data objects in PowerMarker: Dataset, Frequency, Distance, Tree, Table and Text. Each of them can be imported from and exported to text files. Dataset stores several different types of genetic information and is the input for most of the analyses. Frequency data is the input for distance analysis and can be generated from Dataset. Distance data is the input for tree analysis and the output for distance analysis. Tree data is the output for tree analysis and bootstrap analysis. Most analyses generate table data output and/or text data output. Analyses also generate logs, which were stored as text data.

To see the details for the manipulation and input format specification for each data type, refer to chapter 3: Data manipulation.

1.3 Methods implemented in PowerMarker

The analyses implemented in PowerMarker fall into the following six main categories.

Category	Short description
Summary statistics	Summary statistics such as allele number, gene diversity, inbreeding coefficient; estimation of allelic, genotypic and haplotypic frequency; Hardy-Weinberg disequilibrium and linkage disequilibrium
Design	Choose core set of lines or haplotype tagging markers
Population structure	Population differentiation test; classical F-Statistics as well as population specific F-Statistics; population structure estimation
Phylogenetic analysis	Frequency, distance and tree analysis; bootstrap trees
Association analysis	Association test for different designs
Tools	Utility tools such as SNP simulation and identification, Mantel test and exact p-values for contingency tables

More details for each analysis can be found at chapter 4: Data analysis.

1.4 System requirements

- Windows 98/NT/2000/XP
- Microsoft .NET framework redistributable
- A minimum of 64 MB RAM
- At least 100 Mb free hard disk space
- Microsoft Excel 2000 or above is required for Excel integration
- TreeView is required for tree viewing

1.5 Installing and uninstalling PowerMarker

Before you install PowerMarker, make sure you have installed Microsoft .Net framework redistributable. You can get it for free from

<http://msdn.microsoft.com/library/default.asp?url=/downloads/list/netdevframework.asp>

After the download of the single file “PowerMarker installer.msi”, execute the installer. The installer will install PowerMarker and add an icon to the desktop. Then start PowerMarker by double-click on the desktop icon. Registration is strongly recommended. You will get email of upgrade information from PowerMarker if you are a registered user. **Email support is restricted to registered users.** Registration is free and can easily be done by selecting “Help | Register PowerMarker” in the PowerMarker graphic user interface.

If you want to uninstall PowerMarker, uninstall it from control panel instead of simply deleting the folder where you installed PowerMarker.

1.6 How to cite PowerMarker

Citation information can be obtained by selecting **Help | How To Cite PowerMarker** in the PowerMarker program.

1.7 Acknowledgements

This program has been made possible by NSF grants No. DBI-0096033 and No. DEB-9996118.

1.8 Upgrade PowerMarker

PowerMarker is still on its early stage of development. Bugs might be fixed shortly after it was detected. Then a minor version upgrade might be performed and the new version will replace the old one for download. The manual might also be upgraded. To check if you have the latest version of PowerMarker, choose **Help | Check Update**.

In order to upgrade PowerMarker, you must uninstall the old version of PowerMarker at first, and then install the new version of PowerMarker as described in section 1.5.

Chapter 2: Tutorial

This tutorial is designed to demonstrate the graphic interface of PowerMarker. Following the steps in this tutorial will allow the user to learn about using PowerMarker's integrated analysis environment to perform analysis. This tutorial shows how to:

- Create a project for the analyses
- Import a dataset from a text file
- Choose a subset from the dataset
- Produce a table of summary statistics

Launch the PowerMarker application by double-clicking its icon  on the desktop to begin the tutorial.

2.1: Creating a project

Before performing an analysis in PowerMarker, you must first create a project to work in. PowerMarker uses a project file with a **.prj** extension to organize data objects and folders. If this is your first time to run PowerMarker, you will notice a project **Default** is automatically created and displayed in the explorer like this:

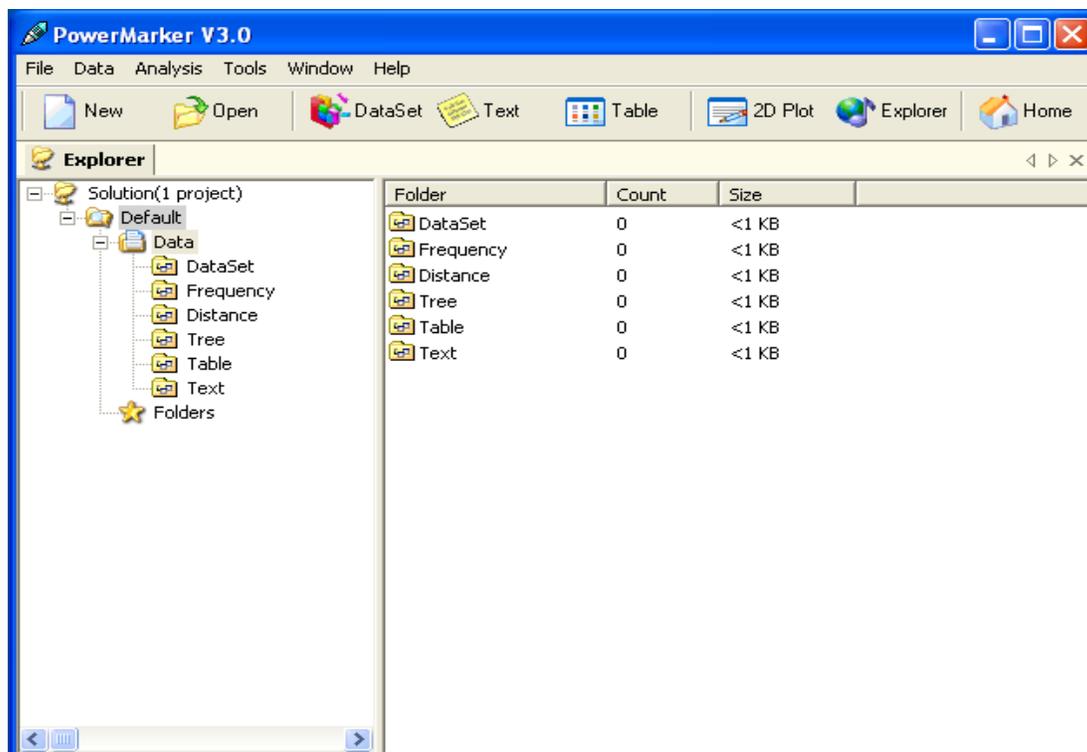


Figure 2.1: The object explorer in PowerMarker

The following steps can be used to create your own project:

- Choose **File | Close All Projects** to close all projects.
- Choose **File | Add New Project** or click the **New** button  on the main toolbar to open the file dialog to save the new project.
- Accept the default directory name, and type **fbi** in the file name field.

- Click the **Save** button to close the dialog.

Now the interface changes back to Figure 5.1 except that the name of the project name has been changed to **fbi**.

2.2: Importing a dataset

The majority of the analyses in PowerMarker works on a Dataset. A Dataset is a serialized object of genetic marker data. To import a Dataset from a text file, follow these steps:

- Choose **File | Import | Dataset** or click the **Dataset** button  on the main toolbar to open the Dataset wizard.
- Click **Browse** button to open the file dialog, Choose the file **fbi.txt** from **<PowerMarker>\Sample\FBI**, where **<PowerMarker>** is the directory where you installed PowerMarker. Step 1 of the Dataset wizard should look like this:

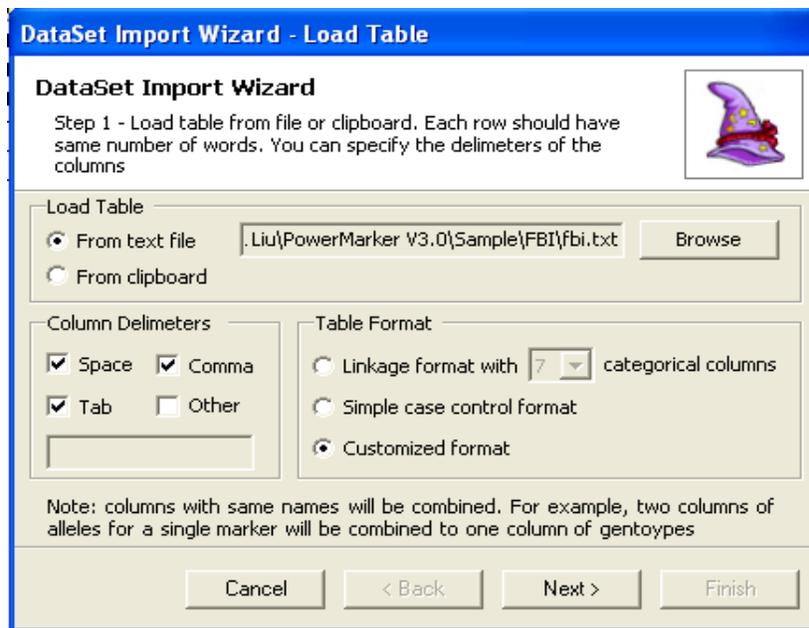


Figure 2.2: Step 1 of Dataset wizard

- Click the **Next** button to go to step 2 of the Dataset wizard.
 - Select the first two columns **Sample** and **ID#**, and click the link **Categorical** to change these two columns to categorical types. All the other columns are accepted as marker types.
 - Select **ID#** from the drop down list of Level-1 Column combobox.
 - Select **Sample** from the drop down list of Level-2 Column combobox. Step 2 of the Dataset wizard should look like this:

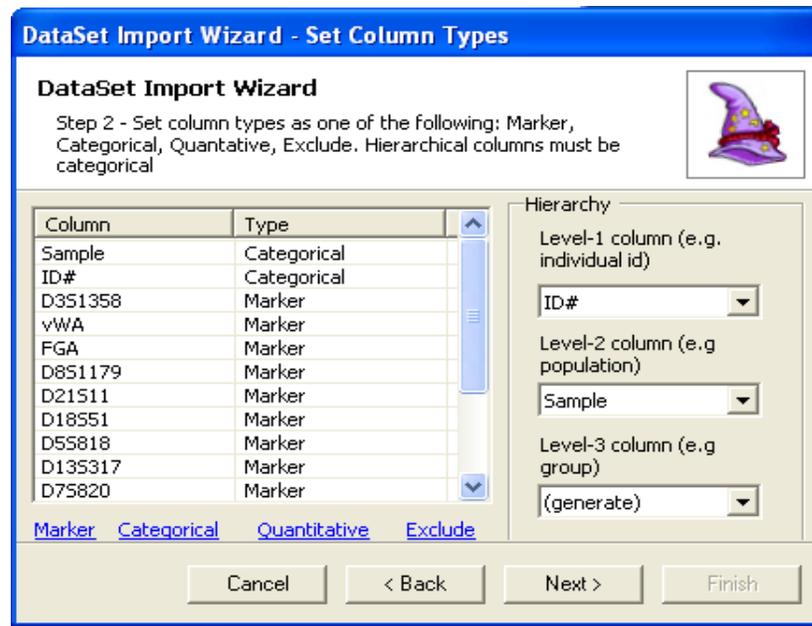


Figure 2.3: Step 2 of Dataset wizard

- Click the **Next** button to go to step 3 of the Dataset wizard. Accept all the settings in step 3.
- Click the **Next** button to go to step 4 of the Dataset wizard. Step 4 of the wizard should like this:

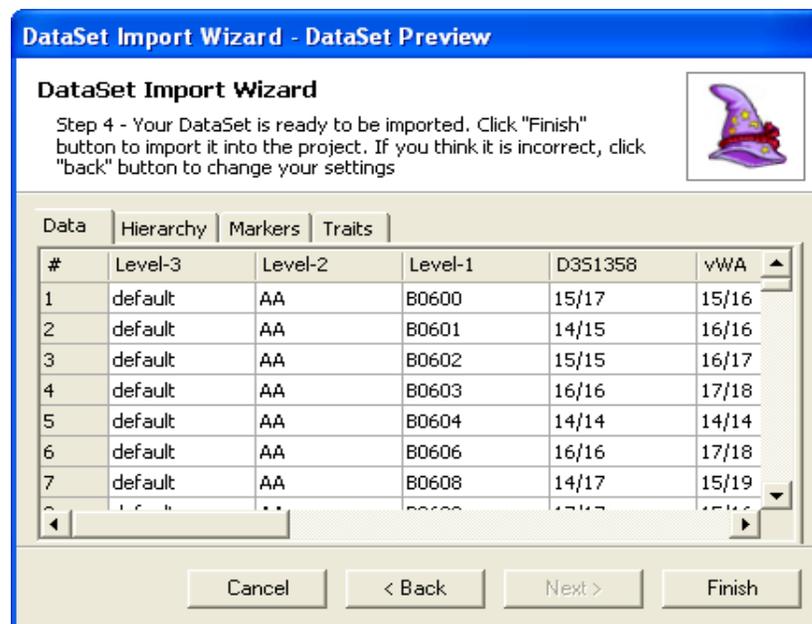


Figure 2.4: Step 4 of Dataset wizard

- Click the **Finish** button to go back to the explorer. You will notice that the Dataset *fbi* has been imported to the project.

2.3: Choosing a subset from the dataset

This step will generate a new dataset from *fbi* by excluding loci with a large missing proportion (>0.05).

- Right click the newly generated Dataset *fbi* and select **Choose Subset** from the pop-up menu. The choose subset dialog will appear.
- Under **Choose assays** tab, click **Select All** button. All the 622 assays will be selected.
- Switch to the **Choose markers** tab. By default none of the markers are selected.
 - Select **Missing proportion** from the drop down list of the combobox.
 - Click the **Compute property** button.
 - Click the header **Missing proportion** in the listview to sort the items.
 - Select the first 6 markers. The dialog should look like this:

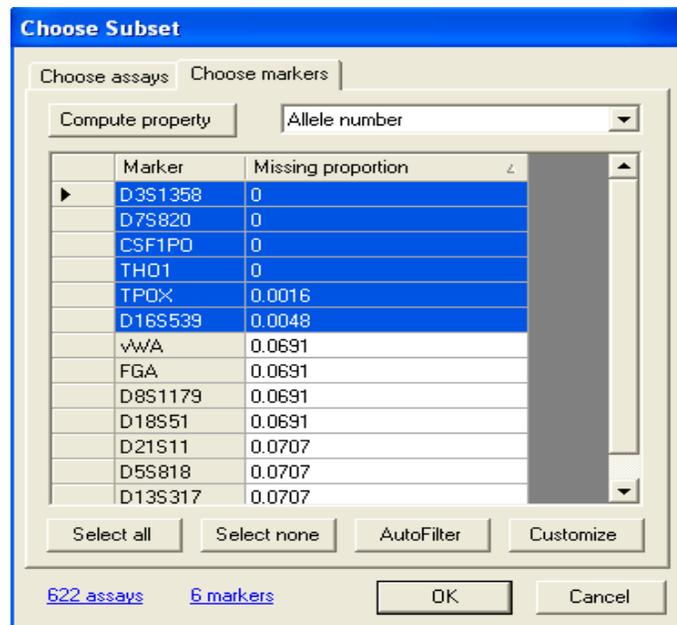


Figure 2.5: Choose subset dialog

- Click **OK** button to close the dialog. A new Dataset *fbi.Subdata* will appear in the explorer.

2.4: Producing a table of summary statistics

In this step we will generate a table of summary statistics for the dataset *fbi.Subdata*.

- Choose **Analysis | Summary | Summary Statistics** to open the analysis dialog. Make the following changes in the appropriate fields in the dialog:
 - Select **fbi.Subdata** in the Data ListBox.
 - Type **Summary** as the name of the result folder. The dialog should like this:

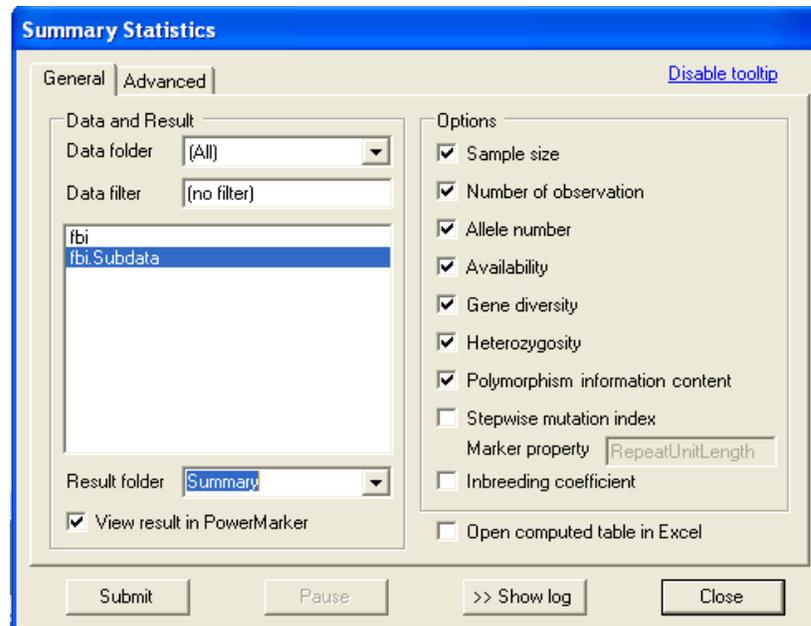


Figure 2.6: Analysis dialog for summary statistics

- Click **Submit** button to perform the analysis. The analysis should be finished immediately. PowerMarker will automatically save the result and open it in TableViewer.

Marker	SampleSize	No. of obs.	AlleleNo	Availability	GeneDiversity	Heter
D3S1358	622	622	10	1	0.7695	0.776
D7S820	622	622	9	1	0.7931	0.779
CSF1PO	622	622	10	1	0.746	0.752
TPOX	622	621	8	0.9984	0.679	0.677
THO1	622	622	8	1	0.7801	0.736
D16S539	622	619	8	0.9952	0.789	0.794
Mean	622	621.3333	8.8333	0.9989	0.7595	0.753

Figure 2.7: Table viewer in PowerMarker

- Right click in the TableViewer and select **Open In Excel** from the pop-up menu. The table will be opened in Excel.

Chapter 3: Data manipulation

3.1 Data import

3.1.1 Import dataset

To import a PowerMarker dataset, you must have a plain text of genetic marker data. PowerMarker does not require a specific input format such as NEXUS format. Instead, PowerMarker supports table-like data format directly.

Dataset format: each row represents all the genetic and trait information for a single assay/individual. There can be any numbers of columns. The first row of the table must be a heading row representing the names of columns. Each column falls into one of the four categories: marker, categorical, quantitative or exclude. Columns with same names will be combined. In other words, two columns of alleles for a single marker will be combined to one column of genotypes. Exclude columns will not be imported. You can use any non-empty string to represent missing data as long as this string is consistent for all the columns of the same type.

There may be up to three hierarchical columns, which are special categorical columns. The three hierarchical columns are: level-1 column, level-2 column and level-3 column. Level-1 column is the column of assay IDs and it must be unique for each assay. Level-2 column is the second level of the hierarchy and usually represents a population or a sample. Level-3 is the next level of hierarchy representing groups or superpopulations. None of these three columns are required. If not defined, the hierarchical columns will be automatically generated.

An example of a dataset:

Sample	Trait1	Trait2	Marker1	Marker2	Marker3
Sample1	+	3.1	0/0	12/14	A/A
Sample1	?	4.5	1/0	12/12	B/B
Sample2	-	3.3	0/1	13/14	A/B
Sample2	-	2.0	1/1	14/14	A/A

This example includes three marker columns, two categorical columns (Sample and Trait2) and one numeric column (Trait1). The sample column can be set as a level-2 column.

A special format supported by PowerMarker is the widely used linkage format. A linkage file can be regarded as a PowerMarker dataset with fixed column properties. To import a linkage file, choose linkage format in the first page of the dataset wizard.

NOTE: Linkage file does not have a header row of marker names. The header row is generated in the importing process.

Dataset consists of four components:

Genotype information: a set of loci, each locus represents all the genetic information for a single genetic marker

Assay information: categorical or numeric traits for assays

Hierarchy: up to three levels of population structure

Marker information: mapping information or other marker properties

PowerMarker dataset consists of four different components. The first component is genotype information, which is just a set of loci. Each locus stores the genotype information of all assays for a single genetic marker. For each assay at a locus, the genetic information can be a single allele for haplotype data or a genotype of two alleles for genotype data. A single marker column in the table will be converted to a single locus. Assay information stores categorical or numeric trait for each assay. Hierarchy stores the population structure of the whole sample. Marker information stores mapping information for each marker. Other marker properties can also be provided (e.g. the repeat unit length for a microsatellite locus). Except for the first component, all the other three components can be changed after the dataset was imported. For example, you can change the hierarchy of the population structure to reflect different types of grouping in your data.

Not all analyses use the information of all the four components although genotype information is always used. Hierarchy information is required for structure analysis such as F-statistics estimation. Association study requires assay information, and some data manipulation uses the marker information. For example, if you want to sort the markers by their chromosomal positions, then the mapping information must be provided.

3.1.2 Import frequency data

In some cases, the allelic composition of the assayed individuals is not specified. The only information we have are the frequencies of the alleles. Distance analysis in PowerMarker is based on frequency. Frequency was imported from a text file with a specific format.

Frequency format: each row represents the frequency of an allele at a single marker for all the observed taxonomy units (OTUs). The table must have a heading row representing OTUs. The first column is the column of marker names; the second column is the column of alleles; the following columns are the allele frequencies.

An example of frequency table:

Marker	Allele	OTU1	OTU2	OTU3
Marker1	A1	0.4	0.3	0.8
Marker1	A2	0.6	0.7	0.2
Marker2	B1	0.1	0.5	0.6
Marker2	B2	0.9	0.5	0.6

This frequency table has two markers, each of which has two alleles. Three OTUs have allele frequency for each allele at each marker.

Frequency can be generated from a dataset object by choosing **Analysis | Phylogeny | Compute Frequency** in PowerMarker.

3.1.3 Import distance data

Distance in PowerMarker needs to be imported if you want to construct NJ/UPGMA tree from your own distance. The program “neighbor.exe” in Phylip package does the same task. Distance was imported from the same format as in Phylip.

Distance format: if you have n OTUs, there should be n rows and n+1 columns. The first column is the column of OTU names. The following n columns are just the distance matrix. This distance matrix should be symmetric and zero-diagonal.

An example from Phylip is here:

Alpha	0.0000	1.0000	2.0000	3.0000	3.0000
Beta	1.0000	0.0000	2.0000	3.0000	3.0000
Gamma	2.0000	2.0000	0.0000	3.0000	3.0000
Delta	3.0000	3.0000	3.0000	0.0000	1.0000
Epsilon	3.0000	3.0000	3.0000	1.0000	0.0000

3.1.4 Import tree data

Tree in PowerMarker can be imported from Phylip format text files. Usually Tree objects were not imported but generated. To get the detail of Phylip format of the tree string, refer to Phylip homepage at <http://evolution.genetics.washington.edu/phylip.html>.

3.1.5 Import table data

You can import tables into PowerMarker as Table objects. Tables can come from files or clipboard. The table importing in PowerMarker is similar to most of other statistical packages such as S+ and R. No specific format is required for the table except that it must be a valid table which means each row in the table has the same number of columns. Heading row can be provided or generated. The first column (column of row names) can also be generated if not provided. By default a column will be parsed to an appropriate type (e.g. a column with all numeric values will be parsed to a numeric column) but the user can customize the parsing procedure.

The easiest way to import a table from Excel is to select the table in Excel and copy the selected table to the clipboard and import Table in PowerMarker from the clipboard.

NOTE: it is very important to make sure the types of columns are parsed as expected. Pay extra attention to columns with only 0s and 1s. Do you want to keep it as a categorical column or numeric column? PowerMarker by default will choose the latter.

3.1.6 Import text data

Text objects in PowerMarker can be imported directly from text files or clipboard. No specific format is required.

3.2 Dataset manipulation

All the following functions are selected from dataset's pop-up menus by right clicking a dataset in the explorer.

3.2.1 Export dataset

Dataset can be exported to a text file in five different formats: raw format, table format, "Structure" format required by Pritchard analysis, NEXUS format required by GDA and Arlequin format required by Arlequin.

3.2.2 Choose subset

PowerMarker provides a variety of approaches to choose a subset from a dataset. Both assays (individuals) and markers can be filtered manually by the user or based on a property of the assay/marker. The user can even provide a list of assays/markers to be selected or excluded.

3.2.3 Partition dataset

A dataset can be partitioned into several datasets by the attributes of individuals or markers, or by the genotypes for a specific marker.

3.2.4 Attach/detach marker information

Marker information, such as chromosome and position, should be provided to a dataset if possible. In order to attach marker information to a dataset, first you need a table with both the header row and a column of row names. The first column of the table must be marker names (which will be mapped to marker names in the dataset), and each column of the table represents a covariate of the markers. Both categorical and numeric covariates are supported. Marker information can be detached from a dataset.

3.2.5 Attach/detach assay information

Dataset wizard provides a user interface to import both genotypes and covariates for assays. Additional assay information can be provided after the dataset is imported to the explorer. First this assay information is imported as a table object (set the column types carefully!), then PowerMarker will try to extract the related information from the table and merge the information into the existing dataset. Like marker information, assay information can also be detached from a dataset.

Chapter 4: Data analysis

For a more detailed description of the equations and methods in this chapter, see the cited pages from Weir (1996) unless otherwise noted. When possible, we use the notation and concepts in Weir (1996).

Suppose we are given n individuals and m polymorphic loci. The symbol A will be used to mean any genetic locus with a series of alleles A_u . For an individual, a single-locus genotype or a single allele is observed for each locus. An allele A_u has a population frequency p_u (or p_{lu} , to indicate the l th locus), and a genotype A_uA_v has a population frequency P_{uv} (or P_{luv}). Sample frequencies will be indicated by tildes, and these observed values are also used as estimates of allelic and genotypic frequencies. Estimates will be indicated by carets. In a sample, counts of alleles and genotypes will be written as n_u and n_{uv} (or n_{lu} and n_{luv} for the l th locus), respectively.

4.1 Summary statistics

4.1.1 Basic statistics

Number of observation

The number of observation for a marker locus is defined as the number of nonmissing alleles (for haploid data) or nonmissing genotypes (for diploid data) observed in the sample. A genotype is regarded as missing if one of its two alleles is missing.

Availability

Availability is defined as $1 - Obs/n$, where Obs is the number of observations and n is the number of individuals sampled.

Stepwise mutation index

Step wise mutation index is defined as the maximal proportion of alleles that follows the stepwise mutation pattern (for microsatellite data only). This statistic needs the length of the repeat unit be specified. This information has to be included in the covariate table of the markers and the column property needs to be numeric.

Within-population inbreeding coefficient

An EM algorithm (pp. 77-78) is used to find the MLE of the within-population inbreeding coefficient. Note that the EM algorithm may fail to converge for negative values of inbreeding coefficient. The same parameter is estimated using the method of moments (pp. 79-80).

Heterozygosity

Heterozygosity (pp. 141-150) is simply the proportion of heterozygous individuals in the population. At a single locus it is estimated as

$$\hat{H}_l = 1 - \sum_{u=1}^k \tilde{P}_{luu}$$

Gene diversity

Gene diversity (pp. 150-156), often referred to as expected heterozygosity, is defined as the probability that two randomly chosen alleles from the population are different. An unbiased estimator of gene diversity at the l th locus is

$$\hat{D}_l = (1 - \sum_{u=1}^k \tilde{p}_{lu}^2) / (1 - \frac{1+f}{n}),$$

where the inbreeding coefficient, f , is estimated from the data using the method of moments (pp. 79-80). The user can also request the common biased estimator of the gene diversity,

$$\widehat{D}_l = (1 - \sum_{u=1}^k \tilde{p}_{lu}^2).$$

Polymorphism information content

A closely related diversity measure is the polymorphism information content (PIC) (Botstein *et al.* 1980). It is estimated as

$$\widehat{PIC}_l = 1 - \sum_{u=1}^k \tilde{p}_{lu}^2 - \sum_{u=1}^{k-1} \sum_{v=u+1}^k 2\tilde{p}_{lu}^2 \tilde{p}_{lv}^2$$

Bootstrap across loci

For all of these summary statistics, the overall estimates are calculated as the average across all loci, whereas variances and confidence intervals are estimated by nonparametric bootstrapping across different loci.

Summary tables for different hierarchical levels

By default PowerMarker generates the summary statistics based on all the samples in the Dataset. The user can request that the summary table to be generated at different hierarchical levels.

4.1.2 Allele and genotype frequencies

The sample allele frequencies are calculated as $\tilde{p}_u = n_u / (2n)$, with the variance estimated as

$$\text{var}(\tilde{p}_u) \hat{=} \frac{1}{2n} (\tilde{p}_u + \tilde{P}_{uu} - 2\tilde{p}_u^2),$$

where $\hat{=}$ means “estimated by”.

The sample genotype frequencies \tilde{P}_{uv} are calculated as n_{uv} / n . Both the \tilde{p}_u s and \tilde{P}_{uv} s are unbiased maximum likelihood estimates (MLEs) of the population frequencies. Confidence intervals for allele and genotype frequencies are formed by resampling individuals from the data set.

4.1.3 Haplotype frequencies*Haplotype frequency for unrelated population*

The EM algorithm (Excoffer and Slaktkin 1995), which is implemented in PowerMarker to estimate haplotype frequencies, is an iterative method to reconstruct haplotypes and find frequencies (F) that maximize the likelihood of the genotype data. Under the

assumption of Hardy-Weinberg equilibrium, the likelihood is the product of the probabilities of each individual conditional on haplotype frequencies:

$$L(F) = \prod_{i=1}^n \Pr(G_i | F),$$

where G_i is the genotype of the i th individual and n is the sample size.

Define S_i as the set of ordered pairs of haplotypes that constitute the genotype G_i . The E-step refreshes the genotype frequencies from the haplotype frequencies as follows:

$$P_i = \sum_{[j,j'] \in S_i} p_j p_{j'},$$

where $[j, j']$ is the ordered pair of j th haplotype and j' th haplotype, and p_j and $p_{j'}$ are the current frequencies of j th and j' th haplotypes, respectively..

At the M-step, maximal likelihood estimates of these haplotype frequencies are obtained and used in turn as the haplotype frequencies at the next iteration:

$$p_k = \frac{1}{2n} \sum_{i=1}^n \sum_{[j,j'] \in S_i} \frac{m_{jj'} p_j p_{j'}}{P_i},$$

where

$$m_{jj'} = \begin{cases} 2 & \text{if } j = j' = k \\ 1 & \text{if } j \neq j', j = k \text{ or } j' = k. \\ 0 & \text{otherwise} \end{cases}$$

The E- and M-step are iterated until the likelihood $L(F)$ converges.

The global convergence of the EM algorithm depends heavily on the starting state. Four different initialization methods are implemented in PowerMarker. **InitMethod=Uniform** assigns equal frequency to all haplotypes, **InitMethod=Random** initializes haplotype frequencies with random values from a Dirichlet distribution. **InitMethod=LinkEq** initializes haplotype frequencies based on the linkage equilibrium assumption. **InitMethod=CHM** initializes haplotype frequencies by the composite haplotype method (Zaykin etc. 2001). The estimation procedure is fixed for *Uniform*, *LinkEq* and *CHM* starts. If multiple random starts are used, the solution with the largest likelihood is retained. The default method is *CHM*. Our experience suggests a single start of *CHM* is usually sufficient.

Care should be taken for the estimation of long haplotypes. Unless the linkage disequilibrium in the whole region is extremely strong, very little value can be gained by estimating haplotype frequencies for dataset with more than 20 markers. Also note that a very large sample size are needed to capture most of the haplotype variation for long chromosomal regions. PowerMarker uses a bisection mechanism to improve the efficiency for large-scale haplotype estimation. When the marker number is above a cutoff (default=20), the whole region is bisected into two subregions and haplotype estimation is performed separately for each subregion. Haplotypes with a small frequency (the cutoff can be changed by the user) are discarded. Then the estimation of the whole region is performed based on the haplotype pool that is generated by the deflated haplotype pools for the two subregions.

The standard error and confidence interval for each haplotype can be estimated by bootstrapping across the samples. The user can also request for a phase assignment table be generated. The table reports the probability of each genotype can be resolved into each of the possible haplotype pairs.

Haplotype frequency estimation for trio families

Statistical methods for unrelated population data consider all possible haplotype pairs consistent with the independent genotypes and provide the complete haplotype pool, which might be larger than the real one. In PowerMarker we developed a new approach to efficiently estimate the MLEs of the haplotype. Incorporating pedigree information can help solve some ambiguous phases and eliminate a majority of impossible haplotypes, which improves both the accuracy and the capability of statistical methods. Consider the trio design, where all children are assumed to be independent. The constraint of parental information on the possible phase of the single child will significantly reduce the number of compatible haplotype pairs for each child. The EM algorithm is then applied to the independent children's genotype data, based on the reduced haplotype set S_i^C for i th individual:

$$\begin{aligned} \text{E-Step: } P_i &= \sum_{[j,j'] \in S_i^C} p_j p_{j'}, \\ \text{M-Step: } p_k &= \frac{1}{2n} \sum_{i=1}^n \sum_{[j,j'] \in S_i^C} \frac{m_{jj'} P_j P_{j'}}{P_i}. \end{aligned}$$

To estimate the haplotype frequency of the parents based on the likelihood of the trios, the method by Rohde and Fuerst (2001) are also implemented. Simulation study shows this method is much more accurate than the EM algorithm based on the unrelated samples formed by parents only.

4.1.4 Hardy-Weinberg disequilibrium

For a single locus, the MLE of the disequilibrium coefficient D_{uv} for alleles A_u and A_v is calculated as

$$\hat{D}_{uv} = \begin{cases} \tilde{P}_{uv} - \tilde{p}_u \tilde{p}_v, & u = v \\ \tilde{p}_u \tilde{p}_v - \frac{1}{2} \tilde{P}_{uv}, & u \neq v \end{cases},$$

and the variance is estimated using the follow formulas:

$$\begin{aligned} \text{Var}(\hat{D}_{uu}) &\triangleq \frac{1}{n} \left[\tilde{p}_u^2 (1 - \tilde{p}_u)^2 + (1 - 2\tilde{p}_u)^2 \hat{D}_{uu} - \hat{D}_{uu}^2 \right] \\ \text{Var}(\hat{D}_{uv}) &\triangleq \frac{1}{2n} \left\{ \tilde{p}_u \tilde{p}_v (1 - \tilde{p}_u)(1 - \tilde{p}_v) + \sum_{w \neq u, v} (\tilde{p}_u^2 \hat{D}_{uw} + \tilde{p}_v^2 \hat{D}_{vw}) \right. \\ &\quad \left. - \left[(1 - \tilde{p}_u - \tilde{p}_v)^2 - 2(\tilde{p}_u - \tilde{p}_v)^2 \right] \hat{D}_{uv} + \tilde{p}_u^2 \tilde{p}_v^2 - 2\hat{D}_{uv}^2 \right\}. \end{aligned}$$

Bootstrap confidence intervals are formed by resampling individuals from the data set. Three different methods are used to test for Hardy-Weinberg Equilibrium. The chi-square goodness-of-fit test is formed by calculating the chi-square statistic

$$X_T^2 = \sum_u \frac{(n_{uu} - n\tilde{p}_u^2)^2}{n\tilde{p}_u^2} + \sum_u \sum_{v \neq u} \frac{(n_{uv} - 2n\tilde{p}_u\tilde{p}_v)^2}{2n\tilde{p}_u\tilde{p}_v}.$$

This statistic has $k(k-1)/2$ degrees of freedom where k is the number of alleles at the marker locus. The same distribution is shared by the likelihood ratio test described in Weir (pp. 105-106). A permutation version of the exact test given by Guo and Thompson (1992) is also implemented (pp. 109-100).

4.1.5 Two-Locus Linkage disequilibrium

Two-locus linkage disequilibrium D_{uv} is defined for two alleles at different loci as $D_{uv} = p_{uv} - p_u p_v$. It is estimated by $\hat{D}_{uv} = \tilde{p}_{uv} - \tilde{p}_u \tilde{p}_v$ for haplotype data or phased genotype data, or by $\hat{D}_{uv} = \hat{p}_{uv} - \tilde{p}_u \tilde{p}_v$ for unphased genotype data. Based on the estimates of D_{uv} , five linkage disequilibrium measures are calculated for each pair of alleles at two loci: the correlation coefficient r^2 , Lewontin's D' , the proportional difference d , the population attributable risk δ , and Yule's Q . These measures are discussed in Devlin and Risch (1995).

For biallelic loci, all the four possible allele-pair linkage disequilibrium (LD) statistics are equivalent (the signs may be different). For multiallelic loci, however, locus-pair linkage disequilibrium has to be defined separately. We define locus-pair LD statistics in PowerMarker as a weighted mean of the absolute allele-pair linkage disequilibria:

$$LD = \sum_u \sum_v p_u p_v |LD_{uv}|,$$

where LD can be linkage disequilibrium, Lewontin's D' or correlation coefficient r^2 .

The chi-square statistic to test that all the pairwise linkage disequilibrium D_{uv} are zero is calculated as follows:

$$X_T^2 = \sum_{u=1}^k \sum_{v=1}^l \frac{(2n)\hat{D}_{uv}^2}{\tilde{p}_u \tilde{p}_v}.$$

This statistic has $(k-1)(l-1)$ degrees of freedom for markers with k and l alleles, respectively. The exact test for testing whether two-locus genotype frequencies are the products of one-locus genotype frequencies (for genotype data), or testing whether two-locus haplotype frequencies are products of allele frequencies (haplotype data), are also implemented. The counts of alleles/genotypes are organized as a contingency table and p-values are evaluated through several different methods depending on the size of the table (see section 4.6.3). The details of these methods can be found in Weir (pp. 127-128).

Pairwise linkage disequilibrium can be summarized in a two-dimensional table and visualized by a 2-D plot in PowerMarker. Sometimes the markers are ordered by their physical positions along the chromosomal. In this case a distance-LD plot can be informative. PowerMarker performs the plot directly in Excel, allowing for the user to easily modify the plots. If the physical position for each marker is available, the plot can be based on the physical distance. Otherwise the physical distance is replaced by the window size between the markers.

4.1.6 Multi-Locus linkage disequilibrium

The exact test for multi-locus association, described by Zaykin *et al.* (1995), is implemented. The null hypothesis of the test says that a multi-locus genotype frequency is equal to the product of corresponding allele frequencies (genotypes are broken down; this implicitly assumes HWE), or of corresponding one-locus genotype frequencies (genotypes are preserved). Permutation is used to examine if the probability of multi-locus genotypes conditional on alleles or one-locus genotypes lies in the tail of the distribution.

4.2 Design

4.2.1 Line selection

In this module we use a simulated annealing algorithm for choosing a core set of lines from a large germplasm collection. Any measure of core set quality can be used in the algorithm. A very attractive feature of our algorithm is that general constraints can be incorporated in the selection. The algorithm can be used to find the minimal set of lines with maximal diversity, or given a sample size find the optimal core set of all possible sets. Our algorithm is computationally efficient and adjustable. With weak convergence conditions the algorithm finds nearly optimal local maxima very quickly. Under strong convergence conditions global maxima are almost guaranteed. The detail of the algorithm can be found online at <http://www.powermarker.net/downloads/coreset.pdf>. A batch script system is developed for supporting user-defined constraints and settings.

4.3 Structure analysis

4.3.1 Population differentiation test

Two different of population differentiation test are implemented in PowerMarker. The first analysis uses a contingency table approach to determine if groups of individuals have significant differences in allele frequencies for each locus (Raymond and Rousset 1995), while the second analysis tests the overall differentiation of groups by using a variant of the Mantel test. The second approach is borrowed from Mark Miller's program MANTEL-STRUCT, which is available online at <http://bioweb.usu.edu/mpmbio/>. For the overall differentiation test, the genetic distance matrix is constructed by calculating the shared allele distance for each pair of individuals. Note PowerMarker always uses level-2 as the grouping variable.

4.3.2 Classical F-statistics

A very useful measure of population subdivision is the F-statistics developed by Wright (1965). F-statistics can be thought of as a measure of the correlation of alleles within individuals and are related to inbreeding coefficients. F-statistics describe the amount inbreeding-like effects within subpopulations (F_{ST} or θ), among subpopulations (F_{IS} or f), and within the entire population (F_{IT} or F).

PowerMarker performs several different types of F-statistics analysis. The first type works on haploid data or diploid data (assuming HWE for diploid data), and reports the overall estimate $\hat{\theta}$ and an estimator for each locus. The details are given in Weir (pp. 170-

174). When analyses are performed at the genotypic level, the same general approach is followed except three levels of F-statistics are now estimated (pp. 176-179). A three-level hierarchical analysis, described by Weir (pp. 184-186), is also implemented under the optional assumption of Hardy-Weinberg Equilibrium.

4.3.3 Population specific F-statistics

Population specific F-Statistics extends the classical F-statistics by allowing different levels of coancestry for different populations, and by allowing non-zero coancestries between pairs of populations. The procedure of estimating population specific F-Statistics and between-population F-Statistics was formulated in Weir and Hill (2002). An extension of the estimation procedure, which works on genotype frequencies instead of allele frequencies, can be found in Appendix A.

4.3.4 Coancestry matrix

A coancestry matrix is formed by calculating θ for each pair of populations. The user can request for the log transformation ($= -\ln(1 - \theta)$) to be performed, which leads to a measure of genetic distance under a drift model (pp. 194-195).

4.4 Phylogenetic analysis

4.4.1 Compute frequency

Dataset can be converted to Frequency data by this module. The user can request that the computation be computed at different hierarchical levels. The output of this module can easily be imported to other statistical package for a principal component analysis.

4.4.2 Frequency-based distance

The calculation of a genetic distance between two populations gives a relative estimate of the time that has passed since the populations have established. Small estimates of distance may indicate population substructure, or indicate the populations have only been separated for a short period of time. When two populations are genetically isolated, mutation and genetic drift lead to differentiation in the allele frequencies at selectively neutral loci. As the amount of time that two populations are separated increases, the difference in allele frequencies are expected to increase. Various distance measures used for frequency data have been described by Nei (1987) and Weir (1996). Appendix B lists the definitions and brief explanations of the distance measures implemented in the package.

4.4.3 Tree reconstruction

For evolutionary studies, the clustering of OTUs leads naturally to a phylogenetic tree. The following two algorithms are used to reconstruct the phylogeny from a distance matrix: UPGMA(unweighted pair-group method using arithmetic average) and Neighbor-joining (pp. 344-356).

4.4.4 Bootstrap

Felsenstein suggests the bootstrapping be performed over the marker loci (1985). Each bootstrap sample consists of same number of markers sampled with replacement from the original data set, and it then is subjected to the same distance calculation and tree

reconstruction. The output is a list of trees that can be summarized to obtain a consensus tree by the program “consensus” in Phylip package (Felsenstein 1993).

4.5 Association analysis

4.5.1 Single-Locus case control test

PowerMarker offers three methods for testing an association between a single marker and the affected status (must be binary). The allele case-control test and genotypic case-control test, implemented using a contingency table analysis, are described in Nielsen and Weir (1999). The allele test assumes HWE. The test statistics have an asymptotic chi-square distribution. Note that the degrees of freedom for the genotypic test will be the number of unique categories of genotypes examined in the data (either in case population or control population). In some cases, this number will not be the same as the theoretical number of genotypes ($= k(k+1)/2$, where k is number of alleles). The third method implemented in the package, the multi-allelic trend test (Slager and Schaid 2001), has the same degrees of freedom as allele test but remains valid even with the violation of HWE assumption.

When the frequency of an allele is small (e.g. the total count of this allele is < 5 in the sample), the asymptotic chi-square test may not be accurate. A permutation-based distance test can be performed to calculate the p-values. There are several different measures of distance provided by PowerMarker, of which the Prevosti distance is recommended by the author. By default PowerMarker calculates the distance of the two populations (case and control) based on allele frequencies. The user may request the distance be calculated based on genotype frequencies.

4.5.2 Single-locus F-test

Case control test is designed for binary traits, whereas F-test is suitable for quantitative traits. In this test each marker is regarded as a factor in a one-way ANOVA layout, as each genotype stands for a different level. F-test is then performed on each marker.

Both case-control test and F-test report a raw p-value for each marker. These p-values are not adjusted for multiplicity. It is up to the user to handle the multiple testing problems.

4.5.3 Haplotype trend regression

Zaykin et al. (2002) shows a regression approach to test haplotype-trait association can be more powerful than the omnibus chi-square test. This approach, called haplotype trend regression (HTR) by the authors, can be applied to both quantitative traits and binary traits. Haplotypes are constructed from the EM algorithm (see section 4.1.3) if the gametic phase is unknown. Asymptotic F-test is used to evaluate the significance, although a permutation based test can be more robust to the violation of model assumptions. Both methods are implemented in PowerMarker. The user can also define a lower bound of the haplotype frequency to be included in the model fitting. Sometimes this leads to a more powerful test.

The window size in HTR could be essential for the testing process. Ideally we wish to test the association between the trait and the whole genome (or a genomic region). In

practice this is not possible due to the lack of power of the test when the degree of freedom of the variables (haplotype number) becomes comparable or larger than the sample size (individual number). Furthermore, the frequency estimation can be very inaccurate for long haplotypes. We suggest a window size of 2 to 5 be used for SNP data.

4.6 Tools

4.6.1 SNP simulation

The coalescence simulation with the hotspot recombination model of Wiuf and Posada (2003) is implemented in PowerMarker. With no hotspot defined, the simulation becomes the classical coalescence model with homogenous recombination (Hudson 1983; Hudson and Kaplan 1990). Mutation is superimposed following an infinite-site model. The user can define a variety of parameters for the hotspot recombination model and the infinite-site model. The optional output of the module includes genealogy trees, the simulated probability distribution function of recombination rate, haplotypes and genotypes. More details of the algorithm and parameters can be found in the SNPSim package (2003).

4.6.2 Mantel test

The Mantel test (Mantel 1967) is a statistical method to determine the significance of correlation between two matrices (with the same dimension). The null hypothesis in the Mantel test is that distance in a matrix A are independent of the distances, for the same objects, in another matrix B. The test involves three different methods to evaluate the distribution of the correlation between the two matrices. For small matrices an exact enumeration of all the permutations can be listed to obtain an exact p-value. Asymptotic results can be applied to large matrices. For middle-size matrices, a random permutation procedure is used to evaluate the p-values.

The alternative hypothesis can be one-sided or two-sided. The default option automatically determines the hypothesis: if the estimated correlation between two distance matrices is negative, a one-tailed test is assumed involving the lower tail of the distribution; if the correlation is positive, the p-value for the upper tail is obtained. The user may want to choose a two-sided test or a specific one-sided test.

4.6.3 Contingency table exact p-value

This small tool calculates an unbiased estimate of the exact p-value of a contingency table using three different approaches: the MCMC approach described by Guo and Thompson (1992), Raymond and Rousset (1995); a permutation approach described as below; the asymptotic chi-square test. By default PowerMarker chooses the method automatically. If the total count of the table is small, a permutation approach is efficient and accurate. The MCMC approach, however, works fine even for very large counts.

For the MCMC approach, dememorization number is the number of steps to move the contingency table prior to starting the analysis. The use of batching procedure permits convergence to be checked automatically. Basically a p-value is calculated for each batch and the coefficient variation (standard deviation / mean) of the p-values is compared with the convergence criterion (default = 0.05) to determine if the convergence has been reached.

The permutation algorithm first recodes the contingency table into two arrays of indices, then the probability of the contingency table conditional on marginal counts is examined to see if it lies in the tail of the distribution generated by permutation (the permutation does not change the marginal counts).

APPENDIX A: Estimating population specific F-statistics

This appendix uses the notation and concepts described by Weir and Hill (2002). Define an indicator variable x_{ijk_u} for the k th allele of the i th individual in the j th population

$$= \begin{cases} 1 & \text{if allele is type } A_u \\ 0 & \text{otherwise.} \end{cases}$$

Then, population specific F-statistics θ_i , between-population F-statistics θ_{ii} , and population specific total inbreeding coefficient F_i are defined as the correlation between x_{ijk_u} and $x_{i'j'k'u}$:

$$\varepsilon(x_{ijk_u}) = p_u$$

$$\varepsilon(x_{ijk_u}^2) = p_u$$

$$\varepsilon(x_{ijk_u}, x_{i'j'k'u}) = \begin{cases} p_u^2 + p_u(1-p_u)\theta_{ii'} & i \neq i', j \neq j', k \neq k' \\ p_u^2 + p_u(1-p_u)\theta_i & i = i', j \neq j', k \neq k' \\ p_u^2 + p_u(1-p_u)F_i & i = i', j = j', k \neq k' \end{cases}.$$

Define

$$n. = \sum_{i=1}^r n_i$$

$$n_{ic} = n_i - \frac{\sum_{i=1}^r n_i^2}{n.}$$

$$\tilde{P}_{iu} = \frac{1}{2n_i} \sum_{j=1}^{n_i} \sum_{k=1}^2 x_{ijk_u}$$

$$\tilde{P}_u = \frac{1}{n.} \sum_{i=1}^r n_i \tilde{P}_{iu}$$

$$\pi_u = p_u(1-p_u)$$

$$\phi_i = \theta_i + \frac{1}{2n_i}(1 + F_i - 2\theta_i)$$

So that

$$\varepsilon(\tilde{P}_{iu}) = p_u$$

$$\varepsilon(\tilde{P}_{iuu}) = P_{iuu} = p_u^2 + \pi_u F_i$$

$$\text{Var}(\tilde{P}_{iu}) = \pi_u \theta_i + \frac{1}{2n_i} \pi_u (1 + F_i - 2\theta_i) = \pi_u \phi_i$$

$$\text{Cov}(\tilde{P}_{iu}, \tilde{P}_{i'u}) = \pi_u \theta_{ii'}$$

$$\varepsilon(\tilde{P}_u) = p_u$$

$$\text{Var}(\tilde{P}_u) = \frac{\pi_u}{n.^2} \sum_{i=1}^r \phi_i n_i^2 + \frac{1}{n.^2} \sum_{i=1}^r \sum_{i' \neq i}^r n_i n_{i'} \theta_{ii'}$$

If the terms in the mean square for individuals within populations (*MSI*) and for alleles within individuals (*MSG*) are weighed by n_{ic} instead of n_i , then the sum of squares corresponding *MSP*, *MSI* and *MSG* have expectations

$$\begin{aligned}\varepsilon(SSP) &= 2\varepsilon\left(\sum_{i=1}^r n_i (\tilde{p}_{iu} - \tilde{p}_u)^2\right) = 2\varepsilon\left(\sum_{i=1}^r n_i \tilde{p}_{iu}^2 - \left(\sum_{i=1}^r n_i\right) \tilde{p}_u^2\right) \\ &= 2\pi_u \left[\sum_{i=1}^r n_{ic} \phi_i - \frac{1}{n_{\cdot}} \sum_{\substack{i,i'=1 \\ i \neq i'}}^r n_i n_{i'} \theta_{ii'} \right] \\ \varepsilon(SSI) &= \varepsilon\left(\sum_{i=1}^r n_{ic} (\tilde{p}_{iu} + \tilde{P}_{iuu} - 2\tilde{p}_{iu}^2)\right) = \pi_u \left[\sum_{i=1}^r n_{ic} - \sum_{i=1}^r n_{ic} (2\phi_i - F_i) \right] \\ \varepsilon(SSG) &= \varepsilon\left(\sum_{i=1}^r n_{ic} (\tilde{p}_{iu} - \tilde{P}_{iuu})\right) = \pi_u \left[\sum_{i=1}^r n_{ic} - \sum_{i=1}^r n_{ic} F_i \right],\end{aligned}$$

suggesting that π_u can be estimated as

$$\hat{\pi}_u = \frac{SS_u}{2(1 - \theta_A) \sum_{i=1}^r n_{ic}}, \text{ where } SS_u = SSP + SSI + SSG, \theta_A = \frac{\sum_{\substack{i,i'=1 \\ i \neq i'}}^r n_i n_{i'} \theta_{ii'}}{\sum_{\substack{i,i'=1 \\ i \neq i'}}^r n_i n_{i'}}.$$

Therefore, from the expectations

$$\begin{aligned}\varepsilon\left(\sum_{u=1}^m \tilde{p}_{iu} (1 - \tilde{p}_{iu})\right) &= \left(\sum_{u=1}^m \pi_u\right) (1 - \phi_i) \\ \varepsilon\left(\sum_{u=1}^m (\tilde{p}_{iu} - \tilde{P}_{iuu})\right) &= \left(\sum_{u=1}^m \pi_u\right) (1 - F_i) \\ \varepsilon\left(\sum_{u=1}^m [\tilde{p}_{iu} (1 - \tilde{p}_{i'u}) + \tilde{p}_{i'u} (1 - \tilde{p}_{iu})]\right) &= 2\left(\sum_{u=1}^m \pi_u\right) (1 - \theta_{ii'}).\end{aligned}$$

A moment estimate of ϕ_i for independent populations is

$$\hat{\phi}_i = 1 - \frac{2(1 - \theta_A) \left(\sum_{i=1}^r n_{ic}\right) \sum_{u=1}^m \tilde{p}_{iu} (1 - \tilde{p}_{iu})}{\sum_{u=1}^m SS_u}.$$

A moment estimate of F_i for independent populations is

$$\hat{F}_i = 1 - \frac{2(1 - \theta_A) \left(\sum_{i=1}^r n_{ic}\right) \sum_{u=1}^m (\tilde{p}_{iu} - \tilde{P}_{iuu})}{\sum_{u=1}^m SS_u}.$$

An estimate of $\theta_{ii'}$ is given by

$$\hat{\theta}_{ii'} = 1 - \frac{(1 - \theta_A) \left(\sum_{i=1}^r n_{ic} \right) \sum_{u=1}^m [\tilde{p}_{iu} (1 - \tilde{p}_{i'u}) + \tilde{p}_{i'u} (1 - \tilde{p}_{iu})]}{\sum_{u=1}^m SS_u}.$$

Define $\alpha_i = \frac{F_i - \theta_A}{1 - \theta_A}$, $T_i = \frac{\phi_i - \theta_A}{1 - \theta_A}$, $\beta_i = \frac{\theta_i - \theta_A}{1 - \theta_A}$, $\beta_{ii'} = \frac{\theta_{ii'} - \theta_A}{1 - \theta_A}$, $f_i = \frac{F_i - \theta_i}{1 - \theta_i} = \frac{\alpha_i - \beta_i}{1 - \beta_i}$,

then θ_A is not involved in the estimates of a_i , β_i , $\beta_{ii'}$, f_i , and T_i :

$$\hat{a}_i = 1 - \frac{2 \left(\sum_{i=1}^r n_{ic} \right) \sum_{u=1}^m (\tilde{p}_{iu} - \tilde{P}_{iuu})}{\sum_{u=1}^m SS_u}$$

$$\hat{T}_i = 1 - \frac{2 \left(\sum_{i=1}^r n_{ic} \right) \sum_{u=1}^m \tilde{p}_{iu} (1 - \tilde{p}_{iu})}{\sum_{u=1}^m SS_u}$$

$$\hat{\beta}_{ii'} = 1 - \frac{\left(\sum_{i=1}^r n_{ic} \right) \sum_{u=1}^m [\tilde{p}_{iu} (1 - \tilde{p}_{i'u}) + \tilde{p}_{i'u} (1 - \tilde{p}_{iu})]}{\sum_{u=1}^m SS_u}.$$

From the definition $\phi_i = \theta_i + \frac{1}{2n_i} (1 + F_i - 2\theta_i) \Rightarrow \beta_i = \frac{n_i}{n_i - 1} T_i - \frac{a_i}{2(n_i - 1)} - \frac{1}{2(n_i - 1)}$,

then $\hat{\beta}_i = 1 - \frac{2 \left(\sum_{i=1}^r n_{ic} \right) \sum_{u=1}^m \left(\frac{n_i}{n_i - 1} \tilde{p}_{iu} (1 - \tilde{p}_{iu}) - \frac{1}{2(n_i - 1)} (\tilde{p}_{iu} - \tilde{P}_{iuu}) \right)}{\sum_{u=1}^m SS_u}$.

These estimates can be simplified by defining

$$S_1 = \sum_{u=1}^m SS_u$$

$$S_{2i} = 2n_c \sum_{u=1}^m \left(\frac{n_i}{n_i - 1} \tilde{p}_{iu} (1 - \tilde{p}_{iu}) - \frac{1}{2(n_i - 1)} (\tilde{p}_{iu} - \tilde{P}_{iuu}) \right)$$

$$S_{3i} = 2n_c \sum_{u=1}^m (\tilde{p}_{iu} - \tilde{P}_{iuu}),$$

then equations for estimating population specific F-statistics are

$$\hat{a}_i = 1 - \frac{S_{3i}}{S_1}$$

$$\hat{\beta}_i = 1 - \frac{S_{2i}}{S_1}$$

$$\hat{f}_i = \frac{S_{2i} - S_{3i}}{S_{2i}}.$$

The weighed average can be estimated as

$$\hat{a}_w = \frac{\sum_{i=1}^r n_i a_i}{n.} = 1 - \frac{\sum_{i=1}^r n_i S_{3i}}{(n.)S_1}$$

$$\hat{\beta}_w = \frac{\sum_{i=1}^r n_i \beta_i}{n.} = 1 - \frac{\sum_{i=1}^r n_i S_{2i}}{(n.)S_1}$$

$$\hat{f}_w = \frac{\sum_{i=1}^r n_i f_i}{n.} = \frac{\sum_{i=1}^r n_i S_{2i} - \sum_{i=1}^r n_i S_{3i}}{\sum_{i=1}^r n_i S_{2i}}.$$

For equal sample sizes these equations reduce to the estimators given by Weir and Cockerham (Weir and Cockerham 1984).

APPENDIX B: List of frequency-based distances

Let p_{ij} and q_{ij} be the frequencies of i th allele at the j th locus in populations X and Y respectively, while a_j is the number of alleles at the j th locus, and m is the number of loci examined.

Geometric distances are not negative, symmetric and satisfy the triangle inequality. The most common distance is the Euclidean distance, defined as:

$$D_{EU} = \frac{1}{m} \sum_j \sqrt{\sum_i^{a_j} (p_{ij} - q_{ij})^2}.$$

Rogers's (1972) distance is a scaled Euclidian distance:

$$D_R = \frac{1}{m} \sum_j \sqrt{\frac{1}{2} \sum_i^{a_j} (p_{ij} - q_{ij})^2}.$$

Prevosti *et al.*'s (1975) distance has statistical properties similar to those of D_R and is defined as:

$$C_p = \frac{1}{2m} \sum_j \sum_i^{a_j} |p_{ij} - q_{ij}|.$$

Cavalli-Sforza and Edwards' (1967) distance gives the chord distance between the two populations if we represent two populations on the surface of a multidimensional hypersphere using allele frequencies at the j th locus:

$$D_C = \frac{2}{\pi m} \sum_{j=1}^m \sqrt{2(1 - \sum_{i=1}^{a_j} \sqrt{p_{ij}q_{ij}})}.$$

Bhattacharyya (1946) and Nei (1987) recommended that the distance between the two populations be measured by

$$\theta^2 = \frac{1}{m} \sum_{j=1}^m (\arccos \sum_{i=1}^{a_j} \sqrt{p_{ij}q_{ij}})^2.$$

The Sanghvi distance (1953) was derived from chi-square goodness-of-fit statistics, and the distance is defined as:

$$X^2 = \frac{2}{m} \sum_{j=1}^m \sum_{i=1}^{a_j} \frac{(p_{ij} - q_{ij})^2}{(p_{ij} + q_{ij})}.$$

Nei *et al.*'s (1983) D_A distance:

$$D_A = 1 - \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^{a_j} \sqrt{p_{ij}q_{ij}}.$$

None of the geometric distances described above involve any evolutionary models. Assuming that there is no mutation, and that all gene frequency changes are by genetic drift alone, the following two quantities are expected to rise linearly with amount of genetic drift.

Cavalli-Sforza's chord distance (1969) is given by:

$$f_v = 2 \sqrt{\frac{\sum_{j=1}^m \left(1 - \sum_{i=1}^{a_j} \sqrt{(p_{ij} - q_{ij})^2} \right)}{\sum_{j=1}^m (a_j - 1)}}.$$

Reynolds, Weir, and Cockerham's (1983) genetic distance (ignoring the terms involving sample size n) is:

$$\theta_w = \frac{\sum_{j=1}^m \sum_{i=1}^{a_j} (p_{ij} - q_{ij})^2}{2 \sum_{j=1}^m (1 - \sum_{i=1}^{a_j} p_{ij} q_{ij})}.$$

Nei's (1972) standard distance has an expected value linearly related to the time since divergence, assuming that all loci have the same rate of neutral mutation, and that the genetic variation is maintained by the equilibrium between infinite-alleles mutation and genetic draft, with the effective population size of each population remaining constant. The quantity is defined as:

$$D_s = -\ln(J_{XY} / \sqrt{J_X J_Y}),$$

where $J_X = \sum_{j=1}^m \sum_{i=1}^{a_j} p_{ij}^2 / m$, $J_Y = \sum_{j=1}^m \sum_{i=1}^{a_j} q_{ij}^2 / m$, $J_{XY} = \sum_{j=1}^m \sum_{i=1}^{a_j} p_{ij} q_{ij} / m$.

Nei's (1973) minimum genetic distance (D_m), Latter's (1972) ϕ^* distance, and Latter's (1973) D_L distance are all defined similarly:

$$D_m = (J_X + J_Y) / 2 - J_{XY}$$

$$\phi^* = \frac{(J_X + J_Y) - J_{XY}}{1 - J_{XY}}$$

$$D_L = -\ln(1 - \phi^*).$$

With the stepwise mutation model (SMM) assumption, Goldstein et al. (1995) proposed that the following distance be used for microsatellite loci:

$$(\delta\mu)^2 = \frac{1}{m} \sum_{j=1}^m (\mu_{X_j} - \mu_{Y_j})^2,$$

where $\mu_{X_j} (= \sum_k k p_{kj})$ and $\mu_{Y_j} (= \sum_k k q_{kj})$ are the average numbers of repeats found, and p_{kj} and q_{kj} are the frequencies of the allele with k repeats at the j th locus in population X and population Y , respectively.

A distance measure closely related to $(\delta\mu)^2$ is the average square distance (ASD, Slatkin 1995), which is given by

$$ASD = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} (u - v)^2 p_{uj} q_{vj}.$$

Another related distance measure is Shriver et al.'s (1995) distance, defined as

$$D_{SW} = W_{XY} - (W_X + W_Y) / 2,$$

where

$$W_X = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} |u-v| p_{uj} p_{vj}, W_Y = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} |u-v| q_{uj} q_{vj}, W_{XY} = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} |u-v| p_{uj} q_{vj}$$

Another commonly used distance, the shared allele distance D_{SA} (Chakraborty and Jin, 1993), is defined as:

$$D_{SA} = \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^{a_j} \min(p_{ij}, q_{ij}).$$

The measure $D_{LS} = -\ln(1 - D_{SA})$ (usually referred as log shared allele distance) has also been proposed.

REFERENCES

- Botstein D, White RL, Skolnick M and Davis RW, Construction of a genetic linkage map in man using restriction fragment length polymorphisms, *American Journal of Human Genetics*, **32**: 314-331 (1980).
- Bhattacharyya A, On a measure of divergence between two multinomial populations. *Sankhya* 7: 401-407 (1946).
- Cavalli-Sforza LL, Human diversity. Proc. 12th Intl Cong. Genet., Tokyo 3: 405-416 (1969).
- Cavalli-Sforza LL and Edwards AWF, Phylogenetic analysis: models and estimation procedures, *American Journal of Human Genetics*, **19**: 233-257 (1967).
- Devlin B and Risch N, A comparison of linkage disequilibrium measures for fine-scale mapping, *Genomics*, **29**: 311-322 (1995).
- Excoffier L and Slatkin M, Maximum-Likelihood estimation of molecular haplotype frequencies in a diploid population, *Molecular Biology and Evolution*, **12**: 921-927 (1995).
- Felsenstein J, Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783-791.
- Felsenstein J, PHYLIP (phylogeny Inference Package), version 3.5c. Depart of Genetics, University of Washington, Seattle.
- Goldstein DB and Ruiz Linares A, Cavalli-Sforza LL and Feldman MM, Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* **92**: 6723-6727 (1995).
- Guo SW and Thompson EA, Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics*, **48**: 361-372 (1992).
- Hudson RR, Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* **23**: 183-201 (1983).
- Hudson RR and Kaplan N, Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147-164 (1990).
- Latter BDH, Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. *Genetics*, **70**: 475-490 (1972).
- Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**:209-220 (1967).
- Nei M, Genetic distance between populations, *American Naturalist*, **106**: 283-292 (1972).
- Nei M, The theory and estimation of genetic distance, p45-54 in Genetic Structure of Populations, edited by Morton NE, University Press of Hawaii, Honolulu (1973).
- Nei M, Molecular Evolutionary Genetics. Columbia University press, New York (1987).
- Nei M and Takezaki N, Estimation of genetic distances and phylogenetic trees from DNA analysis. Proc. 5th World Cong. Genet. Appl. Livstock Prod. 21: 405-412 (1983).
- Nielsen DM and Weir BS, A classical setting for associations between markers and loci affecting quantitative traits, *Genetic Research*, **74**: 271-277 (1999).
- Prevosti A, Ocana J and Alonzo G, Distances between populations for *Drosophila Subobscura* based on chromosome arrangement frequencies. *Theo. Appl. Genet.* **45**: 231-241 (1975).
- Raymond M and Rousset F, An exact test for population differentiation. *Evolution*, **49**: 1280-1283 (1995).

- Reynolds J, Weir BS and Cockerham CC, Estimation of the Coancestry coefficient: basic for a short-term genetic distance. *Genetics* **105**: 767-779 (1983).
- Rohde K, Fuerst R, Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information, *Human Mutation* **17**: 289-295 (2001).
- Rogers JS, Measures of genetic similarity and genetic distance, pp. 145-153 in *Studies in Genetics VII*. University of Texas Publication 7213, Austin, TX (1972).
- Sanghvi LD, Comparison of genetical and morphological methods for a study of biological differences. *Amer. J. Phys. Anthropol.* **11**: 385-404 (1953).
- Shriver M, Jin L, Boerwinkle E, Ferrell R et al., A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol.* **12**: 914-920 (1995).
- Slager SL and Schaid DJ, Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. *American Journal of Human Genetics*, **68**: 1457-1462 (2001).
- Slatkin M, A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457-462 (1995).
- Weir BS, Genetic data analysis II, Sunderland, MA: Sinauer Associates, Inc (1996).
- Weir BS and Hill WG, Estimating F-Statistics, *Annu. Rev. Genet.* **36**: 721-750 (2002).
- Wiuf Carsten and Posada David, A coalescent model of recombination hotspots. To be appear in *Genetics* (2003)
- Posada D, Wiuf C. Simulating haplotype blocks in the human genome. *Bioinformatics.* **19**(2):289-90 (2003).
- Wright, S. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**: 395-420 (1965).
- Zaykin D, Zhivotovsky L and Weir BS, Exact tests for association between alleles at arbitrary numbers of loci, *Genetica*, **96**: 169-178 (1995).
- Zaykin DV, Westfall PH, Young SS, Karnoub MC, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity* **53**:79-91 (2002).
- Zaykin D, Ehm MG and Weir BS. The composite haplotype association mapping of complex traits in out-bred populations. Accepted for publication in Genetic Epidemiology.