Mendel 16.0 Documentation

Kenneth Lange
Rita Cantor
Steve Horvath
Jeanette C Papp
Chiara Sabatti
Janet S Sinsheimer
Hua Zhou
Eric M Sobel

Department of Human Genetics
UCLA School of Medicine
Los Angeles, CA 90095-7088
e-mail: mendel@genetics.ucla.edu

The production of Mendel was supported by
USA NIH grants GM053275, HG006139, MH059490, and RR025521.

August 21, 2016

Contents

0 Introduction

0.1 Overview of Mendel

Mendel is a comprehensive package for the statistical analysis of qualitative and quantitative genetic traits. On pedigree data, it internally incorporates both the Elston-Stewart [29, 63] and the Lander-Green-Kruglyak [56, 57] algorithms. In some applications, it will choose pedigree by pedigree whichever algorithm is faster. Mendel is coordinated with the companion program SimWalk, which performs many of the same tasks by stochastic sampling. Mendel also incorporates an enhanced version of the variance component program Fisher [61] for QTL (quantitative trait loci) mapping and classical biometric genetics. The Macintosh and Windows versions of Mendel include simple front-end applications, "Mendel Launcher", that permit running Mendel without resorting to a command line interface. There is also a more comprehensive web-based interface called Mendel Enterprise that includes database tools for phenotype and genotype data. To get an idea of the scope of applications available through Mendel, glance at the analysis options listed in Table 0.1. These options include many of the common tasks of genetic epidemiology and population genetics. Although Mendel is oriented to pedigrees, several analysis options employ case/control data or random samples of individuals.

This documentation is best understood by consulting the sample input and output files. Each example illustrates some facet of Mendel. To use Mendel, you will need to master the simple input formats of the definition, map, and pedigree files and the keyword strategy of the control file. All sample input files have the extension .in, and all sample output files the extension .out. The files connected with a particular example share a common suffix just before the extension. Thus, the second example of Analysis Option 9 has definition file Def9b.in and standard output file Mendel9b.out. Table 32.1 at the end of this document defines all control file keywords and lists their default values. In interpreting results, background reading in statistical genetics can only help. The books [17, 23, 59, 83, 86, 92, 110] are a good place to start.

0.2 New Features

0.2.1 New Features in Version 16

In version 16 of Mendel, the expiration date has been removed. Also, two new analysis options have been added, both quantitative versions of the Maternal-Fetal Genotype (MFG) incompatibility test. These are QMFG LRT (Analysis Option 30) that performs a Likelihood Ratio Test (LRT) procedure, and QMFG Score (Analysis Option 31) that uses the Score test instead of the LRT. The score test algorithm is much more efficient than the LRT,

Table 0.1: Mendel's Analysis Options

#	Analysis Option	#	Analysis Option
1	MAPPING_MARKERS	17	GENE_DROPPING
2	LOCATION_SCORES	18	COMBINING_LOCI
3	PEDIGREE_HAPLOTYPING	19	VARIANCE_COMPONENTS
4	NPL	20	QTL_ASSOCIATION
5	MISTYPING	21	TRIM_PEDIGREES
6	ALLELE_FREQUENCIES	22	ASSOCIATION_GIVEN_LINKAGE
7	RISK_PREDICTION	23	SNP_IMPUTATION
8	GAMETE_COMPETITION	24	GWAS (SNP_ASSOCIATION)
9	PEDIGREE_SELECTION	25	FILE_CONVERSION
10	KINSHIP	26	MFG
11	GENETIC_EQUILIBRIUM	27	INBRED_STRAINS
12	ASSOCIATION_BY_PERMUTATION	28	SIMULATE_TRAITS
13	TDT	29	PED_GWAS
14	PENETRANCES	30	QMFG_LRT
15	ETHNIC_ADMIXTURE	31	QMFG_SCORE
16	COMBINING_ALLELES		

while the LRT is slightly more accurate. The QMFG Score Option can easily handle dense genome-wide SNP data sets. (There was no public version 15 of Mendel.)

0.2.2 New Features in Version 14

The main change in version 14 of Mendel is faster computation in many of the analysis options, dramatically faster for GWAS (Analysis Option 24), Ped-GWAS (Analysis Option 29), and Kinship Estimation (Analysis Option 10). This was predominantly obtained through parallelization. If desired, the user can limit or turn off this parallelization by setting the new keyword `MAX_THREADS` to the number of processors Mendel should use. The default is to use all available processors.

In addition to making Ped-GWAS and Kinship Estimation much faster, they are also now much easier to use. For example, simply setting the value of the keyword `KINSHIP_SOURCE` in the control file determines if the global kinship coefficients are estimated via (1) only the explicit pedigree structures in the input file, (2) by comparing the SNP genotypes of pairs of individuals within pedigrees, or (3) comparing the SNP genotypes of all pairs of individu-

als in the data set. (The default value is to use the SNP genotypes to compare individuals within the given pedigrees.) If the kinship coefficients are set to be derived from SNP genotypes, the number of SNPs used can be set via the keyword `SNP_SAMPLING_INCREMENT`. The default value of 5 implies 20% of the SNPs are used; to use 100% of the SNPS, set this keyword to the value 1. The user can also trivially set overall minimum numbers of SNPS to use. Finally, the user can use the new keyword `KINSHIP_METHOD` to choose between the Method of Moments (MoM) and the Genetic Relationship Matrix (GRM) algorithms for converting SNP genotypes to global kinship coefficients. On a standard laptop computer the association analysis of 1 million SNPs and 1000 related individuals often takes less than 90 seconds and requires only about 1 GB of RAM.

Many more minor additions are also included. For example, the new keywords `MIN_MAF` and `MAX_MAF` allow one to easily set bounds on the minor allele frequencies of the SNPs allowed to be included in the analysis. Also, Pedigree Trimming (Option 21) has a new sub-option that simply combines all individuals into one pedigree, keeping all stated relationships intact.

The 14.2 revision of Mendel significantly simplifies and decreases the computation time for Ped-GWAS (Analysis Option 29) and Kinship Estimation (Analysis Option 10). For the Ped-GWAS Option, when you list several quantitative traits, each is analyzed separately, unless the new keyword `MULTIVARIATE_ANALYSIS` is set to true. This makes performing many univariate analyses much faster, since the data needs be initialized and the kinship coefficients estimated only once. Also, a new shorthand "all traits" can be used if all variables should be analyzed or a particular predictor should be used for all traits.

In the 14.4 version, the default penalty function used in penalized regression in GWAS (Analysis Option 24) has been changed from the classic lasso to the minimax concave penalty (MCP). The MCP escapes the over shrinkage of the lasso, and thus should improve model selection by better avoiding false positives. To force Mendel to use the lasso penalty instead of MCP, use the new command `LASSO_PENALTY = True` in the Control file. In the 14.5 minor revision of Mendel, the Plink format import is improved.

0.2.3 New Features in Version 13

This new version has significantly expanded several existing options and adds one entirely new analysis option. The new Analysis Option 29: Ped-GWAS provides extremely fast GWAS analysis on extended pedigrees, nuclear families, unrelated individuals, or any combination thereof. It can handle univariate or multivariate quantitative traits with missing data, and allows for covariate adjustment, including correction for population stratification. Pedigree kinships can either be explicitly provided or automatically estimated from dense markers. SNPs can be analyzed under additive (codominant), dominant, or recessive models.

The Ethnic Admixture analysis, Option 15, has been expanded to rank markers by their informativeness to distinguish subpopulations. This option can now also perform very efficient principal component analysis (PCA). These two new features can be used with either text-based or binary data files.

The Kinship analysis, Option 10, has also been expanded to binary files. It can now quickly estimate SNP-based global and local kinship coefficients on dense genome-wide data. It can also use the SNP-based global estimates to cluster individuals and thus construct pedigree groupings without pre-existing relationship information.

The input data formats are now more flexible. For example, Plink binary data sets (.fam, .bin, and .bed files) are now accepted (see Section 0.6.1.2). The SNP definition file can now include allele names. Analogous to the SNP subset file, there is now a sample subset file to easily change which subset of individuals should be included in an analysis. Also, pedigree files without the twin field or without any pedigree fields can be read directly. These options are set using the new keyword `INPUT_FORMAT`.

Many smaller improvements were also implemented. For example, missing SNP genotypes are now replaced with random "average" genotypes when missing data is not allowed in an analysis. Each selection of a random genotype is based on the allele frequencies within the data set. The GWAS output has been expanded and an issue with the Q-Q plot data is fixed. Also in the GWAS analysis, Option 24, one can now set specific values for the penalized regression tuning constants using the command `TUNING_CONSTANT`, if one wishes to override Mendel's automatic adjustments.

0.2.4 New Features in Version 12

In this update we have added one new option, Trait Simulation (see Option 28), and improved several of the existing options. The new option uses either existing genotype data or the output of the genotype simulation option (Option 17) to determine trait values. The traits can be either univariate (simulated via a generalized linear model) or multivariate (simulated via a variance component model). Major locus contributions can be included in the model as mean effects. The list of supported distributional families for the mean effects model is substantial. See Tables 14.1 and 14.2.

Genotype simulation now allows various output styles. In addition, one has considerable control over the degree of missing data for both trait and genotype simulation. This flexibility is accomplished with the introduction of four new keywords: `GENE_DROP_OUTPUT`, `KEEP_FOUNDER_GENOTYPES`, `MISSING_AT_RANDOM`, and `MISSING_DATA_PATTERN`.

Since a grand mean (intercept) is required for each trait in all Mendel's regression analyses, it is now added automatically. Thus, it is no longer necessary to add to your model definition any commands such as `PREDICTOR = GRAND :: trait`. It does no harm if you continue to include such commands explicitly.

An important change has been made to the meaning of the weight that can be assigned to each SNP in the SNP definition file. These predictor weights are used when building the best regression model that includes the user-specified number of predictors. A weight can now be any positive real number. Also, the greater the weight, the more likely the predictor will be included in the model. If the keyword UNIFORM_WEIGHTS has the value True, which is the default, then all predictors with unassigned weights receive weight 1.0. Otherwise, SNPs with unassigned weights receive weight $1/\sqrt{4q(1-q)}$, where q is the minor allele frequency at the SNP. (If you previously included weights in your SNP definition files, we suggest you now use the reciprocal of the old weights.)

Other minor changes include greater flexibility in defining superloci. The maximum number of loci that can be combined into a superlocus has been increased from four to five. The keyword SEED can now equal the special value TIME, which sets the seed based on the current time. The value set is reported in the standard output file.

0.2.5 New Features in Version 11

This revision incorporates two new options and several improvements to existing options. The new options are Maternal-Fetal Genotype Incompatibility Testing (see Section 26) and Inbred Strains Analysis (see Section 27). The first of the new options facilitates modeling of interactions between maternal and child genotypes. The classic example is Rh maternal-fetal incompatibility. The second of the new options represents a new approach to QTL mapping with inbred strains of mice and other animals. The option implements a mixed effects model that correctly captures polygenic background, handles multivariate traits, and copes with pedigrees of arbitrary complexity. The QTL effect is modeled at the mean level as a vector of regression coefficients on the strain origin pair (maternal-paternal) imputed for each animal at the current QTL location along the genome.

To better deal with rare SNPs, the GWAS option (see Section 24) has been significantly improved. In an expanded version of the SNP Definition file, for each SNP one can now assign a weight and membership in a group of SNPs. Default weights can be based on allele frequencies. Penalized regression can now use group penalties, which make it easier to understand association at the gene or molecular pathway level rather than the SNP level. The new keyword PREDICTOR_PENALTY_PROPORTION specifies the relative importance the model puts on individual predictors versus predictor groups. One may also specify individual predictors, or groups of predictors, to always retain in the penalized regression model by using the keywords RETAINED_PREDICTOR and RETAINED_GROUP. Finally, in interaction testing, it is now possible to test for all pairwise interactions in a large base set of predictors.

Among the modification of existing features, we have introduced a better facility for imposing linear constraints on parameters. This is discussed in Section 0.10.2, where

the new keyword PARAMETER_EQUATION is introduced. The deprecated, previous keywords PARAMETER_FIXED_VALUE and PARAMETER_EQUATE will continue to work for now.

The SNP Imputation option (see Section 23) now performs both pedigree based and linkage disequilibrium based genotype and haplotype inference. The combination of these two approaches leads to more accurate imputation results. Better imputation in turn leads to better handling of missing data and more sensitive SNP association tests.

0.3 Random Quotes

Reading program documentation is a necessary evil. As an inducement to dip into this manual more frequently, we have filled any extra space between sections with miscellaneous quotations. These offerings are meant solely for your enjoyment and bear no relation to the sections to which they are appended.

0.4 Special Terminology

The word “gene” like the word “love” has too many meanings for its own good. For our purposes, the old fashioned definition of gene suffices. Thus, a gene is a physical unit of heredity occupying some locus along a chromosome. Genes come in different varieties called alleles. Genes, or more properly copies of genes, are passed from generation to generation. An allele name serves as a label for a gene. For example at the ABO locus, the genes are labeled A, B, and O. Two genes at the same locus with the same allelic label are said to be identical by state. This is to be distinguished from the genes being identical by descent, in which case they are both copies of some common ancestral gene.

Mendel invests certain other words with special meanings. As much as possible, we have tried to be consistent with ordinary English usage, but readers will want to be particularly aware of the terms “affected”, “proband”, “factor”, “category”, “variable”, and “penetrance”. The term “affected” is used as a noun to mean someone affected by a disease or trait. Pedigrees come to the attention of investigators through key affecteds known as “probands”. Many analyses involve non-genetic “factors” such as smoking behavior or health, by which people can be categorized. Relevant “categories” (or “levels”) for smoking behavior might be “non-smoker”, “one pack per day”, and “two packs per day”; for health the categories might be “good”, “fair”, and “poor”. People are also rated on continuous quantitative scales that we simply call “variables”. The same variable, say blood pressure, may be a primary trait of interest or a predictor of a primary trait such as coronary artery disease. Factors and their possible categories are defined after loci in the definition file. Variables are defined after factors. Finally, “penetrance” is the probabilistic relationship between genotype and phenotype. As elaborated in Section 0.5.7, this definition generalizes the traditional notion of the penetrance of a dominant disease.

The term “genetic equilibrium” is a shorthand for Hardy-Weinberg and linkage equilibrium. In “case/control” studies, associations are sought between genotypes, alleles, or haplotypes on the one hand, and disease status on the other hand. Usually, we will use the term “model loci” to indicate those loci held in common by the map file and the definition file. The model loci are precisely the loci that an analysis option considers, either singly in sequence or simultaneously. Analysis options are numbered for easy reference; an option may have different sub-options called “models” attached to it.

0.5 Standard Input Files

0.5.1 Overview

To make your genetic data available to Mendel, the various input files must be in a format that Mendel can read. This section describes these formats. In brief, Mendel usually requires plain text files where data values are separated by commas, blanks, or tabs. For large-scale SNP genotyping data sets, Mendel has efficient binary file formats discussed in Section 0.6. As opposed to the standard list-directed files, Mendel can also read column formatted files where specific spacing is required. Column formatted files were formerly the default, but we no longer encourage them. They are described in detail in Section 0.7.

A few general comments on list-directed files are appropriate. In these files, the extent of spacing is irrelevant, and extra spaces can be added to make the files easier to read. We encourage the use of commas to separate data values since this allows blanks to indicate missing values. For this reason we often refer to list-directed files as comma-separated, but blanks or tabs may also be used to separate values. Files that use commas to delimit values are often referred to as CSV files (Comma-Separated Values), and common application programs such as Microsoft Excel can save to this format. In any comma-separated Mendel input file, all commas after the last non-blank value may be omitted. The maximum line length in any input file is 256 characters, with the exception of the control file with 2048 characters, and the pedigree file with 262,144 characters. The maximum line length of the pedigree file can be reset by the user. The name of a pedigree, person, sex, twin-set, or allele should be input as a character string of eight or fewer characters. Locus names can be up to 16 characters long. Traits can be either qualitative or quantitative. Quantitative values can be up to 64 characters long. Unless you are using scientific notation, the only non-digits in a quantitative value should be an initial + or –, and a single period or comma indicating the start of the decimal digits. In scientific notation, a quantitative value may also have a trailing D or E and an integer exponent. For example, 4.25E–4 represents 4.25×10^{-4}.

As a prelude to more detailed descriptions of the various formats in the following sections, we provide here a brief overview of the role of each input file.

1. **Control file**. In this file, keywords determine the values of the parameters that control a Mendel run. Here one sets the names of the other input files and chooses the analysis option for the current Mendel run. Although the default name for this file is Control.in, any filename may be specified on Mendel's command line as described at the beginning of Section 0.5.3.

2. **Definition file**. This file names and describes the genetic loci in your data. If qualitative traits (factors) or quantitative variables also appear in the data, they must be defined here as well. Sometimes Mendel can deduce descriptive information from other input files, and you can omit it in the definition file. For example, this is the case with allele names and frequencies. The definition file and the pedigree file must be coordinated so that the locus, factor, and variable definitions occur in the same order as their corresponding phenotypes in the pedigree file. Finally, in the definition file you can specify certain loci to be combined into super-loci for further analysis.

3. **Map file**. This file specifies which genetic loci in the definition file to analyze and the distances between them. For almost all analysis options, only loci common to both the definition and map files are analyzed. These loci are referred to as model loci. The map file must contain at least one locus that is also in the definition file; it may contain loci not in the definition file. In the latter case, loci unique to the map file serve solely to determine the distances between adjacent model loci.

4. **Pedigree file**. This file contains data specific to each individual. It names individuals and their parents and reports each person's sex and phenotypes at all genetic loci, categorical factors, and quantitative traits. Data items must be coordinated in the pedigree file with the corresponding items in the definition file.

5. **Penetrance file**. This optional file specifies how genotypes influence phenotypes at certain loci. Analysis Option 14 helps in the construction of penetrance files. Simple penetrance models can be defined using only the control file without resorting to an explicit penetrance file; see Section 0.5.7.2 for instructions.

If you keep in mind that Mendel analyzes only those loci common to both the map and definition files, then you can create definition and pedigree files containing all of your loci and simply modify the map file to determine what subset of loci is used in a given Mendel run. It is worth emphasizing that the order of the loci in the map file determines the order of the loci in an analysis and may be different from the order shared by the coordinated definition and pedigree files.

Mendel always delivers a standard and summary output file. It may create additional output files depending on the analysis option. The possible output files are:

1. **Standard output file**. This file is created for every Mendel run. It gives a more or less complete report of input data and analysis results, depending on the level of output chosen. All processing errors are indicated here.
2. **Summary file**. This file provides a snapshot of analysis results. Unless a Mendel run has failed, you probably will turn to this file first for results.
3. **Plot file**. This optional output file contains a list of results in a simplified columnar structure to make it easy to graph results.
4. **New definition file**. This optional output file defines new loci, phenotypes, and alleles for use in subsequent analyses.
5. **New pedigree file**. This optional output file defines new pedigrees and phenotypes for use in subsequent analyses.
6. **New penetrance file**. Model 2 of Analysis Option 14 creates penetrance files for use in subsequent analyses.

Sections 0.5.3 through 0.5.7 describe common input file conventions in detail, including the use of LINKAGE formatted pedigree files. Section 0.9 describes the standard output file in depth. The remaining output files are discussed in sections devoted to specific analysis options.

0.5.2 Forbidden and Missing-Value Symbols

Table 0.2: Forbidden Symbols

Named Item	Forbidden Symbols
Files	**!** and operating system restrictions
Alleles, Categories, Pedigrees, Individuals, and Phenotypes	internal-space internal-tab **! , /** \ | &
Allele Separators	space tab **! ,** & + **–** is discouraged as Excel converts this to a date **:** is discouraged as Excel converts this to a time
All other names, including Loci, Factors, & Variables	internal-space internal-tab **! , /** \

Forbidden Symbols The input files name many items, including loci, alleles, phenotypes, factors, categories, variables, pedigrees, and individuals. The specific restrictions on which symbols may appear in which names are listed in Table 0.2. In general, avoid names that include any internal spaces, internal tabs, exclamation points, commas, ampersands, forward slashes, backward slashes, and vertical bars. Part of the prohibition on certain symbols stems from the fact that computers often treats commas, spaces, tabs, and forward slashes as field separators.

Genotypes appear in the control, definition, pedigree, and penetrance files as two alleles separated by an allele separator. The default for this single-character symbol is a slash; either the forward or backward slash is accepted, Mendel treats them interchangeably. A vertical bar "|" is the default ordered allele separator. These defaults can be overridden as explained in Section 0.5.4.1. Restrictions on which symbols can be used as allele separators are listed in Table 0.2. For those who use Microsoft Excel to edit their files, we discourage the use of dash, forward slash, or colon as an allele separator since genotypes using these separators are often automatically converted to dates and times when the files are opened in Excel.

Missing-Value Symbols In any Mendel data file that is comma separated, or column formatted, any missing value can be indicated by a blank. Of course when blanks or tabs are used to delimit data values, a symbol other than blank must be used to represent missing values. Additional missing value symbols are defined in the control file, whose format is described below. For example, a command such as `MISSING_VALUE = 0` instructs Mendel to interpret 0 as a missing value when it appears as the complete entry in a non-quantitative field. The keyword `MISSING_QUANTITATIVE_VALUE` does the same for quantitative fields. The string assigned to either of these keywords can be up to four characters long. When the pedigree file is declared to be in LINKAGE format, the default for `MISSING_VALUE` is "0", the zero character; otherwise, the default is blank. The default for `MISSING_QUANTITATIVE_VALUE` is always "–". Again, for any comma separated data file, blanks are always interpreted as missing values.

0.5.3 The Control File

The control file is a plain text file consisting of keywords drawn from Table 32.1 and their assigned values. The default name of the control file is Control.in. In practice, it is often convenient to rename the control file, for example to reference the data set or analysis option it controls. The command line version of Mendel allows you to specify an alternate name for the control file by using the flag –c. For example, typing "`mendel -c my_values.txt`" will cause Mendel to read all its control parameters from the file my_values.txt.

Each line of a control file has the form:

```
KEYWORD = value   !optional comments
```

For instance, the line

```
MAP_FILE = map.in
```

sets the name of the map file. Notice that the assigned value is often case sensitive, while the keyword itself is always case insensitive. On a Unix-derived operating system, Map.in and map.in can be different files, and Mendel needs to know which to use. For the sake of convenience, the possible values of the keyword `ANALYSIS_OPTION` are case insensitive, as are the generic values of true, false, yes, and no. Any text placed after an exclamation point on an input line of the control file is treated as a comment and ignored. Thus, you can nullify a whole command by inserting an exclamation point at its beginning. Blank spaces are ignored in the control file except for internal blanks in file names and in the value of the `TITLE` keyword. No line may exceed 2048 characters.

Some keywords require two values separated by a double colon. For example, the two commands

```
PREDICTOR = Sex :: Left
COVARIANCE_FACTORS = 1 :: Additive
```

imply that sex should be a linear predictor for the quantitative variable Left and that there should be one factor (or dimension) for the additive covariance component.

The order of commands in the control file is irrelevant, but whenever a keyword is set more than once, the last value prevails. An exception to this rule occurs for the keywords indicated in Table 32.1 to be multi-valued. For instance, the commands

```
QUANTITATIVE_TRAIT = Left
QUANTITATIVE_TRAIT = Right
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
```

in a control file would alert Mendel to analyze a bivariate trait, with names Left and Right, under a model with two distinct covariance classes, additive and environmental.

Here is a complete example control file:

```
DEFINITION_FILE = Def0.in     ! Name of the definition file
MAP_FILE = Map0.in            ! Name of the map file
PEDIGREE_FILE = Ped0.in       ! Name of the pedigree file
OUTPUT_FILE = Mendel0.out     ! Name of the detailed output file
SUMMARY_FILE = Summary0.out   ! Name of the summary output file
ECHO = Yes                    ! Echo input files in detailed output file
ANALYSIS_OPTION = Mistyping   ! Analysis option
MODEL = 1                     ! sub-option of the current option
```

The last command is redundant because model 1 is always the default model.

Table 32.1 gives a list of all keywords together with brief definitions and default values. More detailed definitions appear in later sections documenting the various analysis options. Some keywords take the values true (equivalently yes) or false (equivalently no). A title can be attached to a run by defining the keyword `TITLE` in the control file.

0.5.4 The Definition File

The definition file is a plain text file that describes the genetic loci, qualitative factors, and quantitative variables in your data set. These entries must appear in the same order in the definition file as their corresponding phenotypes appear in the pedigree file. After these entries in the definition file, one may also specify super-locus entries that define which loci should be combined when running Analysis Option 18, Combining_Loci. No line in the definition file should exceed 256 characters.

0.5.4.1 Genetic Loci The first entries in a definition file list all genetic loci in your data set. The following lines are required for each locus:

1. Locus identifier line specifying the
   (a) Locus name
   (b) Chromosome, e.g., 4, Autosome, or X-linked [blank $\Rightarrow$ autosome]
   (c) Number of alleles (integer) [blank or 0 $\Rightarrow$ to be determined by Mendel]
   (d) Number of phenotypes (integer) [blank $\Rightarrow$ 0]
2. If item 1(c) is a positive number, then for each allele include a line specifying the
   (a) Allele name
   (b) Allele population frequency (real number) [blank $\Rightarrow$ to be estimated by Mendel]
3. If item 1(d) is a positive number, then for each phenotype include a line specifying the
   (a) Phenotype name
   (b) Number of genotypes associated with the phenotype (integer), or the word ALL if all genotypes are considered compatible with the phenotype [blank $\Rightarrow$ ALL]
4. If item 3(b) is a positive number, then for each genotype include a line specifying the
   (a) two allele names separated by /, \, |, or a user-designated allele separator

Locus names can be up to 16 characters long. In the output these are often shortened to eight characters. The default allele separator symbol is a slash; either forward or backward slash is accepted, Mendel treats them interchangeably. This default value can be easily overridden by setting the keyword ALLELE_SEPARATOR in the control file. See Table 0.2 for restrictions on which characters can be allele separators. Do not use an allele separator as part of an allele name. For example, if the allele separator is *, then it is confusing to Mendel to use * within the name of an allele. Occasionally, it is helpful to input ordered genotypes in the definition or pedigree files. This is possible using an ordered allele separator such as the vertical bar "|", the default ordered allele separator. Thus, the ordered genotype $a|b$ has maternal allele a and paternal allele b. The default can be overridden via the keyword ORDERED_ALLELE_SEPARATOR in the control file. The same character restrictions apply to both types of allele separators. Processing of the definition file stops at the first blank locus identifier line encountered, so be careful to delete any blank lines that precede the end of your data.

In Section 0.5.4.5 we will explain how to vary the input format to accommodate multiple populations, each with its own set of allele frequencies. The keyword POPULATIONS plays a vital role in alerting Mendel to this feature.

Here is a sample definition file:

```
Egomania, , 2, 2
  a, 0.99
  b, 0.01
 NORMAL
 AFFECTED
Marker1
Marker2
```

This sample definition file contains three loci, a disease locus called Egomania and two marker loci called Marker1 and Marker2. All loci are considered to be autosomal since their chromosome fields are blank. The disease locus has two alleles, called a and b, and two phenotypes. In some problems, for example when mutation is allowed, it is necessary to list at the disease loci the normal or wild type allele first and the disease or affected allele second. In the above definition file, the normal allele has frequency 0.99, and the affected allele has frequency 0.01. In the associated pedigree file, an affected person is denoted by AFFECTED and a normal person by NORMAL. In this definition file, both of these phenotypes are consistent with all three possible genotypes, a/a, a/b, and b/b. Since alleles are not listed for the two marker loci, Mendel will determine the number, names, and frequencies of alleles at these two loci from the genotypes in the pedigree file.

On the other hand, if more information were known, the definition file might look like:

```
Egomania,  , 2, 2
```

```
  a, 0.99
  b, 0.01
 NORMAL,   1
   a/a
 AFFECTED, 2
   a/b
   b/b
Marker1, 12, 3
  213
  217
  219
Marker2, XY, 2
  +, 0.545
  -, 0.455
```

Here the trait has been defined as an autosomal dominant. Affected people have either genotype a/b or genotype b/b; normal people have genotype a/a. Marker1 is on chromosome 12 and has three alleles, 213, 217, and 219. Mendel is still requested to estimate their frequencies from the genotypes at this locus in your data set. Marker2 is in the pseudo-autosomal region of the X and Y chromosomes and has two alleles, + and –, with specified frequencies.

Some classical markers such as ABO and Rh have dominant and recessive alleles. Restricting the genotypes corresponding to each phenotype is both logically correct and computationally efficient. In contrast, most trait loci display incomplete penetrance, with each phenotype compatible with any genotype. Thus, it is usually not appropriate to limit which genotypes are compatible with the phenotypes at a trait locus. Instead, only listing the phenotypes, as in the first definition file above, is usually preferred. In applications such as parametric linkage analysis, one must create a penetrance file to convey the different penetrance probabilities. For more information, see Section 0.5.7 on the penetrance file format and Section 2 on the Location_Scores analysis option.

Unless you specifically use Analysis Option 6 for estimating allele frequencies, Mendel automatically estimates frequencies by recursively counting genes. This EM (expectation-maximization) algorithm even works for genes with dominant and recessive alleles. Gene counting treats everyone as unrelated and may give poor estimates, particularly if your data set is small. Therefore, use it with caution. It is better to list allele frequencies when they are available and reliable for your population. If you know allele names, but not allele frequencies, then you can list allele names but leave allele frequencies blank. See Analysis Option 6 for more information. To be consistent, at each locus list either all allele frequencies or none. Mendel gives a warning if listed allele frequencies for a locus do not sum to 1. When this happens, allele frequencies at that locus are rescaled to sum to 1.

Table 0.3 lists possible entries in the chromosome field and their interpretation by

Table 0.3: Chromosome Labels

Chromosome Label (case insensitive)	**Mendel Interpretation**
missing value	autosomal locus
first character 0-9	autosomal locus
first character A	autosomal locus
first character X – second character Y – second character not Y	 pseudo-autosomal X & Y-linked locus X-linked locus
first character Y – second character X – second character not X	 pseudo-autosomal X & Y-linked locus Y-linked locus; excluded from model loci
first character M	mitochondrial locus; excluded from model loci
“Factor”	factor (categorical variable)
“Variable”	quantitative variable
“Superlocus”	super-locus to be formed by combining other loci

Mendel. The default value for this field is autosomal. A locus is considered X-linked if the first character in its chromosome descriptor is an “X” and the second character is not a “Y”; these characters are not case sensitive. If the first two characters are “XY” or “YX”, again case insensitive, then the locus is assumed to be in the pseudo-autosomal region of the X and Y chromosome and is treated as autosomal. Currently Mendel ignores mitochondrial and Y-linked loci.

It is worth emphasizing that Mendel accepts and performs all analyses with X-linked data. Mendel currently does not analyze Y-linked or mitochondria data. As a rule, do not mix X-linked and autosomal loci in the same analysis. If a phenotype for an X-linked locus has a heterozygous genotype listed as one of its possible genotypes, then that genotype is reserved for females and ignored for males. Homozygous female genotypes are treated as legitimate hemizygous male genotypes.

For instance, the sample definition file Def6b.in shows four phenotypes

```
Color-Blind, X-LINKED, 2, 4
    b,.05
    B,.95
  XXBLIND,1
   b/b
  XXNORMAL,2
```

```
   b/B
   B/B
  XYBLIND,1
   b/b
  XYNORMAL,1
   B/B
```

at a hypothetical X-linked color blindness locus. These clearly distinguish females from males. Mendel would accept an alternative form with only the two phenotypes

```
Color-Blind,X-LINKED,2,2
    b,.05
    B,.95
  BLIND,1
   b/b
  NORMAL,2
   b/B
   B/B
```

instead of the four previous ones and eliminate the b/B heterozygous genotype for normal males.

0.5.4.2 Factors As mentioned earlier, the definition file also defines factors and their permitted categories (or levels). These are listed after all genetic locus entries. The rules for constructing a factor entry are even simpler than for a locus entry:

1. Factor identifier line specifying the
   (a) Factor name
   (b) The word “Factor” (case insensitive)
   (c) Number of categories (integer) [blank or 0 $\Rightarrow$ to be determined by Mendel]
2. If item 1(c) is a positive number, then for each category include a line specifying the
   (a) Category name

Factor names can be up to 16 characters long.

For example, a definition file might include the three factors

```
HEALTH,  FACTOR, 2
 Good
 Poor
Race,    Factor
PROBAND, factor, 1
 Proband
```

defining health, race, and proband states. Mendel is alerted to the fact that HEALTH, Race, and PROBAND are factors by the presence of the word Factor (case insensitive) in the second field. Factors are read in the same order in the pedigree and definition files. The HEALTH factor has two defined categories. Mendel will check that no other categories are used at this factor. This type of data integrity checking is always helpful. However, Mendel allows the user to omit the allowed categories; see for example the Race factor. Mendel determines the number and names of all categories in the Race factor from the data in the pedigree file. The proband factor is listed with one defined category. Non-probands are indicated in the pedigree file by blank entries in the proband field.

0.5.4.3 Quantitative Variables After all locus and factor entries, the definition file lists entries for the quantitative variables in the data set. The rules for constructing a quantitative variable entry are similar to those for a factor entry:

1. Variable identifier line specifying the
   (a) Variable name
   (b) The word "Variable" (case insensitive)
   (c) Number of bounds, either 0, 1, or 2 (integer) [blank $\Rightarrow$ 0]
2. If item 1(c) is 1 or 2, then for each bound include a line specifying the
   (a) Bound-type, either "lower", "upper", "minimum", or "maximum" (case insensitive)
   (b) Bound value (real number)

As with loci and factors, variables are read in the same order in the pedigree and definition files. Variable names can be up to 16 characters long, and real-valued bounds can be up to 64 characters long. Either bound value may be left blank. A blank lower bound is interpreted as $-\infty$ and a blank upper bound as $+\infty$.

An example definition file

```
Marker1
Marker2,x-linked
Smoking,factor,3
 heavy
 light
 none
Health,Factor
AgeAtTest,variable,2
 Lower,10
 Upper,120
BMI,VARIABLE
```

illustrates that loci are listed first, factors second, and variables third. For the first quantitative variable listed, AgeAtTest, each corresponding value in the pedigree file will be checked against the two bounds specified here. Any age that is strictly less than 10 or strictly greater than 120 is flagged in the standard output file and causes Mendel to halt. Since there are no bounds listed for the second variable, BMI, no data integrity checks can be performed by Mendel on the BMI values. As we will see in Section 0.9.2, Mendel normally includes in the standard output several summary statistics for each variable, including minimum and maximum. Inspection of these summary statistics can reveal errors in your data.

In Section 0.5.4.6 below we describe how to subject variables to simple transformations. This is particularly helpful in achieving approximate normality. Some analysis options require normally distributed traits.

0.5.4.4 Super-Loci Finally, after the entries for variables, super-loci are defined. These optional entries are pertinent to Analysis Option 18, Combining_Loci. Unlike the locus, factor, and variable definitions that precede them, the super-locus entries have no corresponding fields in the pedigree file.

The rules for constructing a super-locus entry are again similar to those for a factor entry:

1. Super-locus identifier line specifying the
   (a) Super-locus name
   (b) The word “Superlocus” (case insensitive)
   (c) Number of loci-to-combine, either 2, 3, 4, or 5 (integer)
2. For each locus-to-combine include a line specifying the
   (a) Locus name

Super-locus names can be up to 16 characters long. At most five loci can be combined into a super-locus. The contributing loci must appear earlier in the definition file, and each is limited to nine or fewer alleles. The alleles need not be listed in the definition file. As explained in detail in Section 18, the analysis option Combining_Loci outputs new definition and pedigree files that include the super-loci defined here.

An example definition file,

```
SNP1   , X-LINKED
SNP2   , X-LINKED
SNP3   , AUTOSOME
SNP4   , AUTOSOME
```

```
SNP5   , AUTOSOME
HEALTH, FACTOR,2
 AFFECTED
 NORMAL
BMI    , VARIABLE
S1+S2 , SUPERLOCUS,2
 SNP1
 SNP2
3+4+5 , SUPERLOCUS,3
 SNP3
 SNP4
 SNP5
```

illustrates that super-locus entries must be the last entries. When running Analysis Option 18, Combining_Loci, this definition file leads to two new super-loci. The first combines the two X-linked loci, SNP1 and SNP2, and names the super-locus "S1+S2". The second combines the remaining three autosomal loci, and names this new super-locus "3+4+5". X-linked and autosomal loci cannot be combined to make a super-locus.

0.5.4.5 Population-Specific Allele Frequencies Several analysis options can take into account the ethnicity of individual pedigree founders and use the allele frequencies pertinent to a founder. The commands

```
POPULATIONS = n
POPULATION_FACTOR = COUNTRY
```

in the control file alert Mendel to read n population-specific allele frequencies for each locus in the definition file. Mendel will also expect a factor defining a set of n corresponding populations. Each individual's data in the pedigree file must contain a blank or a legal population in the corresponding field of the pedigree file. As an example, the definition file

```
APOE, 19, 3
2, 0.0319, 0.0394
3, 0.8006, 0.7835
4, 0.1676, 0.1772
COUNTRY, FACTOR, 2
JAPAN
COLOMBIA
```

for the case $n = 2$ defines separate allele frequencies for Japan and Colombia at the APOE locus, with the Japanese frequencies appearing on the left and the Colombian frequencies on the right. These conventions guarantee that the correct frequencies will be

used whenever a founder is encountered in computing a pedigree likelihood. If an individual's population field is left blank in the pedigree file, then he or she is considered to belong to the first named population, in this case Japan. Thus, it is a good idea to make the first population the most likely population.

Table 0.4: Allowed Variable Transformations

Transform Name	**Transformation of Data**
Above_Threshold	Convert values at or above threshold to 1; Convert values below threshold to 0
Below_Threshold	Convert values at or below threshold to 1; Convert values above threshold to 0
Degender	Standardize data to mean 0 and variance 1 separately for females and males
Log	Natural logarithm of data value
Rank	Rank of data value among all values
Indicator	Affecteds assigned 1; Unaffecteds assigned 0
Standardize	Standardize data to mean 0 and variance 1
# (any number)	Data value raised to a power

0.5.4.6 Transforming Variables In many situations it is a good idea to transform a variable. For instance, you might want to (a) standardize a trait so that it has mean 0.0 and variance 1.0, (b) transform a trait to eliminate skewness and achieve approximate normality, or (c) convert a qualitative factor into a quantitative variable. The keyword `TRANSFORM` allows you to do this. The possible transformations are listed in Table 0.4.

As examples of these transformation commands, the first of the two control-file commands

```
TRANSFORM = Standardize :: Cholesterol
TRANSFORM = Degender :: Weight
```

standardizes the variable Cholesterol ignoring sex. The second command standardizes the variable Weight separately for females and males. The commands

```
TRANSFORM = Log :: Triglycerides
TRANSFORM = 0.5 :: Lipid
```

replace the variable Triglycerides by its natural logarithm and Lipid by its square root, i.e., the value raised to the power one-half. Any real number qualifies as a power, so you can take reciprocals by specifying the power –1.0. If you specify a log or power transformation, keep in mind that the values of the variable undergoing transformation must either be absent or a positive number. The Rank transformation replaces each value of a variable by its rank. Tied values are assigned averaged ranks.

Some Mendel analysis options require an indicator variable that has the value 1 for one class of individuals and the value 0 for all other individuals of known status. Mendel has transformations that can convert either quantitative variables or qualitative phenotypes to this type of indicator variable. The transformations Above_Threshold and Below_Threshold act on quantitative variables. The Above_Threshold transformation replaces a value with 1 if the value is greater than or equal to a preset threshold; it replaces values below the threshold with 0 and preserves missing values. The Below_Threshold transformation does the same except the values less than or equal to the threshold are set to 1. For either thresholding transformation, the threshold value is set via the keyword `INDICATOR_THRESHOLD`, which has default value 0. For example, the control file commands

```
TRANSFORM = Above_Threshold :: activities
INDICATOR_THRESHOLD = 5
```

will convert any value of activities ≥ 5 to the value 1; any other non-missing value will be converted to 0.

The Indicator transformation turns qualitative phenotypes into an indicator variable. Specifically, the indicator variable is set equal to 1 for an affected individual and 0 for an unaffected. For example, the commands

```
AFFECTED_LOCUS_OR_FACTOR = Health
AFFECTED = ILL
TRANSFORM = Indicator :: Status
```

transfer the affection status encoded in the locus or factor named Health to the variable named Status. Individuals with the phenotype ILL at the locus or factor named Health will be assigned the value 1 in the variable Status; individuals with all other non-missing phenotypes are assigned the value 0; missing phenotypes lead to missing Status values. The named variable, Status in this example, must exist in the definition and pedigree files. Its pre-transform values are ignored.

The summary statistics listed for a variable in the standard output file always pertain to the transformed variable, not to the original variable. You can transform the same variable several times. The sequence of transformations is dictated by the corresponding sequence of keyword assignments in the control file.

0.5.4.7 Summary Despite its many options, the definition file can be very simple. For example, a complete definition file can be as simple as

```
Marker1
Marker2
Marker3
Marker4
Marker5,x-linked
Marker6,x-linked
Disease,factor
Health,factor
Age,variable
BMI,variable
```

Simplicity comes with a price. In this simple form, Mendel can do little data integrity checking. Also, allele frequencies, which for some analyses need to be accurate, are determined solely from your data. On the other hand, if you have confidence in your data formatting, and a large sample for frequency estimation, then the price is small, and the simplicity elegant.

0.5.5 The Map File

The map file is a plain text file indicating the genetic loci to be analyzed and the distances between them. The loci in common between the map and definition file are called the model loci and are the only loci used in analysis. The map file should contain at least one locus in common with the definition file. The map file should not contain any factors or quantitative variables.

There are two map file types, one listing marker positions and the other listing inter-marker distances. Position map files require physical distances. Inter-marker distance map files require genetic distances or recombination fractions. Your choice of distance units informs Mendel which type of map file to expect.

0.5.5.1 Distance Units The allowed distance units are listed in Table 0.5. Although recombination fractions do not meet the mathematical definition of a distance, it is convenient to include them as a choice. You indicate which units are employed by defining the keyword `MAP_DISTANCE_UNITS` in the control file. Either the full unit name or its abbreviation is an acceptable value for the keyword. These values are not case sensitive. The default value is RF, the abbreviation for recombination fractions.

Internally most of Mendel's analyses use recombination fractions between adjacent loci. When genetic distances are input, Mendel converts these to recombination fractions.

Table 0.5: Map Distance Units

Units (Abbreviation)	Measurement type	Style of map file
Bases (bp)	Physical distance	Positions
Kilobases (Kb)	Physical distance	Positions
Megabases (Mb)	Physical distance	Positions
Recombination-Fractions (RF)	Recombination fraction	Interval distances
Morgans (M)	Genetic distance	Interval distances
centiMorgans (cM)	Genetic distance	Interval distances
Kosambi-Morgans (K-M)	Genetic distance	Interval distances
Kosambi-centiMorgans (K-cM)	Genetic distance	Interval distances

If physical distances are input, Mendel first converts these to genetic distances and then the genetic distances to recombination fractions. As a rough rule of thumb, one megabase equals one centiMorgan. You can change this default ratio by using commands such as `MAP_CONVERSION = 0.8` , or for sex-specific ratios

```
MAP_CONVERSION_FEMALE = 1.5
MAP_CONVERSION_MALE = 0.5
```

in the control file. In the former case, one Mb equals 0.8 cM in either sex; in the latter, one Mb is 1.5 cM in females and 0.5 cM in males. Mendel converts genetic distances into recombination fractions using Haldane's model. Haldane's formulas

$$\begin{aligned} \theta &= \frac{1}{2}\left(1 - e^{-2d}\right) \\ d &= -\frac{1}{2}\ln\left(1 - 2\theta\right) \end{aligned}$$

relate the recombination fraction θ and genetic distance d in Morgans between two loci. If you specify Kosambi-Morgans or Kosambi-centiMorgans, Mendel applies Kosambi's map function in place of Haldane's.

0.5.5.2 Interval Format The map file format specifying inter-marker distances has a simple structure alternating two kinds of lines or records. Odd numbered lines contain locus names. Even numbered lines specify genetic distances or recombination fractions between adjacent loci. For example, the simple map file

```
Marker2
```

```
        5.45
Marker1
        3.23
Marker3
```

using cM sets the sex-average distances between loci Marker2 & Marker1 and Marker1 & Marker3 as 5.45 cM and 3.23 cM, respectively. If sex-specific distance information is known, two distances should be listed, female preceding male. For example

```
Marker2
        5.45, 4.62
Marker1
        3.23, 3.79
Marker3
```

If you leave a female distance blank, then Mendel will equate it to infinity, indicating unlinked loci. If you leave a male distance blank, then Mendel will equate it to the female distance. Leaving both distances blank indicates completely unlinked loci.

Some of Mendel's analysis options permit you to specify a variable number of analysis points between adjacent pairs of loci. This facilitates, for example, the plotting of location score curves in Option 2 and NPL curves in Option 4. The default value for this number of points can be set in the control file using the keyword `INTERIOR_POINTS`. To change the number of points for a specific interval, the new value should be listed in the map file after the male distance for that interval. Here is a sample map file with this feature

```
Marker2
        5.45, , 4
Marker1
        3.23
Marker3
        2.0001, 1.119, 2
Marker5
```

For the first interval, 5.45 cM is specified for both female and male distances, and the number of interior analysis points is set to four. In the second interval, the number of interior analysis points is defined by the current value of the keyword `INTERIOR_POINTS`. The third interval specifies sex-specific distance values and two interior analysis points.

The keyword `GRID_INCREMENT` allows you to override the numbers of internal points mandated for each interval and instead specify an evenly spaced grid of map points for analysis purposes. The value specified by `GRID_INCREMENT` must be in the units specified by the keyword `MAP_DISTANCE_UNITS`. For instance, the commands

```
MAP_DISTANCE_UNITS = cM
GRID_INCREMENT = 1.0
```

in the control file determines a spacing of 1 cM between successive grid points. These uniformly spaced points are generally supplemented by marker positions when Mendel analyzes data.

0.5.5.3 Position Format The map file format listing positions for each locus also has a simple structure. Each line starts with a locus name and follows with a position. All positions on a given chromosome must be a physical distance measured from the same arbitrary point near the start of the chromosome. Only one position value is listed at each locus. If a position value is missing, it is set to infinity, indicating that the locus is unlinked to the previous and subsequent loci. Such a gap in distance information also implies that previous loci and subsequent loci are unlinked. Leaving out all position information is equivalent to treating all loci as unlinked. A sample map file using Mb

```
Marker2, 1.11
Marker1, 6.56
Marker3, 9.79
Marker6
Marker7, 1.13
```

sets the distance between loci Marker2 & Marker1 and Marker1 & Marker3 as 5.45 Mb and 3.23 Mb, respectively. Marker6 and Marker7 are unlinked to each other and to the other loci.

If a position value is less than the value at the preceding locus, then the distance between the two loci is set to infinity, again indicating unlinked loci. A completely blank line also indicates a change in chromosome between two loci. An example map file using base pair information on six loci

```
SNP01,  9023
SNP02,557893
SNP03, 10500
SNP04,200389

SNP05,335298
SNP06,444783
```

shows SNP01 and SNP02 are on one chromosome, SNP03 and SNP04 on another, and SNP05 and SNP06 on a third.

Again one can specify the number of interior analysis points for each interval, just as in the interval-based map file format described earlier. The default value for this number is

set in the control file using the keyword INTERIOR_POINTS. To change the number of points for a specific interval, insert the desired number immediately after the listed position for the locus beginning the interval. For example, consider the map file using Mb position values

```
Marker2, 1.11,4
Marker1, 6.56
Marker3, 9.79,2
Marker5,11.7901
```

The interval between Marker2 and Marker1 has length 5.45 Mb and 4 interior analysis points. In the second interval, the number of interior analysis points is set by the current value of the keyword INTERIOR_POINTS. The third interval, between Marker3 and Marker5, has length 2.0001 Mb and 2 interior analysis points. As mentioned above for interval formatted map files, you can override the numbers of internal points indicated for each interval and specify instead an evenly spaced grid of analysis points by assigning a value to the keyword GRID_INCREMENT in the control file. The value specified by GRID_INCREMENT must be in the same units used in the map file.

0.5.5.4 Model Loci As previously mentioned, the model loci are those common to both the definition and map files. Among the genetic loci, only the model loci are used in Mendel's analysis. One of the benefits of having a separate map file is that it allows you to exercise easy control over the model loci. Suppose your original map file was

```
Marker2
        0.10
Marker0
        0.10
Marker1
```

If Marker0 is not in the definition file, it is ignored in analysis. In this case, Mendel correctly converts the two distances into a single distance separating locus Marker2 from Marker1. If Marker0 were in the definition file, you could achieve the same purpose by changing "Marker0" in the map file to "!Marker0". This no longer corresponds to a locus name in the definition file, so again Marker0 is ignored in analysis, and the correct distance is used between Marker2 and Marker1.

0.5.6 The Pedigree File

The pedigree file is also a simple text file. It names individuals, describes their relationships, and records their phenotypes at all genetic loci, factors, and quantitative variables. Usually names in this file should be eight or fewer characters. Technically they can be

longer, but only the first eight characters will be used, so they must be uniquely determined by the first eight. Genotypes and phenotypes at loci and factors can be up to 10 characters long, and variable values up to 64. The default maximum line length in the pedigree file is 262,144 $= 2^{18}$ characters. This value can be overridden using the keyword `PEDIGREE_MAX_LINE_LEN` in the control file. Comment strings (which recall are initiated by an ! and continue to the end of the line) are allowed on any line after any data values. Completely blank or commented out lines are allowed anywhere in the pedigree file, except within person records that span multiple lines (as explained below).

0.5.6.1 Person Records In the default pedigree format, each line in the pedigree file lists an individual's data. Each line is called a person record and contains the following items, in order:

- Pedigree name
- Individual's name
- Name of parent #1, if in the pedigree
- Name of parent #2, if in the pedigree
- Individual's sex
- Identical twin status
- Phenotype or genotype at each genetic locus
- Category at each factor
- Value at each quantitative variable

Missing values are allowed in any of these fields except pedigree name, individual's name, and sex. Acceptable missing value symbols and certain forbidden symbols are discussed in Section 0.5.2. The individuals within a pedigree should all be grouped together in the file, but their order within the pedigree is arbitrary.

An example pedigree file

```
Bush,   George ,      ,     ,M,,AFFECTED,213\217,1946
Bush,   Laura  ,      ,     ,F,,NORMAL,  213\213,1946
Bush,   Barbara,George,Laura,F,,NORMAL,  213\213,1981
Bush,   Jenna  ,George,Laura,F,,AFFECTED,       ,1981
Clinton,Bill   ,      ,     ,M,,AFFECTED,213\217,1946
Clinton,Hillary,      ,     ,F,,AFFECTED,213\217,1947
Clinton,Chelsea,Hillary,Bill,F,,NORMAL,  213\213,1980
```

illustrates the use of these fields. The spacing in comma-separated files is arbitrary and is arranged here simply for readability. To preserve confidentiality, you will almost always want to substitute coded IDs for names.

Mendel provides a mechanism to read in data sets that do not have all these fields. For example, if the twin field is absent, then use the command `INPUT_FORMAT = No_Twins`. If in addition there is no field containing pedigree names or parental names, then use the command `INPUT_FORMAT = No_pedigrees`. Clearly, this data set can only be valid if all the individuals are unrelated and each has a unique name. In this case, Mendel will set each individual's pedigree name to be the same as their individual name. Finally, if even the sex field is absent, then use the command `INPUT_FORMAT = IDs_Only`. Mendel will again set the pedigree name to be equal to the individual name, and set all sexes to female. Such a data set can only be valid if again all individuals are unrelated and have unique names, and in additional either there are no X-linked loci or everyone is actually female. (There are two other specialized input formats, invoked by assigning either the value LINKAGE or PLINK to the keyword `INPUT_FORMAT`. How to use Linkage and Plink format files is described in more detail in Section 0.5.6.3 and Section 0.6.1.2, respectively.

When they are present, each of the fields of a person record have some necessary attributes. For example, all pedigree names should be unique. A name for an individual can be reused across pedigrees, but never within a pedigree. Part of our technical definition of a pedigree is that an individual has either both parents present in the pedigree or neither. Never include one parent, and exclude the other. The order of the parents is arbitrary. Those people without parents in the pedigree are termed the founders of the pedigree. To reconstruct the relationships between individuals, often people must be included in the pedigree who are dead or otherwise unavailable for study.

Each sex has two default case-insensitive names: 1 and M for male, and 2 and F for female. These defaults can be overridden in the control file. For example, the commands

```
MALE = varon
FEMALE = hembra
```

might be useful for a Spanish data set. These sex designators can be up to eight characters and are always case insensitive.

If there are monozygotic twins in a pedigree, then each pair of identical twins should be assigned a unique non-blank identifier in the twin-status field. Identical triplets are handled in the same way. If there are no monozygotic twins in a pedigree, as in the example above, then the field should be left blank or filled with another missing value symbol.

As mentioned earlier, the definition file and the pedigree file must be coordinated in the sense that the order of the phenotypes for individuals in the pedigree file should exactly match the order of the loci, factors, and variables in the definition file.

The unordered genotypes at the second locus in the above sample pedigree file consist of two codominant alleles separated by a slash. It is possible to input ordered genotypes using an ordered allele separator such as the default vertical bar, |. For example, the ordered genotype 213|217 indicates that the maternal allele of a given individual is 213 and the paternal allele is 217. Alternative allele separators can be defined by appropriate keywords in the control file; see Section 0.5.4.1. For an X-linked locus, only homozygous genotypes should be assigned to males in the pedigree file. Otherwise, Mendel has trouble deciding which allele is pertinent.

Some genotyping outcomes leave one allele in doubt. Mendel accordingly accepts partial genotypes in the pedigree file. For instance, if an autosomal marker has three alleles denoted 1, 2, and 3, then an individual listed with the unordered partial genotype 2/ possesses one of the consistent unordered genotypes 1/2, 2/2, or 2/3. The same can be said for an individual listed with the unordered partial genotype /2. If we use the ordered allele separator instead, then 2| requires the maternal allele to be allele 2, and |2 requires the paternal allele to be allele 2. Partial genotyping also works correctly for males at an X-linked marker because an observed allele is always paired with a second allele of the same type. Simply listing a single allele without an allele separator leads to trouble, so either list male genotypes at an X-linked locus as the corresponding female homozygous genotypes or as a single allele with an allele separator. Thus, the genotype of a male with hemizygous genotype 1 can be listed as 1/1, 1/, /1, 1|, or |1, but never as 1 in isolation.

To cope with a very large number of loci, each person record may occupy up to 64 lines. If a person record continues on the next line, the last non-blank character before any comments on the current line must be an ampersand, &. For example, one of the person records of the above presidential pedigree file could be

```
Bush,   Jenna, &   ! Jenna is a DZ, not an MZ, twin of Barbara
 George,Laura,F&
,,AFFECTED,,1981
```

Recall that an exclamation point "!" signals the start of a comment string that extends to the end of the line. The comment string is completely ignored by Mendel. Since the trailing ampersand indicates that the person record continues on the next line, completely blank or commented out lines are not allowed within a person record that spans multiple lines. Normally, a line can end after the last non-blank value, that is, no "trailing commas" are required. However, a person record that uses continuation lines must include explicit fields for each entry, even missing values at the end of the line.

0.5.6.2 Pedigree Records Mendel also recognizes a modified pedigree format that moves pedigree names to a separate line preceding each pedigree. These pedigree

records list the number of individuals in the pedigree and the pedigree name. Here is the presidential pedigree using pedigree records:

```
4, Bush
George ,      ,     ,M,,AFFECTED,213\217,1946
Laura  ,      ,     ,F,,NORMAL,  213\213,1946
Barbara,George,Laura,F,,NORMAL,  213\213,1981
Jenna  ,George,Laura,F,,AFFECTED,       ,1981
3, Clinton
Bill   ,      ,     ,M,,AFFECTED,213\217,1946
Hillary,      ,     ,F,,AFFECTED,213\217,1947
Chelsea,Hillary,Bill,F,,NORMAL,  213\213,1980
```

Mendel is alerted to this form of the pedigree file by the command

```
READ_PEDIGREE_RECORDS = True
```

in the control file. The default value of READ_PEDIGREE_RECORDS is false in the usual list-directed files; the default is true for the column-specific pedigree files discussed in Section 0.7.4.

With pedigree records, you can omit pedigree names altogether. Mendel then assigns names consistent with each pedigree's order of entry. For example, if the second pedigree is missing a name, it is assigned a name such as PED#02.

In some applications it is convenient to attach a copy or repeat number to each pedigree. With pedigree records, you can place this copy number after the pedigree name. To enable Mendel to read pedigree copy numbers, the command

```
READ_PEDIGREE_COPIES = True
```

must be inserted in the control file. With this command inserted, we could read a copy number of 2 for the Bush family via the amended pedigree record

```
4, Bush, 2
```

in the pedigree file. If the pedigree copy field is left blank, then the pedigree is counted as appearing only once in the data. The loglikelihood of a pedigree is multiplied by its copy number. Thus, a pedigree with copy number 0 is effectively skipped in analysis. As a rule, it is better to avoid this tempting device for ignoring certain pedigrees because it inflates computing times. Positive pedigree copy numbers save you from repeating identical pedigrees in the pedigree file and allows you to assign prior probabilities to pedigrees. See Analysis Option 9 for an example of the latter tactic.

0.5.6.3 Linkage Format Mendel can also directly read LINKAGE (pre-Makeped) format pedigree files. Here is the presidential example in LINKAGE format, with a missing birth year for Chelsea:

```
1       George    0       0     1  2        213 217  1946
1       Laura     0       0     2  1        213 213  1946
1       Barbara George  Laura   2  1        213 213  1981
1       Jenna   George  Laura   2  2          0   0  1981
2       Bill      0       0     1  2        213 217  1946
2       Hillary   0       0     2  2        213 217  1947
2       Chelsea Hillary Bill    2  1        213 213   -
```

In LINKAGE pedigree format, recall that by default male = 1 and female = 2, normal = 1 and affected = 2, and a 0 represents a missing non-quantitative value. For quantitative values, Mendel uses a dash "–" as the default missing value symbol. This convention applies to all pedigree files, not just to LINKAGE files. The various LINKAGE defaults can be overridden in the control file by using the keywords `MALE`, `FEMALE`, `AFFECTED`, `MISSING_VALUE`, and `MISSING_QUANTITATIVE_VALUE`. Identical twins are not recognized in LINKAGE format, so omit twin status. Several of the previously mentioned restrictions also hold for LINKAGE format files. For example, each line contains at most 262,144 characters, unless this value is overridden using the keyword `PEDIGREE_MAX_LINE_LEN` in the control file. Person records can be continued on the following line by setting the last non-blank character to an &. The ! character and everything after it on a line are interpreted as a comment and completely ignored.

In accepting LINKAGE format, Mendel still expects the definition and pedigree files to be coordinated. The data may include at most one trait ("affection status") locus. In LINKAGE format a phenotype at the trait locus is listed as one of the numbers 0, 1, or 2 rather than as a pair of alleles. If there is a trait locus, it may be any of the loci, although the usual convention is to make it the first locus. The above file violates the LINKAGE rule that fathers precede mothers, but Mendel does not mind. To read a LINKAGE formatted pedigree file, insert

```
INPUT_FORMAT = Linkage
```

in the control file. Implicit in this command are the further commands

```
PEDIGREE_LIST_READ = True
READ_PEDIGREE_RECORDS = False
READ_PEDIGREE_COPIES = False
MALE = 1
FEMALE = 2
AFFECTED = 2
MISSING_VALUE = 0
```

The first of these implicit commands tells Mendel to read person records as a sequence of delimited items. As mentioned above, each of the last four of these implicit commands may be overridden by specifying an alternative value in the control file. If there is a trait locus, one must indicate it's position, say as the first locus, by including a command such as

```
AFFECTED_LOCUS_OR_FACTOR = <name of first locus>
```

in the control file.

0.5.6.4 Automatic Pedigree Trimming In many applications, the presence of untyped people erodes computational efficiency without increasing information. Therefore, when appropriate, Mendel automatically trims untyped people unless you tell it otherwise by setting the keyword `PRETRIM_PEDIGREES` to false in the control file. Option 21 allows you to influence the trimming process in explicit ways that are discussed in the documentation of that option. Trimming works by identifying a set of core people who must be retained in analysis. Besides these core people, Mendel retains everyone needed to specify the correct relationships between members of the core. Automatic trimming includes as core everyone who is affected, a proband, has a recorded value for a designated quantitative trait, or has a non-blank phenotype or non-trivial penetrance for at least one model locus. These are the same criteria used by model 4 of Analysis Option 21. Analysis Options 7, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20, 21, 23, 24, 25, 27, 28, 29, and model 2 of Option 14 do not invoke automatic pedigree trimming.

If you are curious about which individuals are being trimmed, then insert the command `KEEP_HIDDEN_FILES = True` in the control file. This will allow you to inspect the intermediate pedigree file that omits the trimmed individuals.

0.5.7 The Penetrance File

0.5.7.1 General Penetrance Models The penetrance file is an optional input file used by a few analysis options such as location scores and genetic risk prediction. Penetrance values specify the probabilistic relationship between hidden genotypes and observed phenotypes at a locus. Consider the Egomania locus featured in the above examples. Under a simple dominant model with a as the normal allele and b as the affected allele, we have

$$\begin{aligned} \Pr(\text{affected} \mid a/a) &= 0 \\ \Pr(\text{affected} \mid a/b) &= 1 \\ \Pr(\text{affected} \mid b/b) &= 1. \end{aligned}$$

Mendel accommodates these either/or conditional probabilities by defining phenotypes with prescribed genotypes in the definition file. One of these phenotypes or a missing data symbol such as blank is then entered for each person in the pedigree file at the Egomania locus. To include the possibilities of incomplete penetrance and phenocopies, we need to modify 0/1 penetrances to reflect our uncertainty. Proper coordination of the definition and penetrance files permits this.

For example, suppose we posit for Egomania a dominant model with the phenocopy and penetrance rates

$$\begin{aligned} \Pr(\text{affected} \mid a/a) &= 0.10 \\ \Pr(\text{affected} \mid a/b) &= 0.95 \\ \Pr(\text{affected} \mid b/b) &= 0.95. \end{aligned}$$

Because any genotype is now compatible with either the affected or normal phenotype, we must use in the definition file an entry for the disease locus such as

```
Egomania, , 2, 2
  a, 0.99
  b, 0.01
 NORMAL
 AFFECTED
```

Since here no genotypes are listed, both phenotypes are deemed compatible with all possible genotypes. The Clinton family portion of the corresponding penetrance file is:

```
 Clinton,    Bill,       Egomania,          a/a,          0.10,   AFFECTED
 Clinton,    Bill,       Egomania,          a/b,          0.95,   AFFECTED
 Clinton,    Bill,       Egomania,          b/b,          0.95,   AFFECTED
 Clinton, Hillary,       Egomania,          a/a,          0.10,   AFFECTED
 Clinton, Hillary,       Egomania,          a/b,          0.95,   AFFECTED
 Clinton, Hillary,       Egomania,          b/b,          0.95,   AFFECTED
 Clinton, Chelsea,       Egomania,          a/a,          0.90,     NORMAL
 Clinton, Chelsea,       Egomania,          a/b,          0.05,     NORMAL
 Clinton, Chelsea,       Egomania,          b/b,          0.05,     NORMAL
```

Note that complementary penetrances are input for Chelsea because she is unaffected. For an actual Mendel run, we obviously need the missing Bush portion of the penetrance file.

Every line of a penetrance file contains at least five ordered items: a pedigree name, a person name, a locus name, a genotype, and a penetrance value. The observed phenotype of the named person at the named locus may be added as the last item on the

line. Only the first five values on each line are read by Mendel. Phenotypes are optional because they already known to Mendel from the pedigree file.

The design of the penetrance file trades parsimony for flexibility. Older versions of Mendel relied on a much simpler penetrance file interpreted in a very different fashion. Unfortunately, the earlier conventions limited penetrances to loci with two alleles and to conditional probabilities as just described or to Gaussian densities. The new version lifts the limitation to biallelic loci and encourages broader thinking about penetrances. Two important rules still hold, however. First, multilocus penetrances are always multiplicative. In other words, phenotypes at the different model loci are conditionally independent given genotypes. Second, any genotypic restrictions imposed on a qualitative phenotype in the definition file limit what genotypes are visited in likelihood computations for a person with that phenotype. For this reason we stressed amending the definition file at the Egomania locus. Mendel's convention on genotype restriction works to your advantage in constructing a penetrance file because forbidden genotypes need not appear in the file. Finally, regardless of the default penetrance value, Mendel assigns a penetrance of 1.0 to each legal genotype of someone missing both a penetrance in the penetrance file and a phenotype at the corresponding locus.

Application of model 2 of Option 14 eases the pain in creating penetrance files. Please see the instructions in Section 14 dealing with penetrance estimation and penetrance file creation under generalized linear models. The keyword `NEW_PENETRANCE_FILE` is introduced there.

In order to keep penetrance files small, Mendel presets all penetrances to 1. You can change this default penetrance to another value, say 0, by inserting the command

```
DEFAULT_PENETRANCE = 0.0
```

in the control file. Note that changing the default penetrance does not change the fact that genotypes inconsistent with a stated phenotype in the definition file are excluded in likelihood computation. The excluded genotypes are permanently assigned a penetrance of 0.

0.5.7.2 Simple Penetrance Models In some simple cases, it is possible to omit the penetrance file and list penetrances in the control file. This is the case when the penetrance function is uniform across all individuals. As a concrete example, the commands

```
AFFECTED = AFFECTED
AFFECTED_LOCUS_OR_FACTOR = Egomania
PENETRANCE = 0.10 :: a/a
PENETRANCE = 0.95 :: a/b
PENETRANCE = 0.95 :: b/b
```

in the control file set the penetrances of the AFFECTED phenotype at the Egomania locus for each of the three genotypes. The penetrances of any non-blank, unaffected phenotype is derived by subtraction of these values from 1.0. For example, the penetrance of the NORMAL phenotype for genotype a/a is 0.9. Anyone with a blank or missing phenotype has penetrance 1.0 for all genotypes. If the penetrance of the affected phenotype for a specific genotype is not defined, then that genotype gets the default penetrance. These penetrance values are then applied uniformly to all individuals, based only on their phenotype at the locus or factor indicating affection status.

0.5.8 Ascertainment and Probands

Although we have already briefly mentioned probands in the description of the definition file in Section 0.5.4.2, the subject merits a longer discussion. Each individual who is a proband should be labeled in the pedigree file. The locus or factor containing the proband label, and the label itself, are set using commands such as

```
PROBAND_FACTOR = P-status
PROBAND = proband
```

in the control file. These two commands will assign proband status to each individual that has the phenotype "proband" at the factor named P-status; no other individuals will be considered probands. The value assigned to the keyword `PROBAND` is not case sensitive and has the default value "proband". If you set `PROBAND` equal to the special value "Everyone", then everyone is considered a proband, and you may dispense with the superfluous factor, such as P-status, conveying proband status. Of course as discussed below, it defeats the purpose of ascertainment correction if everyone is treated as a proband.

Ascertainment correction is undertaken in analysis options such as Option 14 that rely on likelihood evaluation. Mendel conducts ascertainment correction by conditioning on proband phenotypes. If one or more probands occur in a pedigree, then Mendel collects the probands and assembles a special proband pedigree with enough connecting people to get the correct relationships among the probands. The loglikelihood of the proband pedigree is subtracted from the loglikelihood of the original pedigree during likelihood evaluation. The phenotypes and genotypes of the connecting people are deleted in the proband pedigree. The phenotypes and genotypes of the probands themselves are left intact. It may waste information to condition on all proband phenotypes and genotypes, so deleting some of the proband phenotypes in the proband pedigree may be advantageous in statistical inference. Two commands read from the control file facilitate this process. The commands

```
DELETE_PROBAND_FIELD = CHOL
DELETE_PROBAND_FIELD = SNP1
```

will delete SNP1 genotypes and CHOL quantitative values in the probands of a proband pedigree. To enable more radical deletion, the keyword DELETE_PROBAND_FIELD can take the values "ALL", "ALL_PHENOTYPES", and "ALL_VARIABLES". These commands may delete more than desired, so they can be partially overridden by commands such as

```
RESTORE_PROBAND_FIELD = CHOL
RESTORE_PROBAND_FIELD = SNP1
```

in the control file. If you are uncertain about what is being deleted, then insert the command

```
KEEP_HIDDEN_FILES = True
```

in the control file. This will allow you to inspect the intermediate pedigree file with proband pedigrees appended.

Option 14 operates somewhat differently than the variance components Options 19 and 20. The latter two options only pay attention to deletion of trait values. Predictors are never deleted since they are not used anyway for deleted traits. This partial deletion strategy suffices for ascertainment correction based on a subset of the component traits forming a multivariate trait.

0.6 Binary SNP Data Files

0.6.1 Overview

The availability of low-cost, high-density, genome-wide SNP genotyping technology is a real paradigm shift in statistical genetic analysis, no matter how hackneyed the cliché. A thousand-fold increase in the amount of data available for analysis, with little to no increase in cost, within just a couple of years, will revolutionize any field of study. As with any new paradigm, there is abundant "low-hanging fruit" that will answer many long-standing questions. (Of course not all susceptibility loci can be found by any one method, which is why Mendel has so many analysis options.)

Unfortunately, the great increase in data and its potential for new discoveries, comes with computing issues that must be addressed. Data sets with billions of genotypes become multi-gigabyte comma-separated, text-based files. Such large files are hard to transport and slow to read into analysis packages. They are also difficult to edit, but with billions of genotypes, editing individual values is no longer a high priority. Once the data has been brought into the analysis program, current computing technology makes working with that much data at once prohibitively slow.

To address the size of the data files, Mendel can use binary SNP data files in which each genotype is stored in two bits of memory, as opposed to the minimum of 32 bits required in our standard pedigree format described above. A 2000 individual and 1 million

SNP data set (with 2 billion genotypes) can then be put into a 500 MB binary file, which is large but much easier to handle than the 8 GB file required without the bit-wise construction. Modern computers usually have more than 500 MB of memory available per CPU core for their applications, although not that large an amount of fast, local cache memory. Not many computers have 8 GB of RAM available per CPU core for use by their application software.

Due to these storage and memory issues, use of the binary files are highly recommended, when feasible. In addition to the binary files, which are discussed in detail below, Mendel will often require short versions of the standard input files described above in Section 0.5. In particular, the control and pedigree files are always required, although the pedigree file will often contain no genotypes. The standard definition file will usually be necessary as well, but rarely will the map file be used with the SNP binary data sets. Thus, the user of the SNP binary files should also be familiar with the standard files previously discussed.

Currently only Analysis Options 10, 15, 21, 23, 24, 25, 27, 28, and 29 can use the binary SNP data sets as input. Mendel understands up to five new types of files in the "binary" SNP data sets. This is a slight misnomer as only two of the five files are binary, the remaining three are plain text files. Before delving into the details of their formats in the following sections, we give an overview of the five new files. Only the first two files listed here are used for most analyses.

1. **SNP Definition file**. This plain text file lists the names of the SNPs genotyped, and optionally, their chromosome and base-pair location. This can be a Plink .bim format file. The list must be in the same order as the genotypes in the SNP data file.

2. **SNP Data file**. This binary file contains the SNP genotypes. This can be a Plink .bed format file. The SNP data file must be coordinated with the pedigree file (for the order of the individuals) and with the SNP definition file (for the order of the SNPs).

3. **SNP Phase file**. This binary file contains the phase information for the SNP genotypes. If all genotypes are unordered or all are ordered, then this file is not necessary. If there is a mixture, then this file will inform Mendel on the phase status of each SNP genotype. Obviously, this file is tightly coordinated with the SNP data file.

4. **SNP Subset file**. This plain text file lists the subset of SNPs that should be analyzed or should be omitted from analysis. If all SNPs in the SNP definition file should be analyzed, then this file is not necessary.

5. **Sample Subset file**. This plain text file lists the subset of individuals that should be analyzed or should be omitted from analysis. If all individuals in the pedigree file should be analyzed, then this file is not necessary.

The two binary files are not designed to be edited in place. Mendel's options for trimming pedigrees (Option 21) and file conversion (Option 25) can manipulate these files. Although it is much easier to edit the SNP definition file, this is discouraged, since this file must be coordinated with the SNP data file. The SNP subset file is designed to be easily edited, and to encourage analyses with different subsets of SNPs. Similarly, editing the pedigree file is discouraged, since this file also must be coordinated with the SNP data file. The sample subset file is designed to be easily edited, and to encourage analyses with different subsets of individuals.

When using a binary SNP data set, the standard pedigree file will not contain any SNP genotypes. It will contain the pedigree names, the parent-child relationships (even if everyone is unrelated), the sex and twin status indicators, and the phenotypes at all non-SNP loci, factors, and variables, as described in Section 0.5.6. In this case, the standard definition file will only contain the definitions for these same non-SNP loci, factors, and variables. If there are no such entries, then the definition file is not necessary. However, more likely will be that the case-control label or other factors and quantitative traits will be included in the data set, and therefore must be included in the pedigree and definition files used in the analyses.

There is an exception to the rule that the genotypes in the SNP data file must be coordinated with the order of the individuals in the pedigree file. One may define a factor, and a label at that factor, that indicates which individuals were genotyped at the SNPs. Any individual that does not have the requisite label will have all his SNP genotypes set to missing, and the SNP data file must not include that individual's genotypes. Set this factor in the control file using the keyword `SNPS_TYPED_FACTOR`. The special label at this factor is set using the keyword `SNPS_TYPED`. The default value for `SNPS_TYPED` is the special value Everyone, which, regardless of the factor, indicates that everyone in the pedigree file is also used in the SNP data file.

0.6.1.1 Filtering on Genotyping Success Rates and Minor Allele Frequencies

In all SNP based analyses, SNPs and individuals with low genotyping rates are ignored. This filtering step is an important tool for removing likely false positives from association testing, as genotyping failure often occurs preferentially in cases or controls, or is correlated with the quantitative trait. The minimum allowed genotyping rates are set via the keywords `MIN_SUCCESS_RATE_PER_INDIVIDUAL` and `MIN_SUCCESS_RATE_PER_SNP`, which both have default values 0.98. These default values are reasonable for mass-produced SNP sets; for custom sets a lower threshold, say 0.9 or 0.95, may be more reasonable. To completely negate this filtering step, set both minimum allowed success rates to 0.0. All skipped individuals and SNPs are listed in the standard output file.

Mendel also allows one to filter SNPs by placing bounds on the minor allele frequencies

of the SNPs to be included in the analysis. The keywords MIN_MAF and MAX_MAF have default values 0.0 and 0.5, respectively, which results in no filtering by minor allele frequency. On the other hand, a command such as

```
MIN_MAF = 0.005
```

in the control file, means that all rare SNPs with minor allele frequency below 0.5% will be excluded from an analysis.

0.6.1.2 Plink format data files Mendel can read sets of standard, binary Plink files directly, that is, so-called .fam, .bim, and .bed files. One must still name the files in the Control file, for example with commands such as

```
DEFINITION_FILE = Def.txt
PEDIGREE_FILE = plink_filename.fam
SNP_DEFINITION_FILE = plink_filename.bim
SNP_DATA_FILE = plink_filename.bed
INPUT_FORMAT = PLINK
!
ANALYSIS_OPTION = GWAS
MODEL = 2
QUANTITATIVE_TRAIT = trait1
```

Assigning the value PLINK to the keyword INPUT_FORMAT informs Mendel to expect Plink formatted pedigree file (.fam file), SNP definition file (.bim file), and SNP data file (.bed file). A standard Definition file will still be required to name any trait, non-SNP genetic loci, factors, or quantitative variables in your data. However, this standard Definition file can be as simple as

```
trait1, factor
qtrait1, variable
qtrait2, variable
```

when the data includes a factor and two quantitative variables.

Any number of the three standard types of loci (genetic loci, factors, or quantitative variables) can be included in the pedigree file after the five standard columns listing the pedigree name, individual's name, father's name, mother's name, and sex. However, Mendel's rule that genetic loci must precede all factors, which in turn must precede all quantitative variables, must still be obeyed. As always, for each individual, the genotypes and phenotypes for the loci must be listed in the pedigree file in the same order the loci appear in the standard Definition file. Of course here Mendel will expect genotypes to be listed in Plink format, that is in two-column format; for example, "1 2" would be used for the genotype 1/2.

Mendel can read Plink files with at most one trait locus, in addition to any number of other genetic loci, factors, or variables. If a trait genetic locus, with phenotypes not genotypes, is included, then Mendel must be informed which locus that is, since their data will be in a single column in the pedigree file. Use the command `AFFECTED_LOCUS_OR_FACTOR` to identify the trait. For example, if the data set includes a genetic locus named epilepsy, which is encoded in the pedigree file using phenotypes (perhaps using "1" for controls and "2" for cases), then the command `AFFECTED_LOCUS_OR_FACTOR = epilepsy` would be appropriate.

Specifying PLINK input format automatically invokes the following commands as well

```
MALE = 1
FEMALE = 2
MISSING_VALUE = 0
MISSING_QUANTITATIVE_VALUE = -9
CASE2_CONTROL1 = True
```

Alternatively, the user can specify their own values for any of these keywords, in which case the user's values take precedence. For example, if your Plink case/control data set uses 1 for cases and 0 for controls, then simply include in your Control file the command `CASE2_CONTROL1 = False`. We note that for some case/control data sets, Plink allows either 0 (zero) or –9 to represent missing data at the locus listing the affection status. Mendel will only allow one symbol to be used: for quantitative variables the symbol assigned to the keyword `MISSING_QUANTITATIVE_VALUE`, which for Plink files has the default value –9; or for all other loci the symbol assigned to the keyword `MISSING_VALUE`, which for Plink files has the default value 0 (zero).

0.6.2 SNP Definition File

The required, plain text SNP definition file has a simple format. The first line of the file contains a single positive real value: the format version number, which is currently 3.0. (Format version 2 is also still accepted by Mendel.) Then for each SNP there is one additional line. Each of these lines contain seven values, only the first of which is required: (1) the name of the SNP; (2) its chromosome; (3) its position in base pairs; (4) name of the first allele; (5) name of the second allele; (6) its group label; and (7) its assigned weight for penalized regression analyses. The latter two values are currently only used for GWAS testing as described in Section 24. All SNPs must be listed in the same order that they appear in the SNP data file. It is *highly* recommended that this order be the known genomic SNP order, indeed it is essential for a proper analysis within Options 23 and 27.

The format version number will indicate what format these SNP files follow. Mendel will warn you if it does not know how to read the format you are using. As mentioned, the

names of the SNPs are required; all other values are optional. The names of the SNPs can be up to 16 characters long. The interpretation of the chromosome label, which is not case sensitive, is listed in Table 0.3 (except that “factor”, “variable”, and “superloci” labels are not permitted in the SNP definition file). The base pair position must be a positive integer.

If only single-point analyses will be performed, then position information is superfluous and may be omitted for all SNPs. However, included position information will be transferred to any plot file, and will be useful for graphing the results of the analyses. On the other hand, for SNP Imputation and Inbred Strains Analyses (Options 23 and 27) Mendel crucially uses the SNP order and information on where linkage breaks occur in the list of SNPs in the SNP definition file. If at some SNP the chromosome and base pair values are left blank, that SNP is treated as unlinked to all other SNPs. Without chromosome labels, if a position value is less than the value at the preceding SNP, then the distance between the two SNPs is set to infinity, again indicating unlinked SNPs. A completely blank line can also always be used to indicate a linkage break between the preceding and subsequent SNPs.

If chromosome labels are supplied for at least some SNPs, then all SNPs assigned to the same chromosome must be listed contiguously in the SNP definition file. For SNPs assigned to a specific chromosome but with ambiguous position, leave their base-pair value missing. SNP Imputation and Inbred Strains Analyses (Options 23 and 27) are multipoint, and thus crucially use the SNP order. Therefore, SNPs with ambiguous order are omitted from these analyses. Obviously, all positions on a given chromosome must be measured from the same arbitrary point near the start of the chromosome. All SNPs with the same chromosome label, and assigned base-pair positions, must be listed in ascending position order. For two SNPs with chromosome labels, a linkage break is placed between them if and only if their chromosome labels differ.

The SNP allele names can be up to four characters long. If left blank, they default to “1” and “2” for the first and second alleles, respectively. When binary data is converted to text-based files using Option 25, to save space in the possibly millions of genotypes output, only the first character of each allele name is used. For each SNP, if the given names are not unique in their first characters, then again “1” and “2” are used instead.

To analyze rare SNPs, it is often useful to group them based on proximity to genes or participation in common molecular pathways. If a SNP should be considered a member of a specific group, list that group as the fourth element on the SNP data line; otherwise, leave the fourth element empty. Group names can be up to 16 characters long. A group can contain any number of SNPs, and there can be any number of groups. The fifth element on the SNP data line is the relative weight assigned to this SNP when building a regression model with the user-specified number of predictors, for example, during a standard penalized regression association analysis. This weight must be a positive real

number. The larger the weight, the more likely the SNP will be retained in the model. If the keyword UNIFORM_WEIGHTS is set to True (the default), then SNPs that have not been assigned weights receive weight 1.0; otherwise, unassigned SNP weights take the value $1/\sqrt{4q(1-q)}$, where q is the minor allele frequency at the SNP. A weight can be assigned to a SNP regardless of whether the SNP is assigned to a group.

The SNP definition file is named using the keyword SNP_DEFINITION_FILE in the control file. For example, the SNP definition file

```
  3    ! FILE FORMAT VERSION NUMBER.
SNPaa
SNP00, 7,        ,  ,  ,       , 0.5
SNP03, 7,    9023, A, G, group1
SNP04, 7,  150389
SNPx1, X,   35298, 1, 2, group2, 0.25
SNPx2, X,   44783, T, C, group2, 0.4
```

lists six SNPs. The second SNP has an assigned weight even though is has no entry for base-pair position, alleles, or group. The spacing of the values on each data line is arbitrary and is aligned here just for readability. Commas should separate each value, but no commas are necessary after the last non-empty value on each line.

Mendel can also read standard, binary Plink input files. (See Section 0.6.1.2 for a full description.) Briefly, the control file command INPUT_FORMAT = PLINK informs Mendel to expect the SNP definition file to be formatted as a standard Plink .bim file (and the pedigree and SNP data files to be in standard Plink .fam and .bed formats, respectively). We note that this includes Plink's required use of the labels X, Y, XY, and MT for X-linked, Y-linked, XY-pseudo-autosomal, and mitochondrial SNPs, respectively. See Table 0.3 for a complete list of the allowed chromosome labels. Also, the missing value symbol is set to "0" (zero) to match Plink's.

0.6.3 SNP Subset File

The SNP subset file is an optional, plain text file that indicates the subset of SNPs to analyze. If all SNPs should be analyzed, then there is no need for this file. The first line of the SNP subset file starts with either the word "Include" or "Omit" (case insensitive). After this first line, the SNP subset file is simply a list of SNP names, one per line. If the first line says to include, then *only* the SNPs in this file are analyzed. If the first line says to omit, then all SNPs *except* the SNPs in this file are analyzed.

The SNPs listed in the SNP subset file must be a subset of those in the SNP definition file. Although not required, data initiation will proceed more quickly if the SNPs in the SNP subset file are listed in the same order as they appear in the SNP definition file, particularly if either file is large.

The SNP subset file is named using the keyword SNP_SUBSET_FILE in the control file. An example SNP subset file is

```
  omit   ! FILE FUNCTION.
SNPbb
SNPx1
```

We encourage users to edit the SNP subset file to set the SNPs used in the analysis, rather than trying to modify the SNP Definition and SNP data files. Note that the fraction of missing genotypes at a SNP may preclude its analysis (see Section 0.6.1.1), independent of the SNP subset file. It is crucial to also note that the SNP subset file only determines the SNPs potentially used in the analysis; all SNPs in the SNP definition file must still be represented in the SNP data file.

0.6.4 Sample Subset File

Analogous to the SNP subset file, the sample subset file is an optional, plain text file that indicates the subset of individuals to analyze. If all individuals should be analyzed, then there is no need for this file. The first line of the sample subset file again starts with either the word “Include” or “Omit” (case insensitive). After this first line, the sample subset file is simply a list of the individuals, one per line. For each individual, two items are required: the pedigree name and the individual name. If the first line says to include, then *only* the individuals in this file are analyzed. If the first line says to omit, then all individuals *except* the ones in this file are analyzed.

The individuals listed in the sample subset file must be a subset of those in the pedigree file. Although not required, data initiation will proceed more quickly if the individuals in the sample subset file are listed in the same order as they appear in the pedigree file, particularly if either file is large.

The sample subset file is named using the keyword SAMPLE_SUBSET_FILE in the control file. An example sample subset file is

```
  omit   ! FILE FUNCTION.
ped003, ID234
ped121, ID121
```

We encourage users to edit the sample subset file to set the individuals used in the analysis, rather than trying to modify the pedigree and SNP data files. Note that the fraction of missing genotypes for an individual may preclude its analysis (see Section 0.6.1.1), independent of the sample subset file. It is crucial to also note that the sample subset file only determines the individuals potentially used in the analysis; all individuals in the pedigree file must still be represented in the SNP data file, unless the SNPs_TYPED_FACTOR has been defined.

0.6.5 SNP Data File

The SNP data file is a binary file, not a text file. That is, the data is put into this file bit-by-bit, not character-by-character, so it is not viewable by normal text editors. This is necessary to obtain a large reduction in the size of the file. The SNP data file contains all SNP genotypes in the data set. The SNP data file is named using the keyword `SNP_DATA_FILE` in the control file.

If all the SNP genotypes are either unordered or missing, then the SNP data file can be in the common, SNP-major Plink binary pedigree format, that is, a so-called ".bed" file. (Plink is a well known statistical genetics package, and there are many tools for creating data files in this binary format.)

The genotypes are listed in the SNP data file in a specific order: genotypes at SNP#1 for each individual, genotypes at SNP#2 for each individual, and so on. This is the so-called SNP-major order. The individuals are represented in the same order they appear in the pedigree file. The SNPs are represented in the same order they appear in the SNP definition file.

Recall that not all individuals in the pedigree file need be represented in the SNP data file. However, those individuals who are represented in the SNP data file must be in the same order as they appear in the pedigree file. All SNPs in the SNP definition file must be present in the SNP data file, and in the same order. The remainder of this section spells out exactly how SNPs genotypes are compressed. Readers may want to skip this material on first reading.

The first two bytes of the SNP data file are special values that are used to authenticate the file. (Recall that one byte equals eight bits, and a bit is either 0 or 1.) If all genotypes are unordered, the first two bytes should be the Plink "magic number": `01101100 00011011`. Otherwise the first two bytes should be the number: `01001001 01010000`. The third byte specifies the overall phase information. If the third byte is `00000001`, then all genotypes are unordered; if `00000011`, then all genotypes are ordered; and if `00000101`, then the genotypes are mixed, some ordered, some not. For the purposes of this categorization, a missing genotype is considered to be unordered. When genotypes are mixed, then there must also be a SNP phase file that tells Mendel the phase of each genotype. The format of the SNP phase file is defined in Section 0.6.6. When the genotypes are not mixed, then no SNP phase file is used.

After the first three special bytes in the SNP data file, all data items represent SNP genotypes. Each SNP genotype is encoded as two bits. Two different encoding schemes are used in the file, depending on whether the genotype is ordered or unordered. The encoding scheme for unordered genotypes appears in Table 0.6. The encoding scheme for ordered genotypes appears in Table 0.7. After all genotypes for a given SNP are written, the bit value 0 is repeated until the next byte boundary.

Table 0.6: Encoding scheme for unordered SNP genotypes

Bit Code	SNP Genotype
00	Homozygous in first allele
01	Heterozygous
10	Missing genotype
11	Homozygous in second allele

Table 0.7: Encoding scheme for ordered SNP genotypes

Bit Code	SNP Genotype
00	Homozygous in first allele
01	Heterozygous 0|1
10	Heterozygous 1|0
11	Homozygous in second allele

These encodings clearly depend on which allele is the "first" allele and which the "second". The names of the first and second alleles at each SNP can be specified in the SNP definition file. For each SNP, the allele labeled as the first allele should be consistent across all data sets using that SNP. We note that the major and minor alleles may swap between ethnic groups. Swapping the meaning of the first and second alleles may manifest as effect estimates of equal magnitude but opposite sign in the results of two GWAS analyses. For human data, a convenient standard is to make the first allele be the allele in the Reference Human Genome.

To better understand the format of the SNP data file, consider the bits within each byte. Number the bits from 0 to 7, where 0 is the bit with the least significant digit and 7 the most significant. The bits are then read from 0 to 7, that is from right to left. A small example may make this clear. Consider a data set with five individuals, named $I1$, $I2$, $\ldots I5$, and two SNPs, named $S1$ and $S2$. Let a be the first allele of S1, and b the second. If all genotypes are unordered, then the SNP data file might look like

```
01101100 00011011 00000001 11011000 00000010 01011111 00000001
```

Ignoring the first three bytes, which do not encode genotypes, consider bytes four and five, which encode all genotypes for SNP $S1$. Since bits are read right to left within each byte, we can decode the genotypes as follows.

The two bytes encoding genotypes of SNP $S1$	11011000 00000010
First two bits, 00, indicate $I1$ is a/a	110110**00** 00000010
Second two bits, 01, indicate $I2$ is a/b	1101**10**00 00000010
Third two bits, 10, indicate $I3$ is not typed at $S1$	11**01**1000 00000010
Last two bits of first byte, 11, indicate $I4$ is b/b	**11**011000 00000010
Next two bits, 01, indicate $I5$ is a/b	11011000 000000**10**
After the last individual, the remainder of the byte is zero	11011000 **000000**10

0.6.6 SNP Phase File

The binary SNP phase file is only used if there are both ordered and unordered genotypes in the SNP data file. The structure of the SNP phase file is very similar to the SNP data file. For each SNP genotype, where the SNP data file uses two bits to encode a genotype, the SNP phase file uses a single bit to encode the phase status. A 0 indicates an unordered genotype, a 1 indicates ordered.

Unlike the SNP data file, the SNP phase file has no special bytes at its beginning. However, exactly the same genotypes are referenced in each of these binary files, and in the same order. Similar to the SNP data file, at the end of each SNP's data, the remainder of the byte is filled with zeros. The SNP phase file, when it exists, will be close to half the size of the SNP data file.

The SNP phase file is named using the keyword `SNP_PHASE_FILE` in the control file.

0.7 Column Formatted Input Files

Mendel can also read the standard definition, map, pedigree, and penetrance files in column-specific formats. The capability to read column-specific files is a legacy from previous versions of Mendel, and ensures backward compatibility with all existing Mendel data sets. However, all new data files should be written in the comma-separated style discussed above in Section 0.5. There is no column-specific analog for the SNP definition, data, phase, or subset files. In column-specific formats, each data value is read from a specific location in the file. Correct spacing is crucial in these files and is defined by the format specifications described below in the sections for each file type. Mendel is alerted to read an input file in column-specific format by the corresponding command

```
DEFINITION_LIST_READ = False
MAP_LIST_READ = False
PEDIGREE_LIST_READ = False
PENETRANCE_LIST_READ = False
```

in the control file. Alternatively, the single command

```
DEFAULT_LIST_READ = False
```

can be substituted for these four separate commands. This blanket declaration can be overridden for a specific input file by setting the corresponding file-specific keyword equal to true. In the past, column-specific files were the norm for Mendel, so you may come across sets of data files using column-specific formats but without the above commands in their control file. The easiest solution is to add the command `DEFAULT_LIST_READ = False` anywhere in the control file. Analysis Option 25 can convert from column-specific files to files in the more modern comma-separated, or list-directed, style discussed in Section 0.5.

The same data are represented in column-specific files as list-directed files; only their placement in the data files is changed. Thus, almost all rules on the data values remain. An exception is that locus, factor, and variable names must be eight or fewer characters in column-specific formats, as opposed to 16 or fewer for list-directed files. Therefore, if any such names are more than eight characters, then all input files containing those names must be list-directed. Quantitative values in list-directed files can be up to 64 characters long; they will usually need to be much shorter in column-specific files, often at most 8 or 10 characters depending on the specific format in use. As with list-directed files, the only non-digits allowed in a quantitative value are an initial + or –, a single period or comma indicating the start of the decimal digits, and a trailing D or E and integer exponent. For example, 4.25E–4 represents the number 4.25×10^{-4}.

The same forbidden and missing value symbols discussed in Section 0.5.2 apply to column-specific files. Blanks may indicate missing values since each value is read from a specific location in the file. The default allele separators remain / and \ for unordered genotypes, and | for ordered. In contrast to list-directed files, comment strings cannot appear in areas of column-specific files reserved for data. Maximum line lengths are not a concern with column-specific files since data is read from wherever the format specifies.

In the following sections, the column-specific formats of the definition, map, pedigree, and penetrance files are defined and illustrated. Before reading those sections, familiarize yourself with the descriptions of the corresponding comma-separated input files in Sections 0.5.4 through 0.5.7 above. The following sections stress the differences between column-specific formats and comma-separated formats. We recommend comma-separated files. Although column-specific files are easier to read, they are more error-prone to produce, edit, and maintain.

0.7.1 Fortran Format Codes

Although column-specific input appears tedious and restrictive, mastery of a few simple rules makes it easy and flexible. Data in almost any consistent format can be read, provided the format is clearly communicated to Mendel. The easiest way to describe Mendel's

formats is to review Fortran format conventions. Fortran uses the following format codes, also called descriptors, to describe data: (A) is used for character strings, (I) for integers, and (F) for decimal numbers. For example, (A8) specifies a string of eight or fewer characters, (I2) specifies an integer with at most two digits, and (F8.5) specifies a decimal number spread over eight or fewer spaces. The 5 in (F8.5) is ignored on input for decimal numbers; on output it specifies that numbers be written with 5 digits to the right of the decimal point. Numbers with more than two digits to the left of the decimal point cannot be output in this format. Fortran uses the descriptor (X) to indicate a blank or unread space, (T) to indicate transfer to a specific column, and (/) to skip to the next line. For example, (3X) specifies skipping three spaces, and (T28) specifies transferring to column 28 of the current input line. The 3 in (3X) represents an example of an edit descriptor preceded by a multiplier. The example (2A8) specifies two successive character strings of eight characters each.

To read multiple data items, format descriptors are concatenated and separated by commas. For example, the input format (2X,I3,T10,2A8,/,A6) specifies skipping the first two positions in a line, reading an integer with three digits, transferring to column 10, reading two character strings with eight characters each, skipping to the start of the next line, and reading a character string with six characters. Column-specific input files should not contain any tab characters. Use only blank spaces to line up data in required columns. The tab character is read in as a single character even though it can take up several columns on your computer screen. This makes it difficult to determine the true column a data value is in. Given this short primer, one can understand the majority of Fortran format statements and all the Mendel file formats listed below.

0.7.2 The Definition File

Recall from Section 0.5.4 that the definition file defines the genetic loci, factors, and quantitative variables in your data set. These entries must appear in the same order in the definition file as their corresponding phenotypes appear in the pedigree file. After the entries on variables, optional entries appear listing loci to be combined into super-loci. In a column-specific definition file, all values are read in character format to allow missing values to be represented by blanks or other symbols.

Genetic Loci Keeping the Fortran format conventions in mind, the following lines are required for each genetic locus:

1. Locus identifier line in (A8,A8,A2,A3) format specifying the
   (a) Locus name (A8)
   (b) Chromosome, e.g., 4, Autosome, or X-Linked (A8) [blank $\Rightarrow$ autosome]

   (c) Number of alleles (A2) [blank or 0 ⇒ to be determined by Mendel]

   (d) Number of phenotypes (A3) [blank ⇒ 0]

2. If item 1(c) is a positive number, then for each allele include a line in (A8,A8) format specifying the

   (a) Allele name (A8)

   (b) Allele population frequency (A8) [blank ⇒ to be estimated by Mendel]

3. If item 1(d) is a positive number, then for each phenotype include a line in (A8,A3) format specifying the

   (a) Phenotype name (A8)

   (b) Number of genotypes associated with the phenotype, or the word ALL if all genotypes are considered compatible with the phenotype (A3) [blank ⇒ ALL]

4. If item 3(b) is a positive number, then for each genotype include a line in (A17) format specifying

   (a) two allele names separated by /, \, |, or user-designated allele separator (A17)

Note that if you need locus names longer than eight characters, then you should use comma-separated input files, where locus names can be up to 16 characters long. Processing the definition file stops at the first blank line encountered, so be careful to delete any blank lines that precede the end of your data.

Here is a sample column-specific definition file:

```
EgomaniaAutosome2 2
a       0.99
b       0.01
NORMAL
AFFECTED
Marker1
Marker2
```

Although the name, chromosome, and number of alleles all run together at the first locus, they are read separately and correctly, because each value is read from specific columns.

On the other hand, if more information were known, then the definition file might look like:

```
Egomania        2 2
  a     0.99
  b     0.01
```

```
NORMAL   1
 a/a
AFFECTED 2
 a/b
 b/b
Marker1 12      3
 213
 217
 219
Marker2 XY      2
 +      0.545
 -      0.455
```

Here the trait has been defined as an autosomal dominant. Affected people have either genotype a/b or genotype b/b; normal people have genotype a/a. Marker1 is on chromosome 12 and has three alleles, 213, 217, and 219. Mendel is still requested to estimate their frequencies from the observed genotypes at this locus. Marker2 is in the pseudoautosomal region of the X chromosome and has two alleles, + and –, with specified frequencies.

Table 0.3 lists possible entries in the chromosome field and their interpretation by Mendel. The default value for this field indicates an autosomal locus. It is worth emphasizing that Mendel accepts and can perform all analyses with X-linked data.

The command `POPULATIONS` = n in the control file alerts Mendel to read n population-specific allele frequencies for each locus in the definition file. For column-specific definition files, each subsequent allele frequency is read using an additional (A8) format. Following the example in Section 0.5.4.5 where n = 2, a column-specific definition file might contain

```
APOE    19      3
2       0.0319  0.0394
3       0.8006  0.7835
4       0.1676  0.1772
```

with two population-specific allele frequencies at the APOE locus. Note here the implicit (3A8) format for each allele record.

Factors As mentioned earlier, the definition file also defines factors and their permitted categories. These are listed after all locus entries. The rules for constructing a factor entry are even simpler than for a locus entry:

1. Factor identifier line in (A8,A8,A2) format specifying the
   (a) Factor name (A8)

(b) The word “Factor” (case insensitive) (A8)

(c) Number of categories (A2) [blank or 0 ⇒ to be determined by Mendel]

2. If item 1(c) is a positive number, then for each category include a line in (A8) format specifying the

   (a) Category name (A8)

For example, a definition file might include the three factors

```
HEALTH  FACTOR  2
 Good
 Poor
Race    Factor
PROBAND factor  1
 Proband
```

defining health, race, and proband states.

Quantitative Variables After all loci and factors, the definition file contains entries for quantitative variables. The rules for constructing a quantitative variable entry are similar to those for a factor entry:

1. Variable identifier line in (A8,A8,A2) format specifying the

   (a) Variable name (A8)

   (b) The word “Variable” (case insensitive) (A8)

   (c) Number of bounds, either 0, 1, or 2 (A2) [blank ⇒ 0]

2. If item 1(c) is 1 or 2, then for each bound include a line in (A8,A8) format specifying the

   (a) Bound-type, either “lower”, “upper”, “minimum”, or “maximum” (case insensitive) (A8)

   (b) Bound value (A8)

The example definition file

```
Marker1
Marker2 x-linked
Smoking factor   3
 heavy
 light
```

```
 none
Health  Factor
BirthYr variable 2
 Lower   1900
 Upper   2000
BMI     VARIABLE
```

illustrates that loci are listed first, factors second, and variables third.

Super-Loci Finally, after the entries for variables, super-locus entries are permitted in the definition file. These optional entries define sets of loci that will be combined into super-loci when running Analysis Option 18, Combining_Loci. Unlike the locus, factor, and variable definitions that precede it, these super-locus entries have no corresponding fields in the pedigree file.

The rules for constructing a super-locus entry are again similar to those for a factor entry:

1. Super-locus identifier line in (A8,A8,A2) format specifying the
   (a) Super-locus name (A8)
   (b) The word “Superloc” (case insensitive) (A8)
   (c) Number of loci-to-combine, either 2, 3, or 4 (A2)
2. For each locus-to-combine include a line in (A8) format specifying the
   (a) Locus name (A8)

Note that the entry identifier, item 1(b) above, is restricted to eight characters, and so must be shortened to “superloc”.

The example definition file

```
SNP1    X-LINKED
SNP2    X-LINKED
SNP3    AUTOSOME
SNP4    AUTOSOME
SNP5    AUTOSOME
HEALTH  FACTOR   2
AFFECTED
NORMAL
BMI     VARIABLE
S1+S2   SUPERLOC 2
 SNP1
```

```
 SNP2
3+4+5   SUPERLOC 3
 SNP3
 SNP4
 SNP5
```

illustrates that super-locus entries must be the last entries in the definition file.

0.7.3 The Map File

Recall from Section 0.5.5 that the map file indicates the genetic loci used in analysis and the distances between them. Also recall that the map file has two styles, one listing the marker interval distances, the other listing the marker positions. The choice of distance units determines the style of map file Mendel expects. Genetic distance and recombination fractions are reserved for interval formatted map files; physical distances are reserved for position formatted map files. The distance units are set in the control file using the keyword `MAP_DISTANCE_UNITS` as discussed in Section 0.5.5.1.

Interval Format Map files using the interval format have a simple structure alternating two kinds of lines or records. Odd numbered lines contain locus names in (A8) format. Even numbered lines specify distances between adjacent loci in (8X,3A8) format. The first distance is the female distance and the second male. Note that both distances are read in (A8) format to allow blanks to indicate missing values. If you leave a female distance blank, then Mendel equates it to infinite distance, implying unlinked loci. If you leave a male distance blank, then Mendel equates it to the female distance. Thus, if only the first distance is listed, it is a sex-averaged distance. If two loci are completely unlinked, then leave both distances between them blank. For example, a simple map file using cM units

```
Marker2
        5.45
Marker1
        3.23    3.79
Marker3

Marker6
```

shows the sex-averaged distance between Marker2 and Marker1 is 5.45 cM; the female distance between Marker1 and Marker3 is 3.23 cM, the male distance 3.79 cM; and finally Marker6 is unlinked to the other loci.

Some of Mendel's analysis options permit you to specify a variable number of analysis points between adjacent pairs of loci. To change the number of points for a specific interval,

the new value should be listed in the map file after the male distance for that interval. Here is a sample map file with interior analysis points defined

```
Marker2
        5.45            4
Marker1
        3.23
Marker3
        2.0001  1.119   2
Marker5
```

For the first interval the distance of 5.45 cM will be used for both female and male distances, and the number of interior analysis points is set to 4. In the second interval, the number of interior analysis points is determined by the value of the keyword `INTERIOR_POINTS`. The third interval has sex-specific distance values and 2 interior analysis points.

Position Format Map files using the position format also have a simple structure. Every line is read in (A8,2A10) format. Each line starts with the name of the locus, lists the locus position, and concludes with the number of interior analysis points in the interval beginning at that locus. All positions on a given chromosome must be a physical distance measured from the same arbitrary point near the start of the chromosome. Mendel accepts three ways to indicate a switch to a new chromosome: 1) a missing position value, 2) a position value less than the previous position value, and 3) a blank line between two markers. In case 1) the locus is treated as unlinked to all other loci. Be careful with switch convention 2). A slight decline in map position can trigger an unintended switch. The loci preceding and following a switch are always unlinked. For example, consider the simple map file using Mb

```
Marker2 1.11      4
Marker1 6.56
Marker3 9.79      2
Marker5 11.7901
Marker6
Marker7 1.13
```

The interval between Marker2 and Marker1 has length 5.45 Mb and 4 interior analysis points. In the second interval, the number of interior analysis points is set by the current value of the keyword `INTERIOR_POINTS`. The third interval, between Marker3 and Marker5, has length 2.0001 Mb and 2 interior analysis points. Marker6 and Marker7 are unlinked to each other and to the other loci.

0.7.4 The Pedigree File

Recall from Section 0.5.6 that the pedigree file names individuals, records their relationships, and lists their phenotypes at all genetic loci, qualitative factors, and quantitative variables. That section also discussed pedigree records; these convey the number of individuals in a pedigree, the pedigree name, and optionally the pedigree copy number. When you use column-specific pedigree files, the default value for READ_PEDIGREE_RECORDS is true. Thus, unless this default is overridden, every pedigree should begin with a pedigree record line, and each person record should omit the pedigree name.

Unlike other column-specific input files, the user must supply the format codes for reading the pedigree file. This burden is balanced by the latitude to use pedigree files written in almost any consistent format. For the user's convenience and to allow arbitrary missing value symbols, all data format codes should be written using character descriptors (A), even when referring to integer or real values. The user supplies the format codes as the first two lines of the file. The first line lists the format for the pedigree record; the second line lists the format for the person record. If READ_PEDIGREE_RECORDS is set to false in the control file, then the format line for the pedigree record is omitted. The person record format then appears as the first line of the pedigree file.

The pedigree record format consists of character descriptors (A) for reading the pedigree size, name, and copy number. A typical example is (A5,A8,A4). Note the use of character descriptors A5 and A4 in this example, for the integer-valued pedigree size and copy number. (For backward compatibility, we allow but discourage the use of an integer descriptor (I) to read the size of the pedigree.) The person record format must include character descriptors (A) for each of the data items in the person record; these items are listed in Section 0.5.6.1. If you attempt to read any name — pedigree, individual, sex, and so forth — with an A format such as A12 that extends beyond eight characters, then only the last eight characters in that field will, in fact, be read. As an exception to this rule, ten characters are reserved for each phenotype. This permits reading of genotypes with allele names up to four characters long.

Here is a sample pedigree file in column-specific format using pedigree records:

```
(A5,A8,A4)
(3A8,2A1,3(1X,A8))
4    Bush
George                  M  AFFECTED 213/217  1946
Laura                   F  NORMAL   213/213  1946
Barbara George  Laura   F  NORMAL   213/213  1981
Jenna   George  Laura   F  AFFECTED          1981
3    Clinton
Bill                    M  AFFECTED 213/217  1946
Hillary                 F  AFFECTED 213/217  1947
```

```
Chelsea Hillary Bill    F  NORMAL   213/213  1980
```

In this example pedigree file, names of individuals and their parents are read in A8 format, sex and monozygotic twin status in A1 format, and the two phenotypes and single variable "BirthYr" in A8 format. To preserve confidentiality, you will almost always want to substitute coded IDs for names. Note the use of a blank field to indicate that Jenna's genotype at the second locus is missing. The file also illustrates the use of the default symbols F and M for females and males and the possibility of listing parents in either order.

To enable Mendel to read pedigree copy numbers, two conditions must be met. First, you should insert a third edit descriptor in the pedigree record format. This is the (A4) descriptor in the above sample pedigree file. We recommend keeping all three edit descriptors in the pedigree record format even if you decide not to use pedigree copy numbers. The second condition is that the command `READ_PEDIGREE_COPIES = True` must be inserted in the control file. With this change, we could read a copy number of 2 for the Bush family via the amended pedigree record

```
4    Bush    2
```

in the pedigree file. If the pedigree copy number is left blank, then the pedigree is counted as appearing only once in the data.

The above pedigree file with the line `READ_PEDIGREE_RECORDS = False` added to the control file looks like

```
(4A8,2A1,3(1X,A8))
Bush    George                  M  AFFECTED 213/217  1946
Bush    Laura                   F  NORMAL   213/213  1946
Bush    Barbara George  Laura   F  NORMAL   213/213  1981
Bush    Jenna   George  Laura   F  AFFECTED          1981
Clinton Bill                    M  AFFECTED 213/217  1946
Clinton Hillary                 F  AFFECTED 213/217  1947
Clinton Chelsea Hillary Bill    F  NORMAL   213/213  1980
```

Note that the pedigree record format specification is now omitted and that the person format has been changed to include another A8 format to read the pedigree ID on each line.

Comment strings should not be used in column-specific pedigree files, except that commented out and blank lines are allowed at the top of the file before the pedigree and person format statements. Completely blank lines will also be ignored before and after pedigree and person records, but not within person records that span multiple lines.

0.7.5 The Penetrance File

Recall from Section 0.5.7 that the penetrance file is an optional input file used with a few analysis options such as location scores and genetic risk prediction. Penetrance values specify the probabilistic relationship between hidden genotypes and observed phenotypes at a locus.

Every line of a penetrance file contains at least five ordered items: a pedigree name, a person name, a locus name, a genotype, and a penetrance value. An optional sixth item may be added as the last item on the line. This optional item conveys the observed phenotype of the named person at the named locus. Each line of a column-specific penetrance file is read using (A8,1X,A8,1X,A16,1X,A17,1X,A16) format.

Consider the trait Egomania featured in the presidential example. Suppose we posit for Egomania a dominant model with the phenocopy and penetrance rates

$$\begin{aligned} \Pr(\text{affected} \mid a/a) &= 0.10 \\ \Pr(\text{affected} \mid a/b) &= 0.95 \\ \Pr(\text{affected} \mid b/b) &= 0.95. \end{aligned}$$

The Clinton family portion of the corresponding column-specific penetrance file is:

```
Clinton,    Bill,       Egomania,          a/a,          0.10,   AFFECTED
Clinton,    Bill,       Egomania,          a/b,          0.95,   AFFECTED
Clinton,    Bill,       Egomania,          b/b,          0.95,   AFFECTED
Clinton, Hillary,       Egomania,          a/a,          0.10,   AFFECTED
Clinton, Hillary,       Egomania,          a/b,          0.95,   AFFECTED
Clinton, Hillary,       Egomania,          b/b,          0.95,   AFFECTED
Clinton, Chelsea,       Egomania,          a/a,          0.90,     NORMAL
Clinton, Chelsea,       Egomania,          a/b,          0.05,     NORMAL
Clinton, Chelsea,       Egomania,          b/b,          0.05,     NORMAL
```

Two things might strike you as odd about this particular file structure. First, the 1X codes in the format list allow Mendel to skip the embedded commas in reading the penetrance file. The above penetrance file is a legitimate comma-separated input file and can be read as such. Another oddity about the format list is that each penetrance value is read in A16 format, that is, as a character string. This makes it easier for Mendel to process the intended numerical value internally. Finally, we note that under the format codes listed above, only the first five values on each line are read; the optional phenotypes are obviously already known to Mendel from the pedigree file.

Application of model 2 of Option 14 eases the pain in creating flexible penetrance files with tailor-made penetrance values. Please see the instructions in the documentation of

Option 14 dealing with penetrance estimation and penetrance file creation under generalized linear models. As seen above, the penetrance files created by that analysis option can be read as column-specific or list-directed files.

For simple penetrance models, it is possible to omit the penetrance file and list penetrances in the control file. See Section 0.5.7.2 for an illustration.

0.8 Obsolete Input Files

Mendel continues to accept two older files, the locus file and the variable file, for backward compatibility. Since the functionality of these two older files has been replaced by the definition file, their support may end in future versions of Mendel. Obviously, any new data sets should be constructed using the definition file and not the locus and variable files.

The locus file, which was named using the keyword `LOCUS_FILE`, originally contained only entries for genetic loci and factors, not variables and super-loci. The locus file is subsumed by the definition file.

The variable file, which was named using the keyword `VARIABLE_FILE`, originally listed all quantitative variables. The specifications for these variables should now be included in the definition file as discussed in Section 0.5.4.3. The old variable file had a simple structure. For each variable there was a single line containing three values: the variable's name, its lower bound, and its upper bound. In column formatted files, these three items were read in (3A8) format. A missing lower bound was interpreted as $-\infty$ and a missing upper bound as $+\infty$.

A third old file is no longer supported by Mendel. The SNP file, which was named using the keyword `SNP_FILE`, listed loci to be combined into super-loci by Analysis Option 18, Combining_Loci. These lists of loci should now be included in the definition file as described in Section 0.5.4.4.

0.9 Output Files

0.9.1 Overview

Mendel has two kinds of output files: files simply listing numerical results, and files that take the form of Mendel input files. The numerical results files include the standard output file, the summary file, and the plot file. Depending on the analysis chosen, Mendel can also produce new definition files, new pedigree files, new penetrance files, and new SNP files. In the example data sets, the output files are named Mendel*.out, Summary*.out, Plot*.out, Def*.out, Ped*.out, Pen*.out, SNP_def*.out, SNP_data*_out.bin, and SNP_phase*_out.bin, where * indicates a number or a combination of a number and a letter. Thus, the standard output file for the second example of Analysis Option 7 is named Mendel7b.out. The

name assigned to an output file is set in the control file. Standard and summary output files are always mandatory; all other output files are sometimes mandatory and sometimes optional. When Mendel generated output files are mandatory, remember to name them in the control file. The standard output file and summary file have default names Mendel.out and Summary.out. Using these default names carries with it the risk that a previous version of the file will be overwritten during a Mendel run. If you wish to make your input or output files double-clickable, then it is often convenient to append the extension .txt to their names. Given this convention, you could replace Mendel7b.out by Mendel7b.out.txt or by Mendel7b.txt. The long name Mendel7b.out.txt preserves the distinction between input and output files.

The standard, more detailed output file has several functions. Foremost among these is to echo the input data in a tidy, readable fashion. Following this information are various tallies and statistics summarizing the pedigrees, the phenotypes, and the quantitative variables recorded on each person. The echoed and summarized data are intended to assist you in screening for input errors. After the summary statistics, the penetrance file appears. If allele frequencies are omitted in the definition file, then preliminary estimates are written. Finally, the standard output file contains all error messages and warnings issued by Mendel. Errors often bring Mendel to a premature halt, so it is important to scan the standard output file carefully for reported errors. The most common errors involve missing input files, incorrect Fortran formats, and genetic inconsistencies. In many cases Mendel will continue processing data beyond the first encountered error or aborted problem. If Mendel grinds to a premature halt, be sure to search for all occurrences of the word “error” in the standard output file.

After echoing input data and noting errors, Mendel commences analysis. Often this means finding maximum likelihood estimates or evaluating likelihoods over a grid in parameter space. Mendel incorporates a powerful optimization engine for maximum likelihood estimation. This engine accommodates parameter upper and lower bounds and linear equality constraints. The loglikelihood and parameter values Mendel outputs at each iteration of a search can be helpful in computing likelihood ratio statistics or diagnosing failure to reach convergence. Sometimes analysis output is so voluminous that it is difficult to interpret quickly. The summary file is designed to make interpretation painless. This ideal is not always achieved, but provided nothing seems amiss in running one of the analysis options, you will likely want to look first in the summary file for important conclusions.

Summary files are constructed by all analysis options, and usually contain the “bottom-line” of the analysis results. This file will usually be the first output file viewed after an analysis. At the bottom of this file are references that can be used to cite Mendel and specific algorithms used in the analysis performed. For example, the general Mendel reference is listed as

```
FOR PUBLICATION PURPOSES, PLEASE CITE THE OVERALL REFERENCE:

Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM (2013)
Mendel: The Swiss army knife of genetic analysis programs.
Bioinformatics 29:1568-1570.
```

Naturally, we would appreciate your acknowledgment whenever you use Mendel [62, 66].

The plot file, when available, contains data arranged for easy plotting by an external package. New definition and new pedigree files are used only in Analysis Options 3, 5, 6, 16, 17, 18, 20, 21, 23, and 25. Because the content of these output files varies so much from analysis option to option, we refer the reader to each relevant option for a description of the content of these output files.

Occasionally, intermediate output is directed to your screen. This may give you an idea of how quickly Mendel is finishing an assigned task such as maximum likelihood estimation. To turn off such output, set the keyword VERBOSE to false.

0.9.2 Detailed Output File

Mendel0.out, the standard output file of the earlier presidential example, illustrates these conventions. It first gives the version number of Mendel, copyright claim, time, and date. Time is measured in hours (in military or European fashion) and minutes. The date appears as month/day/year. Here is the relevant output:

```
                    WELCOME TO MENDEL, VERSION 14.5

                 (c) COPYRIGHT KENNETH LANGE, 1985-2015
                     ALL COMMERCIAL RIGHTS RESERVED
                PROGRAMMED BY KENNETH LANGE & ERIC SOBEL

            DOCUMENTATION AND MOST RECENT VERSION AVAILABLE AT
               http://www.genetics.ucla.edu/software/mendel


TIME OF DAY   : 14:40
MONTH/DAY/YEAR: 09/01/2015
```

After these preliminaries, Mendel0.out echoes the control file stripped of all comments.

```
CONTENT OF CONTROL FILE Control00.in:

DEFINITION_FILE = Def0.in
MAP_FILE = Map0.in
```

```
PEDIGREE_FILE = Ped0.in
OUTPUT_FILE = Mendel0.out
SUMMARY_FILE = Summary0.out
ECHO = Yes
ANALYSIS_OPTION = Mistyping
MODEL = 1
```

The definition file is echoed next. Note how the input data has been reorganized for easier reading.

```
CONTENT OF DEFINITION FILE Def0.in:

LOCUS = Egomania        CHROMOSOME = Autosome    2 ALLELES    2 PHENOTYPES
ALLELE NAMES     FREQUENCIES
   a             0.9900000
   b             0.0100000
PHENOTYPES
     NORMAL       1 GENOTYPE
                       a\a
     AFFECTED     2 GENOTYPES
                       a\b
                       b|b

LOCUS = Marker1         CHROMOSOME = Autosome    2 ALLELES    0 PHENOTYPES
ALLELE NAMES     FREQUENCIES
   213           0.4450000
   217           0.5550000

VARIABLE = Birth
   LOWER BOUND =  1900.0000
   UPPER BOUND =  2004.0000
```

The echoed map file gives recombination fractions for each interval between adjacent model loci. It also lists either the genetic map distances for each interval or the physical map position at each model locus. If any locus names are more than 8 characters long, the output includes a key showing long and short names for each locus.

```
LIST OF MODEL LOCI FROM MAP FILE Map0.in:

MODEL  LOCUS     FEMALE    MALE       FEMALE    MALE
LOCUS  NAME      THETA     THETA      MORGANS   MORGANS

   1  Egomania
                0.10000   0.05000    0.11157   0.05268
   2  Marker1
```

Finally, the pedigree file lists all pedigrees and all people within a pedigree. Pedigrees are numbered for easy reference. Phenotypes and quantitative variables are presented in clumps of three per line.

```
CONTENT OF PEDIGREE FILE Ped0.in:

PEDIGREE NUMBER 1 HAS 4 MEMBERS AND NAME BUSH.

     ID       PARENT   IDS      SEX   TWIN   Egomania   Marker1    Birth

  1  George                     M            AFFECTED   \217       1946.00
  2  Laura                      F            NORMAL     213\217    1946.00
  3  Barbara  George   Laura    F            NORMAL     213|217    1981.00
  4  Jenna    George   Laura    F            AFFECTED              1981.00

PEDIGREE NUMBER 2 HAS 3 MEMBERS AND NAME CLINTON.

     ID       PARENT   IDS      SEX   TWIN   Egomania   Marker1    Birth

  1  Bill                       M            AFFECTED   213\213    1946.00
  2  Hillary                    F            AFFECTED   \213       1947.00
  3  Chelsea  Hillary  Bill     F            NORMAL     213\213    1980.00
```

After the input files are echoed, Mendel0.out presents summary statistics on pedigrees and quantitative variables and tallies of phenotypes. The label FAMILIES refers to the number of nuclear families over all pedigrees. Tallies include missing data when relevant. All summary statistics and tallies incorporate the number of copies of each pedigree. It is almost always worthwhile to examine summary statistics and tallies in detail. Many input errors can be caught if you are well acquainted with your data. Here are the descriptive statistics from Mendel0.out:

```
TOTALS:   PEDIGREES  FAMILIES  PEOPLE  FEMALES   MALES   MZTWINS  FOUNDERS
               2         2        7        5        2        0        4

TALLIES FOR THE PHENOTYPES:

LOCUS      PHENOTYPE      COUNT

 Egomania   AFFECTED         4
 Egomania   NORMAL           3

LOCUS      PHENOTYPE      COUNT
```

```
 Marker1    MISSING          1
 Marker1    213\213          2
 Marker1    213\217          1
 Marker1    213|217          1
 Marker1    \213             1
 Marker1    \217             1


DESCRIPTIVE STATISTICS FOR THE QUANTITATIVE VARIABLES:

VARIABLE           Birth

MEAN            1961.00000
STD DEVIATION     17.03777
SKEWNESS           0.28856
KURTOSIS           1.08604
MINIMUM         1946.00000
MEDIAN          1947.00000
MAXIMUM         1981.00000
VALUES PRESENT           7
VALUES MISSING           0
TRANSFORM             NONE
```

If you want sex specific descriptive statistics as well, then insert the line

```
STATS_BY_SEX = True
```

in the control file. If you tell Mendel where to look for affected people by adding the line

```
AFFECTED_LOCUS_OR_FACTOR = Egomania
```

to the control file, then the standard output file will list the number of affecteds in the pedigree data directly under the number of pedigrees. The same applies to probands. For example, suppose you add the factor P-status defining proband status to the definition and pedigree files. If you now insert the line

```
PROBAND_FACTOR = P-status
```

in the control file, then Mendel would be able to count the number of probands in the pedigree file. This simple example does not include a penetrance file. Sample problems 2b, 7b, and 7c require penetrance files.

One can easily generate output illustrating Mendel's capacity for estimating allele frequencies by gene counting. If we omit allele frequencies at the Egomania locus in Def0.in, then Mendel0.out displays the estimated frequencies

```
 GENE COUNTING ESTIMATES OF UNKNOWN ALLELE FREQUENCIES:

 LOCUS    ALLELE NUMBER    ALLELE NAME      FREQUENCY

 Egomania       1              a              0.6547
 Egomania       2              b              0.3453
```

for the normal and affected alleles. Despite the fact that four out of seven pedigree members are affected, the affected allele has a lower estimated frequency than the normal allele. This just reflects the fact that the trait is dominant. Again it is worth repeating our admonition to enter known allele frequencies in the definition file whenever possible.

On occasion, you may want to modify what data is dumped into the standard output file. The keyword ECHO provides this capability. The default value of ECHO is No (or False), which omits both the input data and the maximum likelihood search iterations in the standard output file; resetting ECHO to Partial (or Some) omits the input data but echoes detailed information on the search iterations; resetting ECHO to Yes (or True) tells Mendel to echo both the input files and the search iteration data.

0.9.3 Search Output

The SEARCH program embedded in Mendel will maximize the loglikelihood of your data or evaluate it over a predefined grid of points in parameter space. The choice of search mode or grid mode is determined by the keyword TRAVEL. To select grid mode rather than the default search mode, you must insert the command

```
TRAVEL = GRID
```

in the control file. Analysis Option 2 on location scores explains how you can exercise some control over the nature of the grid. The output

```
                                PROBLEM 1

GRID OR SEARCH OPTION: GRID
LOG BASE: 10

ITER  STEPS  LOGLIKELIHOOD    XX THETA    XY THETA

  1     0    0.6113161E+00  0.1000E-02  0.1000E-02

  2     0    0.2283546E+01  0.1000E-01  0.1000E-01

  3     0    0.3327551E+01  0.5000E-01  0.5000E-01
```

```
  4     0    0.3467841E+01  0.1000E+00  0.1000E+00

  5     0    0.3380372E+01  0.1500E+00  0.1500E+00

  6     0    0.3185419E+01  0.2000E+00  0.2000E+00

  7     0    0.2372344E+01  0.3000E+00  0.3000E+00

  8     0    0.1136151E+01  0.4000E+00  0.4000E+00

  9     0    0.0000000E+00  0.5000E+00  0.5000E+00

THE MAXIMUM LOGLIKELIHOOD OCCURS AT ITERATION   4.
```

from Mendel2.out is typical. It numbers the problem, informs you that loglikelihoods are base 10 rather than base e, and gives the loglikelihoods at the grid points. Parameters are named for easy interpretation, and the largest loglikelihood is flagged. Exponential notation is used for loglikelihoods and parameter values; thus, a number such as 0.1136151E+01 represents $0.1136151 \times 10^1 = 1.136151$.

In search mode, SEARCH finds maximum likelihood estimates subject to parameter lower and upper bounds and linear constraints. It also provides asymptotic standard errors and correlations of all parameter estimates. On an algorithmic level, SEARCH carries out recursive quadratic programming with quasi-Newton updates of the observed information matrix. Although users are shielded from the computational details, it is important to be able to interpret the output from SEARCH.

The standard output file records all parameter constraints. For instance, the frequency estimation output from problem 1 of Mendel6d.out

```
                                PROBLEM 1

 GRID OR SEARCH OPTION: SEARCH
 LOG BASE:  E

 PARAMETER MINIMA AND MAXIMA:

     FREQ 1      FREQ 2      FREQ 3      FREQ 4      FREQ 5      FREQ 6
     FREQ 7      FREQ 8

    0.000001    0.000001    0.000001    0.000001    0.000001    0.000001
    0.000001    0.000001
```

```
   INFINITY    INFINITY    INFINITY    INFINITY    INFINITY    INFINITY
   INFINITY    INFINITY

PARAMETER CONSTRAINTS:

CONSTRAINT # 1 SUM =   1.000000

CONSTRAINT # 1 COEFFICIENTS:

    FREQ 1      FREQ 2      FREQ 3      FREQ 4      FREQ 5      FREQ 6
    FREQ 7      FREQ 8

   1.000000    1.000000    1.000000    1.000000    1.000000    1.000000
   1.000000    1.000000
```

shows that Mendel imposes a lower bound of 0.000001 on each allele frequency being estimated and forces these frequencies to sum to 1.0. No parameter upper bounds are necessary in this situation. In general, if there are n parameters labeled $p_1, \ldots, p_n$, then a linear constraint has the form

$$c_0 \;=\; c_1 p_1 + \cdots + c_n p_n.$$

The constant c_0 is the constraint value and the multipliers $c_1, \ldots, c_n$ are the constraint coefficients. In the example just given, these numbers are all 1.0.

The output

```
ITER  STEPS  LOGLIKELIHOOD     FREQ 1      FREQ 2      FREQ 3      FREQ 4
                               FREQ 5      FREQ 6      FREQ 7      FREQ 8

  1     0    0.0000000E+00  0.7534E-02  0.6041E-02  0.7909E-03  0.6341E-03
                            0.4947E+00  0.3967E+00  0.5193E-01  0.4164E-01

  2     3   -0.2254806E+00  0.7470E-02  0.5990E-02  0.7842E-03  0.9085E-02
                            0.4906E+00  0.3933E+00  0.5149E-01  0.4129E-01

  3     2    0.3028909E+00  0.7242E-02  0.5807E-02  0.7603E-03  0.8808E-02
                            0.4756E+00  0.3813E+00  0.8049E-01  0.4003E-01
  .     .     .
  .     .     .
  .     .     .

 13     0    0.5566832E+01  0.1000E-05  0.1249E-01  0.1000E-05  0.2509E-02
                            0.4625E+00  0.4300E+00  0.9248E-01  0.1000E-05
```

```
 14     0    0.5566834E+01  0.1000E-05  0.1249E-01  0.1000E-05  0.2503E-02
                            0.4625E+00  0.4300E+00  0.9249E-01  0.1000E-05

 15     3    0.5566831E+01  0.1000E-05  0.1252E-01  0.1000E-05  0.2501E-02
                            0.4625E+00  0.4300E+00  0.9249E-01  0.1000E-05

THE MAXIMUM LOGLIKELIHOOD  0.5566834E+01 OCCURS AT ITERATION  14.

ASYMPTOTIC STANDARD ERRORS OF THE PARAMETERS:

    FREQ 1      FREQ 2      FREQ 3      FREQ 4      FREQ 5      FREQ 6
    FREQ 7      FREQ 8

 0.0000E+00  0.8880E-02  0.0000E+00  0.5419E-02  0.3549E-01  0.3501E-01
 0.2089E-01  0.0000E+00

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS:

    FREQ 1      FREQ 2      FREQ 3      FREQ 4      FREQ 5      FREQ 6
    FREQ 7      FREQ 8

   0.0000

   0.0000      1.0000

   0.0000      0.0000      0.0000

   0.0000     -0.3480      0.0000      1.0000

   0.0000     -0.1446      0.0000      0.0549      1.0000

   0.0000     -0.0872      0.0000     -0.0290     -0.8001      1.0000

   0.0000      0.0571      0.0000     -0.1561     -0.3107     -0.2720
   1.0000

   0.0000      0.0000      0.0000      0.0000      0.0000      0.0000
   0.0000      0.0000
```

of Mendel6d.out immediately follows the bound and constraint summary. This output records the successive iterations of SEARCH in ascending the likelihood surface. Except for iterations 2 and 15, the loglikelihood steadily increases, and parameter values smoothly converge to their maximum likelihood estimates. The barely perceptible decline of the log-

likelihood at iteration 15 is of no consequence. The label STEPS in the output refers to the number of step halvings at each iteration. After the iteration summary, SEARCH gives the asymptotic standard errors of the parameters and the asymptotic correlation matrix. These quantities are helpful in assessing the precision of the parameter estimates and the amount of independent information available on each parameter. In many analysis options, overall conclusions from maximum likelihood estimation are presented in the summary file.

0.9.4 Pedigree Deviances

Once a genetic hypothesis is tested, it is helpful to identify those pedigrees giving the greatest weight in favor of the null or alternative hypotheses. A pivotal pedigree can be identified by computing its deviance, which is defined as twice the difference in its loglikelihood (base e) under the alternative and null hypotheses. These loglikelihoods are evaluated at the corresponding maximum likelihood estimates for the entire sample. A pedigree with a large positive deviance favors the alternative hypothesis, while a pedigree with a large negative deviance favors the null hypothesis. To compute deviances, the command

```
DEVIANCES = True
```

must be inserted in the control file. Analysis Options 2, 8, 14, 19, and 20 act on this command. The standard output file Mendel8a.out contains the fragment

```
DEVIANCES ENCOUNTERED IN MAXIMUM LIKELIHOOD ESTIMATION

PEDIGREE   PEDIGREE  DEVIANCE
 NUMBER      NAME

     1     KUS          2.3159
     5     BOD          0.3688
     2     KRA         -0.1237
     4     STO         -0.3324
     3     NEU         -0.4780

THE ORDERED-SUBSET DEVIANCE STATISTIC HAS APPROXIMATE P-VALUE 0.3588
PLUS OR MINUS 0.009593 BASED ON 10000 RESAMPLES. THE MAXIMUM SUM HAS
1 TERM.
```

describing the deviances of the five pedigrees under analysis. If one or more pedigrees occur in multiple copies, then a third column is output reporting the adjusted deviance per pedigree.

In some genetic studies, clinical evidence such as age of onset can be used to rank pedigrees on how likely each is to depart from the null hypotheses of no association or

no linkage. Mendel outputs an ordered-subset deviance statistic $D_{\max}$ that is helpful in deciding whether a particular prior ranking is justified [42]. Suppose that pedigree i of n pedigrees has deviance d_i. Then the statistic

$$D_{\max} = \max_{1 \le j \le n} \sum_{i=1}^{j} d_i$$

should be large if the ranking is predictive and pedigrees are input according to their ranks. Mendel computes the p-value of the observed $D_{\max}$ by permuting deviances across pedigrees. If pedigrees are ranked by size with the largest pedigrees coming first, then view this p-value with skepticism. If you want to override the implicit assumption in computing $D_{\max}$ that pedigrees are taken in their order of occurrence in the pedigree file, then define the keyword `DEVIANCE_VARIABLE` in the control file. Thus, the command

```
DEVIANCE_VARIABLE = Onset
```

instructs Mendel to re-order deviances by the average of the variable Onset within each pedigree. The pedigree with the smallest average is taken first in computing $D_{\max}$, then the pedigree with second lowest average, and so forth. Pedigrees with no observed value for the designated variable are treated as having average equal to $+\infty$ and taken last. Mendel will report the rank of each pedigree relative to an ordering variable such as Onset. If you define DEVIANCE_VARIABLE, then there is no need to set DEVIANCES equal to true in the control file.

0.9.5 New Definition and Pedigree Files

Some analysis options output new definition files, new pedigree files, or both. These novel output files are named via instructions such as

```
NEW_DEFINITION_FILE = Def18.out
NEW_PEDIGREE_FILE = Ped18.out
```

in the control file. The names of the new definition and pedigree files should not coincide with the names of the old definition and pedigree files. New definition and pedigree files obey certain conventions. For example, if the original definition or pedigree file is comma-separated, then any new file of the same type is also comma-separated. Allele separators in force in the original definition or pedigree file are also in force in the new file. If pedigree records or pedigree copy numbers appear in the original pedigree file, then the same can be said for the new pedigree file.

Despite preserving these distinctions, all person records in a new pedigree field are output in the same comma-separated format. If the files are to be read in column-specific

format, a person format is written at the top of the new pedigree file instructing Mendel to skip over the embedded commas. For example, our standard presidential pedigree would be output as

```
      4,Bush
George    ,          ,          ,M        ,          ,a/b       ,213/217   ,1946
Laura     ,          ,          ,F        ,          ,a/a       ,213/213   ,1946
Barbara   ,George    ,Laura     ,F        ,          ,a/a       ,213/213   ,1981
Jenna     ,George    ,Laura     ,F        ,          ,a/b       ,          ,1981
      3,Clinton
Bill      ,          ,          ,M        ,          ,a/b       ,213/217   ,1946
Hillary   ,          ,          ,F        ,          ,a/b       ,213/217   ,1947
Chelsea   ,Bill      ,Hillary   ,F        ,          ,a/a       ,213/213   ,1980
```

if the pedigrees enter Mendel in comma-separated format with pedigree records. If the pedigrees enter in column-specific format, then the output would be the same except for the two format lines

```
(A8,1X,A8,1X,A8)
(20000(A10,1X))
```

at the top. The 1X codes within the formats tell Mendel to skip the embedded commas in the pedigree and person records. The format reserves ten columns for each data item — individual and parent names, sex, twin status, phenotypes, and variables. Obviously, if READ_PEDIGREE_RECORDS were false, then pedigree names would appear as the first item in each person record.

0.10 Exceptions to Normal Operation

In practice, Mendel runs may fail for one reason or another. Messages indicating mistakes in data input are usually obvious. Other reasons for failure are deeper. We now discuss some of the more subtle ones.

0.10.1 Issues of Computational Complexity

Pedigree analysis is one of the most computationally intensive tasks in modern biology. Mendel can be overwhelmed by the synergistic obstructions of missing data, multiple markers, multiple alleles per marker, and inbreeding. In such situations, Mendel's usual response is to skip the pedigree and print a corresponding warning in the standard output file. For instance, in the NPL output of Analysis Option 4, you might get the message:

```
*** WARNING *** THE FOLLOWING PEDIGREES WERE TOO COMPLEX FOR ANALYSIS:

PEDIGREE  PEDIGREE  ADJUSTED
 NUMBER     NAME    MEIOSES

    5          5       17
   19         19       18

MAXIMUM_ADJUSTED_MEIOSES CURRENTLY = 16
```

Three keywords control Mendel's thresholds for tackling large pedigrees. From the above NPL analysis output, we know immediately that the Lander-Green-Kruglyak algorithm has skipped pedigrees numbered 5 and 19 because their adjusted meioses count was slightly higher than the current maximum allowed. The remedy is to insert the command

```
MAXIMUM_ADJUSTED_MEIOSES = 18
```

in the control file. The keyword MAXIMUM_ADJUSTED_MEIOSES (or equivalently the simpler MAX_ADJUSTED_MEIOSES) has default value 16, except in Options 5 and 19 where its default is 12. Keep in mind that the computation time for a pedigree with f founders and c children is roughly proportional to 2^{2c-f} in the Lander-Green-Kruglyak algorithm. We say "roughly" because the recent enhancements [77] to the algorithm programmed in Mendel can do a lot better on thoroughly typed pedigrees. Even with these improvements, setting the keyword MAX_ADJUSTED_MEIOSES above 23 is apt to lead to very long computation times. With a dense marker map, it is also possible to accelerate likelihood evaluation by discarding highly unlikely descent graphs at each locus [1]. Think of a descent graph as a pattern of gene flow. Although Mendel automatically deletes highly unlikely descent graphs, you can discard a greater fraction by inserting some version of the command

```
GENE_FLOW_CUTOFF = 1.0e-8
```

in the control file. The keyword GENE_FLOW_CUTOFF has the default 1.0e–10 = 10^{-10}. Of course, as the value of this keyword approaches 1, the accuracy of your results will suffer, and it is best to err on the conservative side.

If the Elston-Stewart algorithm fails on one or more pedigrees, then try increasing the keyword COMPLEXITY_THRESHOLD above its default of 50000000.0 = 5.0e7. (Note that numbers should contain at most a single period or comma indicating the start of the decimal digits.) The standard output file of Mendel will ordinarily tell you which algorithm applies. If it appears that both algorithms apply, then try changing all three of the keywords. If these tactics are futile, and often they will be, then break the pedigree into sub-pedigrees, reduce the number of loci in the map file, or combine the alleles within each locus. Analysis

Option 21 addresses the first possibility, and Analysis Option 16 addresses the third possibility. If none of these measures succeeds, you can resort to programs such as SimWalk that rely on MCMC (Markov chain Monte Carlo) sampling [69, 99].

0.10.2 Problems with Maximum Likelihood Estimation

Occasionally, SEARCH, the optimization engine of Mendel, will experience a convergence problem. This usually manifests itself as 1) a failure of the loglikelihood to increase from one iteration to the next, 2) attainment of the maximum likelihood at an intermediate iteration, or 3) premature stopping at iteration 200. Sometimes SEARCH will hit a part of parameter space causing the likelihood to underflow and halting operation of Mendel. (Underflow occurs when a very small number is incorrectly equated to 0.) There are several remedies for failure of convergence. These should be exercised with considerable caution.

First, inspection of the column STEPS in the iteration output may reveal that one or more iterations have attained the default maximum value of 3. This suggests that SEARCH is overshooting the maximum and is trying to correct itself by backtracking along the current search direction. One can try resetting the maximum number of backtracking steps by inserting the command

```
MAX_STEPS = 4
```

in the control file. This will permit an even shorter step in the current search direction. A possibly better remedy is to limit the length of the search direction vector by issuing the command

```
MAX_STEP_LENGTH = x
```

for a value x dictated by the scale of the parameters. For instance, in the sample data set for Analysis Option 16, log cholesterol values vary between 4.9 and 6.4. Too large a change in a genotype mean for log cholesterol may land SEARCH in a region where Gaussian densities underflow. This suggests that x be 1.0 or less. The default maximum step length is ∞.

One can increase the maximum allowed number of iterations from 200 to 400, say, via the command

```
MAX_ITERATIONS = 400
```

in the control file. SEARCH declares convergence when the loglikelihood changes by less than 0.0001 for 4 successive iterations. Tighter or looser convergence is governed by the keywords `CONVERGENCE_TESTS` and `CONVERGENCE_CRITERION`. For instance, if you want SEARCH to declare convergence when the loglikelihood changes by less than 0.01 for 2 successive iterations, then insert the commands

```
CONVERGENCE_TESTS = 2
CONVERGENCE_CRITERION = 0.01
```

in the control file.

Fine-tuning of convergence criterion is not the only way to improve maximum likelihood estimation. Because the optimization engine of Mendel depends on local information such as derivatives, it can converge to a local maximum rather than to the global maximum. You can specify your own parameter starting values and bounds through commands such as

```
PARAMETER_INITIAL_VALUE = 0.2 :: a
PARAMETER_MIN = 0.0 :: b
PARAMETER_MAX = 1.0 :: b
```

in the control file. These three commands set the initial value of the parameter named "a" to 0.2 and the minimum and maximum values of the parameter named "b" to 0.0. and 1.0, respectively. You can deduce Mendel's parameter naming conventions by running the current analysis option and looking at the standard output file.

If you want to place a linear constraint on the parameters during maximum likelihood estimation, then insert a command such as

```
PARAMETER_EQUATION = -4.4*d + 2*e - 10.3*b :: 2.5
```

in the control file. This forces the parameters named d, e, and b to satisfy the linear equation $-4.4d + 2e - 10.3b = 2.5$. There are a few restrictions on the format of this command: (1) the coefficients of all parameters must be integers or real numbers written in standard, not exponential, format; (2) there must be an asterisk between each coefficient and its parameter, except if the coefficient is 1, in which case the parameter name alone is sufficient; and (3) between each pair of linear terms there must be a plus or minus sign. To fix a parameter at a particular value, for example to fix m at -1.5, use a command such as

```
PARAMETER_EQUATION = m :: -1.5
```

To constrain two parameters named U and V to be equal, use a command such as

```
PARAMETER_EQUATION = U - V :: 0.0
```

When you take advantage of these features of Mendel, make certain that your initial parameter values are consistent with all bounds and linear constraints imposed. If a linear constraint or bound is not satisfied, then Mendel will perturb the initial parameters values so that the constraint or bound is satisfied.

0.11 Stochastic Operations

Several of Mendel's analysis options use permutation tests or simulation, for example, NPL scores in Option 4 and trait simulation in Option 28. These are stochastic operations that depend on the generation of random numbers. If you want to rerun an analysis with the same data but different random results, try resetting the seed of Mendel's random number generator by issuing a command such as

```
SEED = 17237
```

in the control file. You may reset the keyword SEED to any integer between 1 and 30,000. Using a fixed value for SEED ensures that each Mendel run with the same input data will generate the same results. Alternatively, you may set SEED to the value "Time", which is not case sensitive. This will force Mendel to use an almost surely unique random number seed during each run. In this case, each stochastic run will produce unique, non-reproducible results.

0.12 Parallelization

Several of Mendel's analysis options can simultaneously use multiple computer processors to speed-up their computations. This is accomplished by using highly optimized linear algebra packages, such as BLAS and LAPACK, and by directly programming some sections of the code to run multi-threaded, i.e., in parallel. Mendel can make use of all processors within a computer that share access to a common pool of memory. However, if you wish to limit or turn-off this parallelization, then set the keyword MAX_THREADS to the number of processors that Mendel should use. For example,

```
MAX_THREADS = 1
```

in the control file will turn-off all parallelization. By default, Mendel will use all available processors for maximum speed. The results will be the same no matter how many processors are used, only the run-time will change.

0.13 Germane Keywords

Keywords naming the input and output files:

```
DEFINITION_FILE
MAP_FILE
PEDIGREE_FILE
PENETRANCE_FILE
SAMPLE_SUBSET_FILE
```

```
SNP_DATA_FILE
SNP_DEFINITION_FILE
SNP_PHASE_FILE
SNP_SUBSET_FILE

NEW_DEFINITION_FILE
NEW_MAP_FILE
NEW_PEDIGREE_FILE
NEW_PENETRANCE_FILE
NEW_SNP_DATA_FILE
NEW_SNP_DEFINITION_FILE
NEW_SNP_PHASE_FILE
OUTPUT_FILE
PLOT_FILE
SUMMARY_FILE
```

Keywords determining how to read the input files:

```
DEFAULT_LIST_READ
DEFINITION_LIST_READ
MAP_LIST_READ
PEDIGREE_LIST_READ
PENETRANCE_LIST_READ

ALLELE_SEPARATOR
FEMALE
INPUT_FORMAT
MALE
MAP_DISTANCE_UNITS
MISSING_VALUE
MISSING_QUANTITATIVE_VALUE
ORDERED_ALLELE_SEPARATOR
PEDIGREE_MAX_LINE_LEN
POPULATIONS
READ_PEDIGREE_COPIES
READ_PEDIGREE_RECORDS
```

Keywords directly altering the output files or screen:

```
ECHO
KEEP_HIDDEN_FILES
MULTIPLE_PLOT_FILES
STATS_BY_SEX
TITLE
VERBOSE
```

Keywords that manage the interpretation of the input data:

```
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
DEFAULT_PENETRANCE
DELETE_PROBAND_FIELD
GRID_INCREMENT
INDICATOR_THRESHOLD
INTERIOR_POINTS
MAP_CONVERSION
MAP_CONVERSION_FEMALE
MAP_CONVERSION_MALE
MAX_MAF
MIN_MAF
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
PARAMETER_EQUATION
PARAMETER_INITIAL_VALUE
PARAMETER_MAX
PARAMETER_MIN
PENETRANCE
POPULATION_FACTOR
PRETRIM_PEDIGREES
PROBAND
PROBAND_FACTOR
RESTORE_PROBAND_FIELD
SNPS_TYPED
SNPS_TYPED_FACTOR
TRANSFORM
UNIFORM_WEIGHTS
```

Keywords that manage the maximum likelihood search procedure:

```
COMPLEXITY_THRESHOLD
CONVERGENCE_CRITERION
CONVERGENCE_TESTS
DEVIANCE_VARIABLE
DEVIANCES
GENE_FLOW_CUTOFF
MAX_ADJUSTED_MEIOSES
MAX_ITERATIONS
MAX_STEP_LENGTH
MAX_STEPS
TRAVEL
```

Keywords that manage stochastic operations and parallelization:

SEED
MAX_THREADS

Random Quotes

I know no method to secure the repeal of bad or obnoxious laws so effective as their stringent execution.

Ulysses S. Grant

When getting someone else to do a task is more work than just doing it yourself, do it yourself.

Susan Elgin in *The Grandmother Principles*

In truth, it is not knowledge, but learning, not possessing, but production, not being there, but traveling there, that provides the greatest pleasure. When I have completely understood something, then I turn away and move on into the dark; indeed, so curious is the insatiable man, that when he has completed one house, rather than living in it peacefully, he starts to build another.

Karl Friedrich Gauss, in a letter to W. Bolyai

I do not know what I may appear to the world; but to myself I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay undiscovered before me.

Isaac Newton in his memoirs

If you would not be forgotten,
As soon as you are dead and rotten,
Either write things worthy reading,
Or do things worth the writing.

Benjamin Franklin

No experiment should be believed until it has been confirmed by theory.

Arthur S. Eddington

Other things being equal, finish the job that is nearest done.

Fred Mosteller

Would I lay down my life to save my brother? No, but I would to save two brothers or eight cousins.

J B S Haldane

1 Analysis Option 1: Mapping Markers

1.1 Background

Analysis Option 1 orders a set of markers and estimates the distances between adjacent marker pairs, assuming that the markers lie on the same chromosome. Genetic distance is proportional to the expected number of recombination events per meiosis separating two loci. For reasons that geneticists do not fully understand, genetic distance is not proportional to physical distance in base pairs. In addition to this mystery, there are also unexplained differences between female and male recombination rates and the complicating feature of chromatid interference. In adopting Haldane's Poisson mapping model, Mendel ignores chromatid interference and supposes recombination events fall independently on disjoint intervals.

The recombination events displayed in typed pedigrees are the raw material for gene mapping. With just two markers, estimation of genetic map distances is largely a matter of counting recombinants and non-recombinants and converting recombination fractions into genetic distances via Haldane's formula. With more than two markers, there are different competing orders, and estimation of recombination fractions becomes more difficult for any given order. Because the potential number of orders, $n!/2$ for n markers, increases so rapidly, model 1 of Option 1 will try all possible orders for only six or fewer markers. With more than six markers, you will probably want to invoke models 2 or 3. If you insist on using model 1, then candidate orders can be input one by one and compared through their maximum LOD scores. Here LOD stands for "Logarithm of the ODds" to the base 10. The numerator in the odds ratio is the likelihood (probability) of the pedigree data for a given set of recombination fractions between adjacent marker pairs. The denominator in the odds ratio is the likelihood of the pedigree data assuming the markers are completely unlinked. Mendel's maximum likelihood procedure climbs the likelihood mountain and finds the most likely pattern of recombination fractions for a given marker order.

Model 2 of Option 1 computes LOD scores for all marker pairs common to both the definition and map files. In contrast to model 1, model 2 operates in either grid or search mode. If you elect grid mode, then you can specify a standard grid of points over which Mendel will compute LOD scores. Because execution of model 2 is usually much faster than model 1, the default model, you may want to try model 2 first to gain a rough idea of the order of the markers. Model 3 incorporates model 2, but does much more. Model 3 takes the pairwise estimates of the recombination fractions and identifies those orders that give roughly the smallest total map length in a sex-averaged sense. This is accomplished via simulated annealing [87] using the preliminary ranking procedure of Falk [33]. The candidate orders identified by simulated annealing are then ranked by a full multipoint analysis exactly as with model 1. The difference is that model 3 focuses only on the

promising candidate orders and ignores the rest.

1.2 Appropriate Problems and Data Sets

Analysis Option 1 is pertinent when reliable map information is unavailable. If you want to situate a new marker on a pre-existing map of high reliability, then Option 2 for computing location scores is more relevant than Option 1. Of course, now that the complete sequence of the human genome is available, all human markers can in principle be ordered with certainty. However, order does not automatically convey genetic distances. Even more pertinent is the fact that most animal and plant species have not been sequenced. Incorrect maps and genotyping errors complicate the process of mapping trait genes. Many geneticists are disheartened to learn that as they add more markers to a promising candidate region, genetic distances tend to inflate, and multipoint evidence for linkage declines. These anomalies are caused by unreliable maps and error-prone genotyping. Analysis Option 1 addresses the first issue and Analysis Option 5 the second issue.

1.3 Input Files

Analysis Option 1 requires typical definition, map, and pedigree files. Remember to name a summary file in the control file. In model 1, the keyword `TRAVEL` should take its default value of SEARCH. Section 0.9.3 describes Mendel's capacity for maximum likelihood estimation. If you want to estimate different female and male recombination fractions, set the keyword `GENDER_NEUTRAL` to false. Standard errors of the estimated recombination fractions can be recovered setting the keyword `STANDARD_ERRORS` to true.

To select grid mode in model 2, set `TRAVEL` to GRID. If the keyword `STANDARD_GRID` is set to true, then Mendel will evaluate LOD scores at the preselected recombination fractions 0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, and 0.5. If you do not specify the standard grid, then Mendel evaluates the LOD score curve at $n + 2$ equally spaced recombination fractions between 0.0 and 0.5. Here n is the number of interior points, which has default value set by the keyword `INTERIOR_POINTS`.. In either case of grid mode, Mendel assumes equal female and male recombination fractions.

1.4 Examples

Our first example for this analysis option uses the input files Control01a.in, Def1a.in, Map1a.in, and Ped1a.in. Mendel1a.out and Summary1a.out are the output files. Two of the three markers are the blood groups Rh and Radin; the third marker is the enzyme PGM1 [74]. These data are mainly of historical interest. Even though the markers have

codominant alleles, the definition file spells out in detail the relationship between phenotype and genotype. Later examples will demonstrate the shorthand of using genotypes in the pedigree file instead of phenotypes. We could include or exclude echoing the input data in the standard output file by resetting the value of the keyword ECHO. See Section 0.9.2 for details.

The most relevant output occurs at the bottom of Summary1a.out:

```
ORDER NUMBER = 1,  MAXIMUM LOD SCORE = 4.87914.

MODEL LOCI   XX THETA    XY THETA  XX MORGANS  XY MORGANS

 1  RADIN
             0.18627     0.18627    0.23304     0.23304
 2  PGM1
             0.31366     0.31366    0.49352     0.49352
 3  RH

ORDER NUMBER = 2,  MAXIMUM LOD SCORE = 9.43973.

MODEL LOCI   XX THETA    XY THETA  XX MORGANS  XY MORGANS

 1  RH
             0.13383     0.13383    0.15576     0.15576
 2  RADIN
             0.20523     0.20523    0.26420     0.26420
 3  PGM1

ORDER NUMBER = 3,  MAXIMUM LOD SCORE = 7.61599.

MODEL LOCI   XX THETA    XY THETA  XX MORGANS  XY MORGANS

 1  PGM1
             0.29399     0.29399    0.44335     0.44335
 2  RH
             0.11623     0.11623    0.13228     0.13228
 3  RADIN

THE BEST ORDER IS ORDER NUMBER 2 ABOVE.
```

Here all $3 = 3!/2$ orders are displayed together with their maximum LOD scores in log base 10 units and their best estimates of the recombination fractions (θ's) between the adjacent pairs of loci. Female recombination fractions are prefaced by XX and male recombination fractions by XY. Corresponding genetic distances are given in Morgans rather

than centiMorgans. Choosing Morgans emphasizes the fact that recombination fractions and genetic distances are nearly the same on short intervals. We prefer the order Rh–Radin–PGM1 because its maximum likelihood is almost two orders of magnitude higher than the next best order PGM1–Rh–Radin. In general, the best order is almost certainly the correct order when its maximum likelihood exceeds the maximum likelihood of the next best order by three or more orders of magnitude.

Even with six or fewer loci, it is possible to consider a single order rather than march through all possible orders. The command

```
ALL_MAP_ORDERS = False
```

in the control file instructs Mendel to consider only the order specified by the map file. If a single order is desired, it is advantageous to omit the line

```
STANDARD_ERRORS = False
```

in the control file. Mendel then reverts to the time-consuming default of computing asymptotic standard errors (precision) of the estimated parameters. The asymptotic standard errors are listed immediately after the search iterations in the standard output file. Finally, you can permit male recombination fractions to differ from female recombination fractions by inserting the command

```
GENDER_NEUTRAL = False
```

in the control file. In grid mode, this command is ignored to avoid the awkwardness of defining a two-dimensional grid.

Our second example shows models 2 and 3 in action on five X-liked markers. The results

```
1ST MARKER  2ND MARKER  MAXIMUM     FEMALE
   NAME        NAME    LOD SCORE    THETA

 DXS6810     DXS7132    1.59870    0.14491
 DXS6810     DXS6800    0.31781    0.31016
 DXS6810     DXS6789    0.10410    0.30782
 DXS6810     DXS6797    0.02359    0.42812
 DXS7132     DXS6800    3.08994    0.05219
 DXS7132     DXS6789    1.08956    0.16218
 DXS7132     DXS6797    0.63783    0.20404
 DXS6800     DXS6789    2.32177    0.07636
 DXS6800     DXS6797    1.28034    0.19585
 DXS6789     DXS6797    3.33039    0.04571
```

from the top of Summary1b.out display the maximum LOD scores and corresponding recombination fractions for all 10 marker pairs. Male recombination fractions are omitted because they are irrelevant for X-linked markers. If you set TRAVEL to GRID and STANDARD_GRID to true in the control file and rerun this example, then Summary1b.out will display LOD scores at all standard grid points. If you omit STANDARD_GRID and define INTERIOR_POINTS instead, then the output in Summary1b.out winds up looking like that just displayed, except that the maximum is taken only over the grid points encountered. In any case, all calculated LOD scores are listed in the detailed output file Mendel1b.out.

Model 2 stops at this point. If you elect model 3 instead, the summary file will display the results of multipoint analysis for each of the candidate orders identified by simulated annealing under the Falk criterion. You can change the default number of candidate orders by inserting a command such as

```
CANDIDATE_ORDERS = 50
```

in the control file. For this example data set under model 3, at the bottom of Summary1b.out, you find the output

```
   BEST ORDERS ENCOUNTERED

  ORDER  MAXIMUM LOD SCORE

     8       10.9477
     3       10.0671
     4        9.3883
     6        9.3852
     9        9.3652
     5        9.2503
     1        8.6583
     7        8.5048
     2        8.4433
    10        8.3096
```

ranking the candidate orders by their maximum LOD scores. The orders are numbered to coordinate with the order-by-order output immediately above this message in the summary file. Notice here that more than 10 candidate orders are mentioned. In fact, model 3 attempts to improve on the best candidate order by subjecting it to all possible path reversals. For example, if the loci are numbered 1 through 5 and (1,2,3,4,5) is the best order among the candidate orders, then Mendel will check orders such as (1,4,3,2,5) arising from a reversal of the segment (2,3,4).

1.5 Germane Keywords

```
ANALYSIS_OPTION = Mapping_Markers
ALL_MAP_ORDERS
CANDIDATE_ORDERS
GENDER_NEUTRAL
GENE_FLOW_CUTOFF
INTERIOR_POINTS
MODEL
STANDARD_ERRORS
STANDARD_GRID
TRAVEL
```

Random Quotes

To explain all nature is too difficult a task for any one man or even for any one age. 'Tis better to say a little with certainty and leave the rest for others that come after you.

Isaac Newton

I still often go for walks on the (Appalachian) trail near my home, especially if I am stuck on something I am working on. Most of the time I am sunk in thought, but at some point on each walk there comes a moment when I look up and notice, with a kind of first-time astonishment, the amazing complex delicacy of the woods, the causal ease with which elemental things come together to form a composition that is — whatever the season, wherever I put my besotted gaze — perfect. Not just very fine or splendid, but perfect, unimprovable.

Bill Bryson in *A Walk in the Woods*

Professors are inclined to attribute the intelligence of their children to nature and the intelligence of their students to nurture.

Roger Masters

"I thought the number of my taxicab was 1729. It seemed to me a rather dull number." To which Ramanujan replied, "No, Hardy! No, Hardy! It is a very interesting number. It is the smallest number expressible as the sum of two cubes in two different ways."

G.H. Hardy and Ramanujan as quoted by C.P. Snow in the Introduction to *A Mathematician's Apology* by Hardy

Life is good for only two things, discovering mathematics and teaching mathematics.

Simon Poisson

2 Analysis Option 2: Location Scores

2.1 Background

The marker map is usually taken as given in linkage analysis. To position a trait locus on the marker map, we slide the trait locus along the map and look for the point maximizing the joint likelihood of the trait and markers [72]. Distance along the map is measured in genetic map units rather than in recombination fractions. The location score curve plots the joint loglikelihood of the trait and markers (base 10) standardized by its value when the trait locus is infinitely distant from the markers. Ordinarily, the method of location scores invokes the fiction of equal female and male recombination fractions, but this assumption is not necessary. For that matter, neither are the assumptions of full trait penetrance and genetic homogeneity. For the sake of simplicity, Analysis Option 2 does invoke Haldane's Poisson model of crossing over [59]. Haldane's model makes it possible to ignore genetic interference and is a requirement of the Lander-Green-Kruglyak algorithm of likelihood computation. If the trait is simply another marker, and it is genotyped on the same set of pedigrees that provide the existing marker map, then you should probably revert to Analysis Option 1. In this situation, the trait locus may bring substantial information to bear on marker order and intermarker distances.

To deal with genetic heterogeneity, Analysis Option 2 assumes the usual admixture model in which a fraction α of all pedigrees are explained by a trait locus in the vicinity of the current markers. No pedigree is permitted to harbor both linked and unlinked disease genes. If you suspect this is happening in one of your very large pedigrees, then you might want to break the pedigree into different branches or exclude it from analysis altogether. This is, of course, a delicate matter since excluding unlinked pedigrees is always bound to increase the evidence for linkage. Let your statistical conscience be your guide and always report such maneuvers honestly.

The admixture parameter α (proportion of linked markers) can be either estimated or fixed. Its default value is 1. If $\alpha < 1$, then it is useful to compute the posterior probability of each pedigree being linked. Option 2 extracts these posterior probabilities at the most likely position of the trait locus. Reduced penetrance of the trait locus can be handled via a penetrance file, as explained in Section 0.5.7.1, or via `PENETRANCE` commands in the control file, as explained in Section 0.5.7.2.

Analysis Option 2 operates by sliding the trait across the marker map. Mendel generates a segment of the overall location score curve marker interval by marker interval. Each segment of the curve is evaluated on a grid of points or optimized on the underlying interval. Due to the computational complexity of handling large numbers of markers simultaneously, Option 2 asks users to specify how many flanking markers are to be included in each likelihood evaluation. The markers involved in a likelihood computation always flank

the current interval as symmetrically as possible. If the number of markers involved is odd, then a central interval will have one more right flanking marker than it has left flanking markers. The leftmost interval will involve only right flanking markers, and the rightmost interval will involve only left flanking markers. In the extreme case of a single flanking marker at a time, Option 2 reverts to ordinary LOD scores and replaces genetic map distances by recombination fractions in the output files. Analysis Options 1 and 2 overlap in this circumstance, but Option 2 is capable of dealing with genetic heterogeneity.

2.2 Appropriate Problems and Data Sets

The method of location scores has a good track record in mapping rare Mendelian disease genes. It applies to sib pairs, nuclear families, and large pedigrees. Unlike case/control association testing, it does not apply to isolated individuals or to parent-offspring trios. For complex traits, many statistical geneticists recommend computing LOD and location scores using models of dominant, additive, and recessive inheritance with reduced penetrance. Analysis Option 4 implements nonparametric linkage analysis and is a competitor of Option 2 for such traits. The published debate [43, 113] emphasizes the advantages of each method. Successful use of Option 2 depends on having an accurate marker order and decent estimates of map distances and trait and marker allele frequencies. All genotyping errors should be corrected prior to analysis.

2.3 Input Files

Analysis Option 2 imposes certain restrictions on the input files. Mendel sets all recombination fractions between loci to be at least 0.0001. The trait locus, which is being mapped, should be defined using a command such as

```
AFFECTED_LOCUS_OR_FACTOR = trait-locus-name
```

in the control file. Despite the name of the keyword, in a location score analysis the trait must be a locus, not a factor, and have values specified in both the pedigree and definition files. As mentioned earlier, during analysis the trait locus is slid along the marker map, starting to the left of the leftmost marker and ending to the right of the rightmost marker. The number of analysis points and distances used for these flanking intervals can be set using commands such as `FLANKING_POINTS = 4` and `FLANKING_DISTANCE = 0.05`, or for sex-specific distances

```
FLANKING_DISTANCE_FEMALE = 0.08
FLANKING_DISTANCE_MALE   = 0.03
```

The flanking distances must be in the units specified by keyword MAP_DISTANCE_UNITS. (If the keyword AFFECTED_LOCUS_OR_FACTOR is not defined, then Analysis Option 2 will map the first locus in the map file against all other model loci. The distances and points separating the first locus from the second will be used as the flanking distances and analysis points.)

Analysis Option 2 has a large number of potential keywords. Inspection of the contents

```
!
!  File Names
!
DEFINITION_FILE = Def2a.in
MAP_FILE = Map2a.in
MAP_DISTANCE_UNITS = cM
PEDIGREE_FILE = Ped2a.in
OUTPUT_FILE = Mendel2a.out
SUMMARY_FILE = Summary2a.out
!
!  Analysis Options
!
ANALYSIS_OPTION = Location_scores
AFFECTED_LOCUS_OR_FACTOR = RADIN
NUMBER_OF_MARKERS_INCLUDED = 1
TRAVEL = GRID
DEVIANCES = True
STANDARD_GRID = True
!GRID_INCREMENT = 5
!FLANKING_DISTANCE = 25
!INTERIOR_POINTS = 19
LINKED_PROPORTION = 0.5
PROBABILITY_PEDIGREE_LINKED = True
!GENDER_NEUTRAL = False
!ESTIMATE_LINKED_PROPORTION = True
```

of Control02a.in shows that some keyword assignments have been nullified by a beginning exclamation point. The summary file captures Mendel's synopsis of Option 2 analysis results.

The keyword NUMBER_OF_MARKERS_INCLUDED determines the number of markers included in each sliding window of analysis and has default value 1, so technically there is no need to assign it in the control file when you want to compute LOD scores rather than location scores. Using too may markers in each window will slow Mendel down and may force it to omit the most demanding pedigrees from analysis. Hence, you may want to experiment first with low values such as 2 or 3 for the keyword NUMBER_OF_MARKERS_INCLUDED. Mendel

accepts the value ALL for this keyword, which will result in a single analysis run using all the markers simultaneously.

If you increase the complexity threshold and the maximum number of adjusted meioses as described in Section 0.10.1, then you may be able to recover some of the omitted pedigrees for analysis purposes. You may notice discontinuities in the location score curve at the marker loci. If there is a definite recombinant between the trait and a marker, then the location score curve will plunge to -∞ as the trait approaches the marker. In the absence of a definite recombinant, discontinuities may still occur because different sets of markers and/or pedigrees figure in the likelihood computations on each interval. As the number of flanking markers increases and the pedigrees used on the various intervals coincide, these second kinds of discontinuities will disappear.

The keyword TRAVEL specifies whether the likelihood is evaluated over a grid of points or optimized. Section 0.9.3 describes Mendel's capacity for maximum likelihood estimation. If TRAVEL takes its default value of SEARCH, then maximum likelihood estimation is performed. If you want to estimate different female and male recombination fractions or genetic map locations for the trait locus, set the keyword GENDER_NEUTRAL to false.

In grid mode, chosen by setting TRAVEL equal to GRID, there are three ways of selecting the grid. With LOD scores (one marker in addition to the trait locus in each sliding analysis window), the keyword STANDARD_GRID imposes the standard grid, and results in likelihood evaluation at the recombination fractions 0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, and 0.5. If the standard grid is not specified for LOD scores in grid mode, then Mendel evaluates the LOD score curve at $n + 2$ equally spaced recombination fractions between 0.0 and 0.5. Here n is the number of interior points determined by the value of the keyword INTERIOR_POINTS. For location scores (more than one additional marker per analysis window), Mendel evaluates the location score curve at $n + 2$ equally spaced map points between every pair of adjacent markers. The first and last of these grid points nearly coincides with the markers on the left and right, respectively. Note that it is possible to divide a marker interval into equal subintervals even when the interval's female and male map lengths differ. You may want to vary the value of n from marker interval to marker interval if the spacings between markers are far from uniform. For one or more particular intervals, you can override the universal n determined by INTERIOR_POINTS by following the instructions in Section 0.5.5 for creating the map file. Alternatively, you can specify an evenly spaced grid of map points through the keyword GRID_INCREMENT, which has the units specified by keyword MAP_DISTANCE_UNITS. For the intervals outside the left and rightmost markers, the number of analysis points and distances can be set using the keywords FLANKING_POINTS and FLANKING_DISTANCE, or for sex-specific distances FLANKING_DISTANCE_FEMALE and FLANKING_DISTANCE_MALE. The flanking distances must also be in the units specified by keyword MAP_DISTANCE_UNITS. Obviously, the conventions

about the standard grid, grid increment, and interior points do not apply in search mode.

Setting the keyword ESTIMATE_LINKED_PROPORTION equal to true enables estimation of the admixture parameter α in search mode. To compute the posterior probability that each pedigree is linked, set the keyword PROBABILITY_PEDIGREE_LINKED equal to true. The keyword LINKED_PROPORTION determines the fixed or initial value of α. If α is estimated, then its converged value is employed in computing the posterior probability of linkage for each pedigree.

2.4 Examples

Our first Option 2 example uses the same data set as the Option 1 example. The trait locus is the blood group antigen Radin. Summary2a.out records the following LOD scores for Radin

```
                    LOD SCORES AT STANDARD GRID POINTS

MARKER      MAX    0.00   0.01   0.05   0.10   0.15   0.20   0.30   0.40   0.50

PGM1       3.47    0.61   2.28   3.33   3.47   3.38   3.19   2.37   1.14   0.00
RH         6.12    5.16   5.98   6.12   5.85   5.42   4.82   3.23   1.42   0.00

THE HIGHEST LOD SCORE:    6.1153
OCCURS WITH MARKER: RH.
```

versus the markers PGM1 and Rh on the standard grid.

If we use a different grid or optimize the LOD score function over the recombination interval $[0, 1/2]$, then it is cumbersome to list LOD scores for so many points. Summary2a.out therefore reports maximum LOD scores in the form

```
MARKER     MAXIMUM     FEMALE       MALE          LINKED
NAME       SCORE       THETA        THETA       PROPORTION

PGM1       3.4694      0.0938       0.0938       0.5000
RH         6.1633      0.0305       0.0305       0.5000

THE HIGHEST LOD SCORE:    6.1633
OCCURS WITH MARKER: RH.
```

Regardless of how we sample the interval $[0, 1/2]$, the LOD score results favor closer linkage of Radin to Rh than to PGM1.

If we perform a location score analysis using all the markers, a grid increment of 5 cM, and flanking distances beyond the markers of 25 cM, then the summary file delivers the results:

```
                                                    LOCATION
POINT    MARKER            CHR         NEUTER       SCORE
NUMBER    NAME             NAME      MAP (cM)     (LOG-10)

   1     --                -          -25.000       3.375
   2     --                -          -20.000       3.497
   3     --                -          -15.000       3.546
   4     --                -          -10.000       3.525
   5     --                -           -5.000       3.333
   6     PGM1              AUTO        -0.005      -0.852
   7     PGM1              AUTO         0.005      -0.426
   8     --                -           5.000       5.079
   9     --                -          10.000       6.974
  10     --                -          15.000       7.917
  11     --                -          20.000       8.394
  12     --                -          25.000       8.501
  13     --                -          30.000       8.176
  14     RH                AUTO       34.995       5.673
  15     RH                AUTO       35.005       5.370
  16     --                -          40.000       7.295
  17     --                -          45.000       7.011
  18     --                -          50.000       6.646
  19     --                -          55.000       6.207
  20     --                -          60.000       5.730

THE BEST LOCATION SCORE OF  8.5014
OCCURS AT POINT   12.
```

One can easily graph the location score curve from these numbers.

Finally, if we elect to maximize the location score curve on each interval, then Mendel produces in this example the summary output

```
MAX LOCATION SCORE =   8.8735       LINKED PROPORTION =   0.8316

  MODEL  LOCUS            RECOMBINATION FRAC   DISTANCE IN MORGANS
  LOCUS  NAME              FEMALE    MALE       FEMALE    MALE

      1  PGM1
                          0.18584   0.18584    0.23236   0.23236
      2  RADIN
                          0.10483   0.10483    0.11764   0.11764
      3  RH
```

for the interval flanked by the markers PGM1 AND RH. This is the best interval as noted in

the summary file. Note here that α is being jointly estimated with the best map position on each interval.

After every problem, Mendel2a.out lists, if asked, each pedigree's posterior probability of being linked and its contribution to the maximum LOD score. For marker PGM1, these appear as

```
POSTERIOR PROBABILITIES OF PEDIGREES BEING LINKED:

PEDIGREE      LOD          POSTERIOR
  NAME       SCORE        PROBABILITY

KUS          2.7838          0.9992
KRA          1.2049          0.9688
NEU         -0.2693          0.0705
STO         -0.1675          0.2647
BOD         -0.0841          0.3932
```

near the bottom of Mendel2a.out. Note the strong positive correlation between contributed LOD scores and posterior probabilities. The second form of the summary file displayed earlier reminds you that the prior proportion of linked pedigrees is 0.5. The LOD scores reported above are proportional to pedigree deviances. As discussed in Section 0.9.4, you can compute the ordered-subset deviance statistic $D_{\max}$ by setting the keyword DEVIANCES equal to true in the control file. In this case, the output

```
PEDIGREE   PEDIGREE  DEVIANCE
 NUMBER      NAME

     1      KUS        12.8197
     2      KRA         5.5490
     5      BOD        -0.3872
     4      STO        -0.7713
     3      NEU        -1.2401

THE ORDERED-SUBSET DEVIANCE STATISTIC HAS APPROXIMATE P-VALUE 0.09660
PLUS OR MINUS 0.005908 BASED ON 10000 RESAMPLES. THE MAXIMUM SUM HAS
2 TERMS.
```

from Mendel2a.out shows that $D_{\max}$ is non-significant as anticipated. If you suspect genetic heterogeneity, say based on an early age of onset of a disease, then $D_{\max}$ can come in handy in checking whether a pedigree's LOD score is correlated with its average age of onset.

Our second and third examples for Option 2 demonstrate Mendel's ability to model incomplete penetrance as described in Section 0.5.7. This is obviously a concern for many

disease traits. Other penetrance examples are discussed in the documentation of Analysis Options 7, 14, and 22. We first demonstrate the use of the simple PENETRANCE keyword in the control file to define penetrance functions that remain constant for all individuals in the pedigree file. More detail on this construct is provided in Section 0.5.7.2. Finally, in the third example data set, we demonstrate the use of a penetrance file to define a more complex penetrance function customized for each individual.

The second example data set uses data from a study on episodic ataxia type 1[75], abbreviated EA1 in the data files. The commands

```
!
! Input Files
!
DEFINITION_FILE = Def2b.in
MAP_FILE = Map2b.in
MAP_DISTANCE_UNITS = cM
PEDIGREE_FILE = Ped2b.in
!
! Output Files
!
SUMMARY_FILE = Summary2b.out
OUTPUT_FILE = Mendel2b.out
ECHO = Yes
!
! General Analysis Parameters
!
AFFECTED_LOCUS_OR_FACTOR = EA1
AFFECTED = 2
MAX_ADJUSTED_MEIOSES = 20
!
! Location Score Analysis Options
!
ANALYSIS_OPTION = Location_Scores
PENETRANCE = 0.001 :: A/A
PENETRANCE = 0.99  :: A/B
PENETRANCE = 0.99  :: B/B
NUMBER_OF_MARKERS_INCLUDED = All
TRAVEL = GRID
GRID_INCREMENT = 0.5
FLANKING_DISTANCE = 2
FLANKING_POINTS = 3
INTERIOR_POINTS = 3
```

in Control02b.in name the data files and specify that all distances are in cM units. The trait

locus is named using the keyword AFFECTED_LOCUS_OR_FACTOR and the phenotype of the affecteds is defined using the keyword AFFECTED. Setting MAX_ADJUSTED_MEIOSES to 20, which is greater than its default, allows more pedigrees to be included in the analysis.

After setting the analysis option, the PENETRANCE keywords define a uniform penetrance function that is fixed for all individuals. Here we have used an almost dominant model with A the wildtype allele and B the disease allele, as designated in the definition file. For all individuals we set $\Pr(\text{affected} \mid \text{A/A genotype at the trait locus}) = \text{phenocopy rate} = 0.001$, and therefore $\Pr(\text{unaffected} \mid \text{A/A}) = 0.999$. We also set nearly complete penetrance, $\Pr(\text{affected} \mid \text{A/B or B/B}) = 0.99$ for the remaining two genotypes.

The remaining commands in the control file mandate that all markers simultaneously enter into the analysis and that location scores be computed on a grid of points spaced 0.5 cM apart, starting 2 cM to the left of the leftmost marker and ending 2 cM to the right of the rightmost marker. These results

POINT NUMBER	MARKER NAME	CHR NAME	NEUTER MAP (cM)	LOCATION SCORE (LOG-10)
1	--	-	-2.000	1.787
2	--	-	-1.500	1.754
3	--	-	-1.000	1.713
4	--	-	-0.500	1.660
5	S91	12	-0.005	1.589
6	S91	12	0.005	1.588
7	--	-	0.500	1.467
8	S100	12	0.995	1.298
9	S100	12	1.005	1.295
10	--	-	1.500	1.266
11	CACNL1A1	12	1.995	1.233
12	CACNL1A1	12	2.005	1.232
13	--	-	2.500	1.202
14	--	-	3.000	1.167
15	--	-	3.500	1.126
16	--	-	4.000	1.079
17	--	-	4.500	1.024
18	S372	12	4.995	0.961
19	S372	12	5.005	1.026
20	--	-	5.500	2.124
21	--	-	6.000	2.316
22	--	-	6.500	2.363
23	--	-	7.000	2.311
24	--	-	7.500	2.109

```
  25    pY2-1            12          7.995       0.546
  26    pY2-1            12          8.005       0.301
  27    --               -           8.500       0.332
  28    pY21-1           12          8.995       0.342
  29    pY21-1           12          9.005       0.342
  30    --               -           9.500       0.332
  31    KCNA5            12          9.995       0.301
  32    KCNA5            12         10.005       0.301
  33    --               -          10.500       0.311
  34    S99              12         10.995       0.301
  35    S99              12         11.005       0.301
  36    --               -          11.500       0.293
  37    S93              12         11.995       0.264
  38    S93              12         12.005       0.284
  39    --               -          12.500       1.012
  40    --               -          13.000       1.249
  41    --               -          13.500       1.386
  42    --               -          14.000       1.479

 THE BEST LOCATION SCORE OF  2.3630
 OCCURS AT POINT   22.
```

can be seen in Summary2b.out.

The pedigree data of the third example data set, as seen in Ped2c.in, record two polymorphisms in the angiotensin-1 converting enzyme (ACE) gene that may be associated with high plasma ACE activity [51]. The penetrance values given in the penetrance file Pen2c.in assume a normal density for each of the three genotypes at the locus ACE_LOC. In this example, genotypes 1\1 and 1\2 have mean 1.550, and the genotype 2\2 has mean 3.883. These genotypes share the common standard deviation 0.7148. Option 14 describes how the penetrance file Ped2c.in is constructed. The commands

```
MAP_DISTANCE_UNITS = cM
AFFECTED_LOCUS_OR_FACTOR = ACE_LOC
ANALYSIS_OPTION = Location_scores
QUANTITATIVE_TRAIT = ACE
TRAVEL = Grid
GRID_INCREMENT = 5
NUMBER_OF_MARKERS_INCLUDED = 2
FLANKING_DISTANCE = 25
```

in Control02c.in name the quantitative trait and dictate that location scores be computed on a 5 cM grid as the hypothetical ACE locus slides across the map determined by the two markers ID and GH.

2.5 Germane Keywords

```
ANALYSIS_OPTION = Location_Scores
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
DEVIANCE_VARIABLE
DEVIANCES
ESTIMATE_LINKED_PROPORTION
FLANKING_DISTANCE
FLANKING_DISTANCE_FEMALE
FLANKING_DISTANCE_MALE
FLANKING_POINTS
GENDER_NEUTRAL
GRID_INCREMENT
INTERIOR_POINTS
LINKED_PROPORTION
MAP_DISTANCE_UNITS
MAX_ADJUSTED_MEIOSES
NUMBER_OF_MARKERS_INCLUDED
PENETRANCE
PROBABILITY_PEDIGREE_LINKED
QUANTITATIVE_TRAIT
STANDARD_GRID
TRAVEL
```

Random Quotes

Full fathom five thy father lies;
Of his bones are coral made;
Those are pearls that were his eyes;
Nothing of him that doth fade,
But doth suffer a sea-change
Into something rich and strange.

William Shakespeare in *The Tempest*

If people destroy something replaceable made by mankind, they are called vandals; if they destroy something irreplaceable made by God, they are called developers.

Joseph Wood Krutch

Everywhere I go, I'm asked if I think the universities stifle writers. My opinion is that they don't stifle enough of them. There's many a best-seller that could have been prevented by a good teacher.

Flannery O'Connor

3 Analysis Option 3: Pedigree Haplotyping

3.1 Background

When we haplotype pedigrees, we reconstruct the flow of multiple linked markers through a pedigree. Such reconstructions cannot be done with certainty, so we must be content with best possible guesses. Visual displays of haplotype assignments are helpful in following segregation and recombination. The usual convention is to paint each founder chromosome a different color and give each allele a unique label. As gametes pass from parents to their children, these founder chromosomes recombine and form multicolored hybrid chromosomes. Analysis Option 3 of Mendel supplies the raw material for such displays. Interpretation of the output of Option 3 will assist you in locating recombination events and haplotype segments shared by people afflicted with the same disease. If many affecteds from different pedigrees possess the same haplotype segment, then this is evidence that they descend from a common ancestor who carried a mutation covered by the segment. In principle, pedigree haplotyping can also be used to screen for false double recombinants created by genotyping errors, but Analysis Option 5 is better suited for this purpose.

Option 3 has a second more modest purpose. Embedded in Mendel is a fast algorithm for genotype elimination [64]. This algorithm reduces a person's set of compatible genotypes at a locus by examining the person's phenotype and the phenotypes of related pedigree members. Model 2 of Option 3 deposits partial results from Mendel's genotype inferencing in a new pedigree file. In particular, model 2 replaces phenotypes by genotypes and unordered genotypes by ordered genotypes whenever these are the only configurations consistent with the rest of the observed data.

3.2 Appropriate Problems and Data Sets

In creating haplotypes, model 1 of Analysis Option 3 assumes Hardy-Weinberg and linkage equilibrium. Because it employs a variant of the Lander-Green-Kruglyak algorithm [56, 57], it is capable of handling only small pedigrees as discussed in Section 0.10.1. To some extent, you can diminish the size of a pedigree by pruning irrelevant branches. For large pedigrees, we suggest that you turn to SimWalk [69, 99] for haplotyping. Although codominant markers are recommended, Option 3 does not require such. If you wish to haplotype a random sample of unrelated people at just a handful of markers, then consider applying model 2 of Option 6 to the output of Option 18. For dense SNP maps, Option 23 serves the same purpose.

The genotype elimination algorithm of model 2 operates locus by locus, can accommodate large pedigrees, and does not depend on assumptions of genetic equilibrium. How-

ever, loci permitting mutation are ignored. For SNP data Analysis Option 23 combines genotype elimination with linkage disequilibrium phasing.

3.3 Input Files

Analysis Option 3 requires typical definition, map, and pedigree files. Remember to name a summary file in the control file. Model 2 also requires naming a `NEW_PEDIGREE_FILE`; for model 1 this is optional.

3.4 Example

Our model 1 example features haplotypes for a pedigree segregating Krabbe disease, an autosomal recessive mapped to chromosome 14q [85]. The corresponding input and output files are Control3a.in, Def3a.in, Map3a.in, Ped3a.in, Ped3a.out, Mendel3a.out, and Summary3a.out. As dictated by the loci listed in the map file, only the marker loci are included in this analysis. The trait locus is omitted. At the bottom of Mendel3a.out, you will find the output

```
HAPLOTYPES FOR PEDIGREE NUMBER 1 WITH NAME Krabbe.

OPTIMAL MATERNAL/PATERNAL HAPLOTYPES

PERSON            HAPLOTYPES
NAME          (ALLELE NUMBERS)
  1         5  5  4  7  3  4  1  5
  1         4  7  3  8  2  3  2  3
  2         5  5  2  4  2  1  5  5
  2         4  5  9  6  1  1  4  4
  8         5  5  9  4  2  7  1  4
  8         5  5  9  8  2  5  5  4
  3         4  5  4  7  3  4  1  5
  3         4  5  9  6  1  1  5  5
  4         5  5  4  7  3  4  1  5
  4         4  5  9  6  1  1  4  4
  5         4  7  3  8  2  3  2  3
  5         4  5  2  4  2  1  5  5
  6         5  5  4  7  3  4  2  3
  6         5  5  2  4  2  1  4  4
  7         4  7  3  8  3  4  1  5
  7         4  5  9  6  1  1  4  4
  9         4  7  3  8  3  4  1  5
  9         5  5  9  4  2  7  1  4
```

```
OPTIMAL MATERNAL/PATERNAL SOURCES

PERSON
NAME                SOURCES
  1          1  1  1  1  1  1  1  1
  1          2  2  2  2  2  2  2  2
  2          3  3  3  3  3  3  3  3
  2          4  4  4  4  4  4  4  4
  8          5  5  5  5  5  5  5  5
  8          6  6  6  6  6  6  6  6
  3          2  1  1  1  1  1  1  1
  3          4  4  4  4  4  4  3  3
  4          1  1  1  1  1  1  1  1
  4          4  4  4  4  4  4  4  4
  5          2  2  2  2  2  2  2  2
  5          4  4  3  3  3  3  3  3
  6          1  1  1  1  1  1  2  2
  6          3  3  3  3  3  3  4  4
  7          2  2  2  2  1  1  1  1
  7          4  4  4  4  4  4  4  4
  9          2  2  2  2  1  1  1  1
  9          5  5  5  5  5  5  5  5

OPTIMAL MATERNAL/PATERNAL PHASES

PERSON
NAME                PHASES
  3          1  0  0  0  0  0  0  0
  3          1  1  1  1  1  1  0  0
  4          0  0  0  0  0  0  0  0
  4          1  1  1  1  1  1  1  1
  5          1  1  1  1  1  1  1  1
  5          1  1  0  0  0  0  0  0
  6          0  0  0  0  0  0  1  1
  6          0  0  0  0  0  0  1  1
  7          1  1  1  1  0  0  0  0
  7          1  1  1  1  1  1  1  1
  9          0  0  0  0  0  0  0  0
  9          0  0  0  0  0  0  0  0
```

In this output, we encounter first the most likely maternal (upper) and paternal (lower) chromosomes for each person. People are listed by name, while alleles are listed by number. You can recover allele names by checking the definition file or the locus data

echoed earlier in the output file. Keep in mind that the listed haplotypes are imputed. There may be other haplotypes that are equally or nearly as likely. The next block of output gives the founder source of each gene. Think of the source numbers as colors assigned to the founder genes. Each founder has two chromosomes and consequently two unique colors. Founders appear first in the output and have constant source rows. The last block of output gives phase, 0 for a contribution from a grandmaternal chromosome and 1 for a contribution from a grandpaternal chromosome. Because phase is arbitrary for founders, the output is limited to nonfounders. You can quickly see where recombination is most likely occurring by looking in the phase output for a switch from a string of 0's to a string of 1's or vice versa. The summary file, Summary3a.out, gives the same output listed by column rather than by row. This tactic permits data on large numbers of markers to be displayed without interruption. The optional new pedigree file, Ped3a.out, will provide these optimal haplotypes in pedigree input format with all genotypes ordered.

Our model 2 example continues the presidential example. The new pedigree file produced by this example, Ped3b.out,

```
Bush   , George  ,         ,         , M   ,   , a/b    , 213/217 , 1946
Bush   , Laura   ,         ,         , F   ,   , a|a    , 213|213 , 1946
Bush   , Barbara , George  , Laura   , F   ,   , a|a    , 213|213 , 1981
Bush   , Jenna   , George  , Laura   , F   ,   , a|b    ,         , 1981
Clinton, Bill    ,         ,         , M   ,   , a/b    , 213/217 , 1946
Clinton, Hillary ,         ,         , F   ,   , a/b    , 213/217 , 1947
Clinton, Chelsea , Bill    , Hillary , F   ,   , a|a    , 213|213 , 1980
```

displays inferred genotypes rather than input phenotypes at the egomania locus. The file substitutes ordered genotypes wherever possible for unordered genotypes at Marker1. To fit the file in the current text, some of the spaces implied by the pedigree and person formats at the top of the file have been deleted. Although you can easily reach the same genetic conclusions by visual inspection of the original definition and pedigree files, Def3b.in and Ped3b.in, genotype inference is very tedious on large pedigrees. Genotype inference also plays a key role in detecting genotyping errors as explained in Option 2.

3.5 Germane Keywords

```
ANALYSIS_OPTION = Pedigree_Haplotyping
MODEL
NEW_PEDIGREE_FILE
```

Random Quote

Violence does even justice unjustly.

Thomas Carlyle

4 Analysis Option 4: NPL

4.1 Background

Allele sharing (NPL) statistics are used to map susceptibility genes for complex diseases whose modes of inheritance depart from classical Mendelian patterns [55, 99, 112]. Excess sharing of marker alleles among the affected people in a pedigree in a chromosomal region suggests that the region harbors a disease susceptibility gene. For a qualitative trait, this option of Mendel computes four allele sharing statistics at points across a genotyped region and approximates their p-values by simulation. Depending on the kind of pedigrees input, each of these nonparametric statistics tends to have superior power against one or more of the specific Mendelian alternatives of recessive, additive, or dominant inheritance listed in Table 4.1. Of course, none of these alternatives holds precisely for any complex disease. The references [58, 59] discuss the rationale for these statistics and how pedigrees with different numbers of affected people are weighted. The "additive all" and "additive pairs" statistics will be familiar to users of the Genehunter or Merlin programs.

For quantitative traits, this option computes the Q-NPL statistic. Although Q-NPL is motivated by standard variance component models with random effects and fixed heritability, it avoids parameter estimation and dubious normality assumptions. This test statistic looks for excess allele sharing among individuals with correlated trait values. It is computationally efficient and applicable to general pedigrees of moderate size [25].

Table 4.1: Best Allele-Sharing Statistics for Various Alternatives

Disease Model	Pedigrees with a Single Generation of Affecteds	Pedigrees with Multiple Generations of Affecteds
Recessive	Recessive Blocks Statistic	Recessive Blocks Statistic
Additive	Additive All and Pairs Statistics	Additive All and Pairs Statistics
Dominant	Additive All and Pairs Statistics	Dominant Blocks Statistics

4.2 Appropriate Problems and Data Sets

The trait of interest should be either dichotomous, with each individual coded as affected, normal, or unknown (blank), or a quantitative variable. As a rule of thumb, try to include the marker genotypes of as many people in the pedigrees as possible; all genotypes help identify allele sharing. Of course, missing data is preferable to incorrect data, so be confident of your genotypes. Although all individuals in a pedigree contribute gene flow information,

only affected individuals directly determine the NPL statistics for qualitative traits. Thus for these traits, labeling a person as normal or unknown has no effect on the NPL statistic per se. Keep in mind, however, that listing an affected person as having an unknown disease phenotype can occasionally improve the accuracy of p-values. For example, if an affected person lacks both marker phenotypes and descendants in a pedigree, then he or she adds no information on allele sharing. The noise generated by including these dangling people among the affecteds degrades the performance of the sharing statistics. If both parents of a nuclear family lack marker phenotypes, then it is similarly harmful to include either of them among the affecteds.

On the other hand, the Q-NPL statistic for quantitative traits is not an affecteds-only statistic. If affectation status is converted to a 0/1 quantitative trait, then information on both affecteds and normals, but not unknowns, contributes to the statistic. Thus, when you compute Q-NPL statistics for 0/1 transformed traits, it is important to assign a normal trait phenotype only when you are confident in that status. If there is much doubt, assign an unknown (blank) phenotype at the trait.

4.3 Input Files

In preparing input files, follow the format requirements described in Section 0.5. Accurate map information is particularly important when dealing with biallelic markers such as SNPs. Although the size of pedigrees acceptable for analysis is limited by the constraints of the Lander-Green-Kruglyak algorithm as discussed in Section 0.10.1, the pedigree file may contain either nuclear families or extended pedigrees. Except in the case of homozygosity mapping of recessive diseases, each pedigree should contain more than a single affected person. In homozygosity mapping, only the recessive blocks statistic has any chance of detecting linkage.

To compute the above allele sharing statistics for a qualitative trait, some variation of the following command must be included in the control file

```
AFFECTED_LOCUS_OR_FACTOR = HEALTH
```

instructing Mendel where to look for the affected designator. In this example, the affected people are designated by one of the default values, AFFECTED, of the keyword `AFFECTED`. The other default value is 2. If we want to change the value of this designator, say to the value D, then the keyword phrase

```
AFFECTED = D
```

must be inserted in the control file. Only those pedigree members designated as affected directly count in computing allele-sharing statistics. People with blank or normal disease

phenotypes do not count. However, all genotyped people at least indirectly contribute sharing information on the affecteds.

To compute the Q-NPL statistic for a quantitative trait, use the command `MODEL = 2` and name the quantitative variable in the control file. For example,

```
MODEL = 2
QUANTITATIVE_TRAIT = trait01
```

instructs Mendel to calculate Q-NPL statistics on the quantitative variable trait01.

Approximation of p-values is done in two stages by two different methods. These methods share their second stage but diverge in their first stage. Let us call these the full enumeration method and the replicate pool method. The full enumeration method is the default: it is faster but more conservative. The replicate pool method is slower but can yield more impressive p-values. To activate the replicate pool method, you must give the keyword `REPETITIONS` a positive value by inserting a command such as

```
REPETITIONS = 50
```

in the control file. In the first stage, the full enumeration method constructs the entire distribution of each statistic for each pedigree, ignoring marker types. Actual NPL statistics are smoother and show less variation because they average over all underlying patterns of gene flow consistent with observed marker types.

The replicate pool method constructs a small number of replicate pedigrees by gene dropping new marker alleles and recomputing all four statistics. These replicates give an idea of the range of variation of each statistic in each pedigree. Once these small replicate pools are created, we can simulate a statistic for the entire collection of pedigrees by sampling each pool and adding the pedigree specific statistics. The full enumeration method draws from the constructed theoretical distribution for each statistic rather than from the replicate pools.

The number of trials in the second-stage simulation is determined by the keyword `SAMPLES`. In the example below, we override the default value of 10000 with the command

```
SAMPLES = 100000
```

in the control file. The second stage is computationally quick, so it pays to take a large value of `SAMPLES`, particularly prior to publication of positive findings. If none of the statistics in the sampled data are as or more extreme than the observed statistics, then the p-value is reported as less than the reciprocal of the number of samples.

To find p-values at a single intermediate point between each pair of adjacent markers, we insert the command

```
INTERIOR_POINTS = 1
```

in the control file. The results of this command are evident in the output file Mendel4.out. Depending on the distance between makers, you might want to use a larger value for INTERIOR_POINTS as discussed in Section 2.3 or define a uniform grid of points via the keyword GRID_INCREMENT as described in Section 0.5.5. If you want to rerun an analysis with the same data but different random results, you can reset Mendel's random number generator using the keyword SEED as described in Section 0.11.

4.4 Example

Our Analysis Option 4 examples examine the breast cancer families used in mapping BRCA1 [40]. You can inspect the input files Def4a.in, Map4a.in, and Ped4a.in for the details of the input data. Note that no disease locus is explicitly included in any of these files. Disease status is handled through the factor HEALTH. In our data set all unaffecteds are listed as normal, as opposed to possibly unknown. This is not realistic for the breast cancer trait. Although this will not affect the NPL statistics, it will affect the Q-NPL statistic when we convert the affection status into a 0/1 quantitative trait. To the extent that some of the "normals" are in fact affected, the Q-NPL statistic will be degraded.

In this example, and all NPL examples, there must be at least two markers common to the definition and map files. If you have data on a single marker, then add a second dummy marker with a single allele. Note the absence of genotypes at MARKER2, the dummy marker, in our pedigree file Ped4a.in. Section 0.10.1 specifically mentions Analysis Option 4 in its discussion of how to handle large pedigrees.

The summary output file Summary4a.out contains the results:

```
                        ALLELE SHARING OPTION

          P-VALUES OF NONPARAMETRIC MARKER ALLELE SHARING STATISTICS
            OBTAINED USING THE CONSERVATIVE, FULL ENUMERATION METHOD
                      AND WITH 100000 OVERALL TRIALS

LOCUS     NEUTER MAP      RECESSIVE    ADDITIVE     ADDITIVE     DOMINANT
NAME    LOCATION (cM)  BLOCKS STAT PAIRS STAT    ALL STAT    BLOCKS STAT

MARKER1        0.00       0.007240    0.000370     0.001580     0.000160
--            17.33       0.115860    0.044990     0.040600     0.062550
MARKER2       34.66       0.269350    0.195800     0.170540     0.236490
```

It is worth drawing attention to several features in this output. First, the evidence for linkage is impressive in the vicinity of Marker 1. Second, as instructed, Mendel has computed p-values at one interior point equally distance from the two markers. Third, because

of the two-stage nature of simulation, the output does not provide confidence intervals for the various p-values. Fourth, the p-values are correlated, so the output says nothing about global p-values. This is more of an issue in genome scans. Fifth, the replicate pool method yields even more striking p-values, but computation times are extremely long.

When you publish your own findings for qualitative traits, we suggest that you report all four statistics. This is more honest and informative than singling out the most significant statistic. If your pedigree file contains a reasonable number of small pedigrees, say 20 or more, then we also recommend that you redo all analyses using the replicate pool method. Keep doubling the number of repetitions until p-values stabilize.

Our second example uses the same data set but calculates the Q-NPL statistic. Since we have no quantitative variable, we must convert the qualitative factor HEALTH into a quantitative variable that has the value 1 for affecteds and 0 for unaffecteds. Unknown phenotypes remain unknown under transformation. As discussed in Section 0.5.4.6, Mendel provides an easy mechanism for this transformation. We simply use the two commands

```
AFFECTED_LOCUS_OR_FACTOR = HEALTH
TRANSFORM = INDICATOR :: Qtrait
```

in the control file to transform the factor HEALTH to the discrete variable Qtrait. Q-NPL analysis also requires the quantitative variable to be standardized, which we accomplish with the subsequent command

```
TRANSFORM = STANDARDIZE :: Qtrait
```

Of course, Qtrait is so bimodal that even after standardization it will not resemble a normal curve. This is not a problem for the Q-NPL statistic. Finally, to inform Mendel to run the Q-NPL analysis on the variable Qtrait, add

```
MODEL = 2
QUANTITATIVE_TRAIT = Qtrait
```

to the control file.

The Q-NPL results listed in Summary4b.out

```
 P-VALUES OF NON-PARAMETRIC Q-NPL STATISTIC
 OBTAINED USING THE CONSERVATIVE, FULL ENUMERATION METHOD
 WITH 100000 OVERALL TRIALS

  LOCUS     NEUTER MAP         Q-NPL
  NAME    LOCATION (cM)      P-VALUE

  MARKER1        0.00        0.000040
  --            17.33        0.038310
  MARKER2       34.66        0.199030
```

again show a highly significant p-value at MARKER1. It is interesting to note that the Q-NPL p-value is more significant than the qualitative NPL p-values even with such a rough quantitative variable. In explanation, recall that everyone not assigned an affected phenotype was cavalierly marked as normal rather than possibly unknown. We warned earlier that such a course of action might skew the results toward less significance. In fact, the pedigree data is roughly half affecteds. Thus, the affecteds-only NPL statistics are based on roughly half as much data as the Q-NPL results.

4.5 Germane Keywords

```
ANALYSIS_OPTION = NPL
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
GRID_INCREMENT
INTERIOR_POINTS
MAX_ADJUSTED_MEIOSES
MODEL
QUANTITATIVE_TRAIT
REPETITIONS
SAMPLES
SEED
TRANSFORM
```

Random Quote

> The greatest pleasure of a dog is that you may make a fool of yourself with him, and not only will he not scold you, but he will make a fool of himself too.
>
> *Samuel Butler*

> Night is a dead monotonous period under a roof; but in the open world it passes lightly, with it stars and dews and perfumes, and the hours are marked by changes in the face of Nature. What seems a kind of temporal death to people choked between walls and curtains, is only a light and living slumber to the man who sleeps afield.
>
> *Robert Louis Stevenson*

> Never put off till tomorrow, what you can do the day after tomorrow.
>
> *Mark Twain*

> Depend upon it, sir, when a man knows he is to be hanged in a fortnight, it concentrates his mind wonderfully.
>
> *Samuel Johnson*

5 Analysis Option 5: Mistyping

5.1 Background

Virtually all genotyping procedures are error prone. Because mistyping rates as low as 1% to 2% can distort linkage and association findings, detection and elimination of typing errors is of paramount importance in genetic epidemiology and genetic risk prediction. Typing errors can be divided into those that are inconsistent and those that are consistent with Mendelian transmission. The former (or Mendelian) errors can be found by examining each marker separately. The latter errors often reveal themselves as unlikely double recombinants. Analysis Option 5 is designed to find both kinds of errors. It is quick and dirty in flagging Mendelian errors, provided you are prepared to deal with considerable ambiguity in pinpointing the source of each error. It is slow and sure in flagging Mendelian consistent errors and in computing posterior mistyping probabilities for both kinds of error. Option 5 even allows you to estimate typing error rates. Users interested in the theoretical underpinnings of Option 5 may consult the article [100].

When a genotype is assigned a high posterior probability of being mistyped, it is wise to recheck and reevaluate it against its original data image, if available. If the called genotype does not change, and retyping is prohibitively expensive, then it is probably best to set the genotype as unknown. For example, it might be that one allele of the marker is amplifying improperly, and the genotype is consistently called a homozygote when it is actually a heterozygote. One should never alter a genotype, except to change it to unknown, based solely on statistical arguments. Specifying a new value other than unknown should only be based on laboratory results. Thus, all Mendelian errors must be corrected or the offending phenotypes set to unknown before further statistical analysis. Deletion of suspect phenotypes is easy if you request a new pedigree file.

5.2 Appropriate Problems and Data Sets

Applying Option 5 is a good idea before any kind of pedigree analysis is undertaken. Model 1, which detects Mendelian errors, can be carried out on both small and large pedigrees [64]. The remaining models involve Mendelian consistent errors and require small to medium-sized pedigrees. For larger pedigrees, consider using SimWalk [69, 99].

5.3 Input Files

Analysis Option 5 requires typical definition, map, and pedigree files. Table 5.1 summarizes five possible models. In model 1, detection of Mendelian errors is done locus by locus for all loci shared by the definition and map files. Allele frequencies in the definition file and

recombination fractions in the map file are irrelevant. In the four remaining models, allele frequencies and recombination fractions are pertinent.

Table 5.1: Mistyping Models

Model Number	Consistent Errors	Estimate Rates	Error Threshold	Large Pedigrees
1	No	No	No	Yes
2	Yes	No	Yes	No
3	Yes	Yes	No	No
4	Yes	Yes	No	No
5	Yes	No	Yes	No

Models 2 and 5 provide posterior probabilities of mistyping for each observed genotype. Genotypes with posterior error probabilities exceeding the threshold are flagged. You can change the threshold default of 0.25 by inserting a command such as

```
GENOTYPING_ERROR_THRESHOLD = 0.5
```

in the control file. Models 2 and 5 require an overall probability of typing error per genotype. You can modify the default value of 0.025 of this prior probability (or genotyping error rate) by inserting a command such as

```
GENOTYPING_ERROR_RATE = 0.01
```

in the control file. Suspect genotypes with error probabilities exceeding the threshold are listed in the summary file. Model 2 assumes that typing errors are uniformly distributed across available genotypes. Model 5 assumes that typing errors are distributed across available genotypes according to their population frequencies. Finally, if you name a new pedigree file with a command such as

```
NEW_PEDIGREE_FILE = Ped5b.out
```

in the control file, then Mendel will output a pedigree file with all suspect phenotypes blanked out. Under model 1, if a pedigree shows an error at a locus, then all phenotypes at that locus will be deleted in the pedigree. Under the remaining models, Mendel deletes only those phenotypes with posterior error probabilities above the threshold. We do not necessarily recommend automated phenotype deletion since it is better to find and correct errors prior to statistical analysis. However, given the time and expense of retyping, you may need to settle for less than the ideal.

Models 3 and 4 estimate error rates from typing data, which obviously should not be preprocessed to eliminate Mendelian errors. Model 3 estimates an omnibus error rate appropriate to all markers; model 4 estimates marker specific error rates. Both models assume uniformly distributed errors.

5.4 Examples

The four examples of Option 5 are all organized around the same basic data set. The output

```
PEDIGREE   PEDIGREE    LOCUS      ERROR NEAR
NUMBER     NAME        NAME       PERSON NAMED

     1      1          D21S1256       2
     7      12         D21S1256       77
     1      1          D21S1914       2
     1      1          D21S263        2
```

from the bottom of Mendel5a.out shows model 1 in action. Column 4 of this output suggests the vicinity of a Mendelian error for the cited pedigree and locus. There is no guarantee that the listed person is the source of the problem. It may be that a near relative is mistyped instead. Note that pedigrees are listed by both sequential position in the pedigree file and by name, while loci and people are listed by name only. If no inconsistencies are found, then none is reported at the bottom of the standard output file.

The second example involves model 2. The output

```
PEDIGREE   PEDIGREE     PERSON     LOCUS     PHENOTYPE       ERROR
NUMBER     NAME         NAME       NAME                   PROBABILITY

     1      1           ANYONE     D21S1256                  1.00000
     1      1           1          D21S1256  01/04           0.94253
     1      1           ANYONE     D21S1914                  1.00000
     1      1           1          D21S1914  01/02           1.00000
     1      1           ANYONE     D21S263                   1.00000
     1      1           4          D21S263   02/07           0.90936
     6      11          ANYONE     D21S263                   0.67063
     6      11          74         D21S263   01/06           0.62346
     7      12          ANYONE     D21S1256                  1.00000
     7      12          78         D21S1256  04/04           0.79051
```

from Summary5b.out gives just those posterior probabilities that fall above the threshold. A row labeled “ANYONE” provides a pedigree-wide posterior probability of typing error at the

given marker. This probability is typically spread over several people. Only those people with posterior probabilities above the threshold are listed in the summary file. A posterior probability of 1 indicates a virtually certain Mendelian error.

Model 3 is implemented in the third example. The output

```
 ITER   NSTEP   LOGLIKELIHOOD     ERROR

   1      0    -0.2136062E+03  0.1000E-01
   2      2    -0.2090457E+03  0.1165E+00
   3      0    -0.2088225E+03  0.1087E+00
   4      0    -0.2082867E+03  0.7471E-01
   5      0    -0.2082843E+03  0.7131E-01
   6      0    -0.2082838E+03  0.7233E-01
   7      0    -0.2082838E+03  0.7230E-01
   8      1    -0.2082838E+03  0.7230E-01
   9      0    -0.2082838E+03  0.7230E-01
  10      0    -0.2082838E+03  0.7230E-01

 THE MAXIMUM LOGLIKELIHOOD -0.2082838E+03 OCCURS AT ITERATION  10.

 ASYMPTOTIC STANDARD ERRORS OF THE PARAMETERS:

    ERROR

  0.3114E-01
```

at the bottom of Mendel5c.out shows that Mendel estimates an overall error rate of 0.0723. The asymptotic standard error 0.0311 of this large estimate is itself large. Lack of precision in estimated parameters is always a feature of small data sets. Section 0.9.3 describes in more detail how to interpret this kind of output on maximum likelihood estimation.

Finally, the model 4 output

```
ITER   NSTEP   LOGLIKELIHOOD    ERROR  1     ERROR  2     ERROR  3

  1      0      -213.606         0.0100       0.0100       0.0100
  2      2      -209.045         0.1165       0.1165       0.1165
  .      .          .              .            .            .
  .      .          .              .            .            .
  .      .          .              .            .            .
 17      0      -207.627         0.0815       0.0349       0.1279

 THE MAXIMUM LOGLIKELIHOOD OCCURS AT ITERATION  17.
```

```
ASYMPTOTIC STANDARD ERRORS OF THE PARAMETERS:

   ERROR  1    ERROR  2    ERROR  3

   0.0526      0.0343      0.0852

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS:

   ERROR  1    ERROR  2    ERROR  3

    1.0000
    0.0043      1.0000
    0.0015     -0.0131      1.0000
```

at the bottom of Mendel5d.out shows wide variation in estimated error rates at the three markers. However, the small difference in maximum loglikelihoods between models 3 and 4 (-208.283 versus -207.627) indicates that a single error rate common to all markers cannot be rejected. In this case, model 3 is nested within model 4, and the likelihood ratio statistic $2\times(208.283-207.627) = 1.312$ follows an approximate χ^2 distribution with $3-1=2$ degrees-of-freedom. It would take much more data to resolve the issue of whether error rates differ among the markers.

5.5 Germane Keywords

```
ANALYSIS_OPTION = Mistyping
GENOTYPING_ERROR_RATE
GENOTYPING_ERROR_THRESHOLD
MODEL
NEW_PEDIGREE_FILE
```

Random Quotes

> I had a feeling once about mathematics, that I saw it all ... but it was after dinner and I let it go.
>
> *Winston Churchill*

> If a man hasn't discovered something that he will die for, then he isn't fit to live.
>
> *Martin Luther King Jr.*

> Nothing so needs reforming as other people's habits.
>
> *Mark Twain* in *The Tragedy of Pudd'nhead Wilson*

6 Analysis Option 6: Allele Frequencies

6.1 Background

Estimation of allele frequencies is an important step in linkage analysis and association testing. Section 0.5.4.1 describes a quick method for estimating allele frequencies that treats all pedigree members as unrelated. Analysis Option 6 provides a more sophisticated procedure that uses full pedigree data and respects relationships [12]. Although the computations are more demanding, the resulting estimates are better and have attached standard errors indicating their precision. Option 6 also permits a) testing of Hardy-Weinberg equilibrium, b) comparison of allele frequencies between two or more populations, for example, cases versus controls, c) estimation of haplotype frequencies, d) testing of linkage equilibrium, and e) testing for selection pressure. Analysis Options 11 and 12 take up the two themes of equilibrium and association testing again in a nonparametric setting. Model 2 of Analysis Option 23 estimates haplotype frequencies for dense SNP maps.

In Option 6, estimation of haplotype frequencies can be done in two different ways. The more roundabout way involves preprocessing the data to construct a super-locus between closely spaced markers. Because of phase ambiguity, any super-locus that combines a sequence of nonrecombinant markers will exhibit dominance relations that must be spelled out in detail in the definition file. Fortunately, Analysis Option 18 creates the required definition and pedigree files for combined loci. The advantage of dealing with super-loci is that Mendel can exploit them in estimating haplotype frequencies with large numbers of haplotypes and in testing for distorted transmission under the gamete competition model. The more direct way of estimating haplotype frequencies avoids the explicit construction of super-loci and does not require a blanket assumption of no recombination between adjacent markers.

If there are too many alleles or haplotypes, then estimation of frequencies may be prohibitively time consuming with pedigree data. Option 6 accordingly incorporates a quick EM (Expectation-Maximization) algorithm procedure that treats all pedigree members as unrelated. This procedure is invoked by setting the keyword `MODEL` equal to 2. If you have a random sample of fully typed individuals and a moderate number of alleles or haplotypes, then we suggest that you use instead either model 1 for allele frequency estimation, the default model, or model 3 for haplotype frequency estimation. Models 1 and 3 provide better frequency estimates and asymptotic standard errors associated with these estimates. If you insist on using pedigree data to estimate allele or haplotype frequencies under models 1 or 3, and if the number of alleles or haplotypes is large, then make certain that all pedigree members are typed or consider applying Analysis Option 16 to consolidate rare alleles prior to invoking Option 6.

In haplotype frequency estimation, all three models can incorporate a Bayesian prior

that steers estimates toward linkage equilibrium [59]. This will be most useful for data sets with a small number of observations or a large number of potential haplotypes. Maximum likelihood rather than maximum *a posteriori* is the default estimation method of Option 6. Under the Bayesian alternative, all estimated haplotype frequencies are positive.

Table 6.1: Allele Frequency Estimation Models

Model Number	Alleles or Haplotypes	Pedigrees Preserved	Tests	Multiple Populations	Bayes Prior
1	Alleles	Yes	HW	Yes	Optional
2	Alleles	No	None	Yes	Optional
3	Haplotypes	Yes	LD	No	Optional

6.2 Appropriate Problems and Data Sets

As with all analysis options, markers can be autosomal or X-linked. Pedigrees and random samples of individuals are both accepted. Because each pedigree can be assigned a copy number, a random sample can be quickly summarized by listing each phenotype as a separate single-person pedigree, with a copy number indicating the number of people in the sample having that phenotype. If several populations are to be compared, then each founder in the pedigree file must be assigned to one of the possible populations. The loci common to the definition and map files are analyzed one by one in models 1 and 2. Because it is designed for haplotype frequency estimation, model 3 analyzes all model loci simultaneously and ignores population differences. Table 6.1 summarizes the different models. For lack of space, the table does not mention that testing for population heterogeneity and selection pressure are possible with model 1.

6.3 Input Files

Section 0.5.6 describes how to include pedigree copy numbers. Readers can consult the control and pedigree files of the Option 6 examples to see how this works in practice. Option 6 can also invoke three other keywords. In Control6a.in the commands

```
POPULATIONS = 2
POPULATION_FACTOR = HEALTH
```

appear, alerting Mendel to the facts that two populations are involved and population homogeneity testing should be undertaken. In testing for identical allele frequencies across

populations, each founder must be assigned to one of the permitted populations defined by the factor HEALTH. The default value of blank for the keyword POPULATION_FACTOR entails just a single population. Mendel deposits its estimated allele frequencies for the combined population in the summary file. If you want to save the estimated allele frequencies for each separate population, then include a command such as

```
NEW_DEFINITION_FILE = Def6a.out
```

in the control file. This command creates a new definition file with the estimated allele frequencies incorporated. For model 2, Mendel omits homogeneity testing and deposits the population specific allele frequency estimates in the summary file. It is still possible to output them to a new definition file as well.

Mendel automatically tests for Hardy-Weinberg equilibrium when there is only one population. Traditionally, Hardy-Weinberg tests have relied on contingency table analysis of unrelateds. To avoid this sampling straitjacket, Mendel introduces a new pedigree-based model. This model depends on a parameter γ reflecting the ratio of heterozygous genotypes to homozygous genotypes among pedigree founders. Thus, if f_i is the frequency of allele i, then the frequency of genotype i/j is

$$\Pr(i/j) = \begin{cases} \frac{f_i^2}{\sum_k f_k^2+\gamma(1-\sum_k f_k^2)}, & j = i \\ \frac{2\gamma f_i f_j}{\sum_k f_k^2+\gamma(1-\sum_k f_k^2)}, & j \neq i. \end{cases}$$

The likelihood ratio test of the null hypothesis $\gamma = 1$ permits non-codominant loci.

You can turn on the Bayesian prior for estimating allele and haplotype frequencies by inserting a command such as

```
PSEUDO_ALLELES = 10.0
```

in the control file. Use this command only when you are analyzing a single population. The command can be invoked in any option provided you are willing to settle for gene counting estimates treating related people as unrelated. Note that the value p assigned to the keyword PSEUDO_ALLELES can be either a whole number or a decimal number. The p pseudo genes are apportioned among all alleles according to their frequencies in the definition file. Thus, if we view the estimation process as gene counting, then Mendel increases the count of allele i by the fractional number pf_i, where f_i is the initial allele frequency. A prior of this form is known as a Dirichlet prior. If you want to steer haplotype frequency estimates toward linkage equilibrium, Analysis Option 18 will take the products of the allele frequencies at the participating loci to form the haplotype frequencies dictated by linkage equilibrium and insert these equilibrium frequencies in the new definition file for

the combined locus. In any event, you still have to choose your own subjective value of PSEUDO_ALLELES to impose a Dirichlet prior. The default value of 0.0 for PSEUDO_ALLELES corresponds to maximum likelihood estimation.

With two or more populations, it is instructive to test for homogeneity of allele frequencies across populations. Mendel does this by a likelihood ratio test. Mendel will also compute Wright's F_{ST} statistic quantifying the extent of population substructure. If f_{ij} represents the frequency of allele j in population i of k populations, and f_j represents the frequency of allele j across all populations, then

$$\begin{aligned} F_{ST} &= \frac{H_T - H_S}{H_T} \\ H_T &= 1 - \sum_j f_j^2 \\ H_S &= \frac{1}{k} \sum_{i=1}^{k} \left(1 - \sum_j f_{ij}^2\right). \end{aligned}$$

Here H_T is the heterozygosity across all populations, and H_S is the average heterozygosity within populations. If $f_j = \frac{1}{k} \sum_{i=1}^{k} f_{ij}$ for all alleles j, then F_{ST} almost always lies between 0 and 1. The value 0 is attained when all populations share the same allele frequencies, and the value 1 is attained when each population is monoallelic for a different allele.

For a codominant marker in a single population, Option 6 will also assess selection pressure. Here each pedigree must consist of a single person. Watterson [108] views the test statistic $W = \sum_j f_j^2$ as a random variable from the Ewens sampling distribution conditioned on the number of different alleles seen in the population [31, 32]. Different departures from neutral evolution affect Watterson's homozygosity statistic in different ways. Homozygosity tends to be low in the presence of balanced selection, population bottlenecks, and ascertainment biases that lead to an underrepresentation of rare alleles. Homozygosity tends to be high in the presence of directional selection and rapid population growth. For these reasons, Mendel reports both a lower p-value $\Pr(W \leq W_{\text{obs}})$ and an upper p-value $\Pr(W \geq W_{\text{obs}})$.

6.4 Examples

The first of the six Option 6 examples has input files Def6a.in, Map6a.in, and Ped6a.in and output files Def6a.out, Mendel6a.out, and Summary6a.out. These contain data on populations consisting of duodenal ulcer patients and controls [21]. Summary6a.out conveys the allele frequency estimates and their standard errors at the ABO locus.

```
ANALYSIS RESULTS FOR LOCUS ABO:
```

```
FREQUENCY ESTIMATES UNDER THE NULL: HOMOGENEOUS POPULATIONS

ALLELE       ESTIMATED    STANDARD
 NAME        FREQUENCY     ERROR

A             0.2335       0.0093
B             0.0588       0.0049
O             0.7077       0.0099


FREQUENCY ESTIMATES UNDER THE ALTERNATIVE: NON-HOMOGENEOUS POPULATIONS

POPULATION NAME : CASE

ALLELE       ESTIMATED    STANDARD
 NAME        FREQUENCY     ERROR

A             0.2136       0.0135
B             0.0501       0.0069
O             0.7363       0.0145


POPULATION NAME : CONTROL

ALLELE       ESTIMATED    STANDARD
 NAME        FREQUENCY     ERROR

A             0.2492       0.0127
B             0.0656       0.0068
O             0.6852       0.0135



HOMOGENEITY    DEGREES OF   ASYMPTOTIC    WRIGHT'S
CHI-SQUARE      FREEDOM      P-VALUE        FST

   7.0106          2        0.03003807     0.0097
```

The first block of output contains allele frequency estimates under the null hypothesis that the cases and controls are a single homogeneous population. Mendel next analyzes the alternative hypothesis that these are non-homogeneous populations. Frequency estimates are provided for each defined population. Finally, the homogeneity chi-square and p-value listed in the summary output file indicates if there is evidence against the null. In Sum-

mary6a.out we see there is evidence of non-homogeneity and thus an association between duodenal ulcer and alleles at the ABO locus. The low value of Wright's F_{ST} measure of population substructure is less interesting in this example than it would be with geographically distinct populations. Because of the presence of two distinct populations, testing for Hardy-Weinberg equilibrium is impossible. Finally, the output definition file Def6a.out repeats the estimated allele frequencies for the combined population in preparation for another analysis option. Ordered genotypes appear in Def6a.out rather than the original unordered genotypes of Def6a.in. This should not cause alarm because the same information is conveyed in both cases.

Our second example shows Option 6 at work estimating allele frequencies at an X-linked locus and testing for Hardy-Weinberg equilibrium in a single population. At the bottom of Summary6b.out dealing with color blindness, we find the messages

```
ANALYSIS RESULTS FOR LOCUS BLIND:

ALLELE        ESTIMATED     STANDARD
 NAME         FREQUENCY      ERROR
b              0.0772        0.0025
B              0.9228        0.0025

WATTERSON TEST REQUIREMENTS NOT MET.

 EXPECTED       EXPECTED      HARDY-WEINBERG     ESTIMATED
HOMOZYGOTES  HETEROZYGOTES      P-VALUE           GAMMA
 7780.5812     1291.4188       0.023667          4.0866
```

giving us the expected numbers of homozygous and heterozygous founders in the data. For large sample theory to apply, these expectations should not fall too low, say, not below 5. The Hardy-Weinberg test is highly significant, with heterozygotes about 4 times more likely than anticipated. In fact, these data represent an amalgamation of cases from two distinct forms of color blindness [23]. Protanopia, or red blindness, is determined by one X-linked locus, and deuteranopia, or green blindness, by a different X-linked locus. Presumably, if the data were properly separated, then Hardy-Weinberg equilibrium would hold for each locus separately. Note that Watterson's test for selection is not undertaken in this example because in females the postulated normal allele is dominant to the color blindness allele. Further output at the bottom of Mendel6b.out gives the precise value 5.119 of the likelihood ratio statistic and the one degree-of-freedom figuring in the asymptotic χ^2 approximation.

Our third example, involving data from Weir (1996) [110], illustrates the use of pseudo-alleles in steering haplotype frequencies toward linkage equilibrium. Examination of the definition file Def6c.in echoed in Mendel6c.out

```
LOCUS = Idh-Mdh  CHROMOSOME = AUTOSOME  4 ALLELES  1 PHENOTYPES
ALLELE NAMES     FREQUENCIES
     AB           0.67000
     Ab           0.13000
     aB           0.16750
     ab           0.03250
PHENOTYPES
     AaBb        2 GENOTYPES
     AB/ab
     Ab/aB
```

shows how passing to haplotypes creates phase uncertainty. Because we are treating haplotypes as alleles in this context, Mendel lists haplotype names and frequencies. Except for the doubly heterozygous phenotype AaBb noted in the definition file, all other phenotypes are safely written as genotypes in the pedigree file Ped6c.in. The double homozygote Ab/Ab is a typical genotype that involves no phase ambiguity.

At the bottom of Mendel6c.out, the output

```
THE EXPECTED FRACTION OF HOMOZYGOTES AT LOCUS Idh-Mdh IS 0.5413.

HARDY-WEINBERG TESTING IS IMPOSSIBLE AT LOCUS Idh-Mdh BECAUSE BAYES ESTIMATES ARE
INVOLVED.
```

gives the fraction of homozygotes and a warning that Hardy-Weinberg equilibrium cannot be assessed. In this case, we have imposed 10 pseudo alleles to steer allele frequency estimates. This warning is repeated in abbreviated form at the bottom of the summary file. Summary6c.out also echoes the allele frequency estimates given in Mendel6c.out, namely:

```
ANALYSIS RESULTS FOR LOCUS Idh-Mdh:

ALLELE       ESTIMATED
 NAME        FREQUENCY

AB            0.7166
Ab            0.0834
aB            0.1209
ab            0.0791

WATTERSON TEST REQUIREMENTS NOT MET.

HARDY-WEINBERG TESTING IS IMPOSSIBLE.
```

Standard errors are missing here because Mendel refuses to compute them in the presence of a prior. They can be retrieved by reverting to the default value 0.0 for the keyword PSEUDO_ALLELES in the control file. Readers are encouraged to experiment with other values for PSEUDO_ALLELES. As this number is increased, the estimates output by Mendel converge to the initial allele frequencies present in the definition file.

Our fourth example demonstrates how haplotype frequencies can be estimated without combining the participating markers. In this model 3 example, we consider three adjacent SNPs as instructed by the super-locus entries in the Def6d.in file. At these SNPs the alternative alleles are A and G, A and C, and C and T, respectively. In the definition file, Def6d.in, we leave all allele frequencies blank. Mendel fills these in by gene counting and then commences a search for the best estimates of the eight haplotype frequencies. In this example, three of the frequencies are estimated to be zero as seen in the relevant portion of the summary file, Summary6d.out,

```
MARKERS:   SNP1      SNP2      SNP3

LD CHI-SQUARE      =  11.1337
-LOG_10(P-VALUE)   =   3.0717
DEGREES OF FREEDOM =   1

HAPLOTYPE FREQUENCY ESTIMATES

FREQUENCY  STD_ERROR    HAPLOTYPE

 0.00000    0.00000      A-A-C
 0.01252    0.00888      G-A-C
 0.00000    0.00000      A-C-C
 0.00250    0.00542      G-C-C
 0.46249    0.03549      A-A-T
 0.42999    0.03501      G-A-T
 0.09249    0.02089      A-C-T
 0.00000    0.00000      G-C-T
```

Whenever a frequency estimate falls on the zero boundary, Mendel automatically assigns the estimate a standard error of zero. The haplotype attached to an estimate is constructed by alternating allele names and dashes. The likelihood ratio statistic for testing linkage equilibrium and its corresponding p-value are listed in both the standard output file and the summary file. At the bottom of the standard output file, Mendel6d.out, the lines

```
THE LIKELIHOOD RATIO STATISTIC FOR LINKAGE EQUILIBRIUM EQUALS
11.13366.  THIS CHI-SQUARE STATISTIC HAS  1 DEGREE OF FREEDOM
AND P-VALUE  0.000848.
```

basically repeat the conclusion in the summary file that linkage equilibrium fails.

Our fifth example illustrates an advantage of using model 2 of Option 6. In many situations we would like to know the most likely genotype corresponding to a given phenotype. This is just the haplotyping problem in the absence of pedigree information. The output

```
MOST LIKELY GENOTYPES FOR AMBIGUOUS PHENOTYPES:

LOCUS     PHENOTYPE      MOST LIKELY        CONDITIONAL
                           GENOTYPE        PROBABILITY

ABO         A              A/O               0.84614
ABO         B              B/O               0.95435
```

from the bottom of Mendel6e.out resolves this question for the two phenotypes A and B of the ABO blood group. The two unordered genotypes A/O and B/O make their appearance here. If the most likely genotype is homozygous or only one pair of an ordered pair of equivalent genotypes is allowed, then the most likely genotype reported is ordered. For a male at an X-linked locus, the most likely allele is reported. Phenotypes corresponding to a single genotype, either ordered or unordered, are skipped. The probability listed for the most likely genotype is conditional on the given phenotype and depends on the allele frequencies supplied in the definition file. If these frequencies are absent, then the frequencies estimated by the EM algorithm are used.

Our last example uses Coyne's data on evolution at the xanthine dehydrogenase locus of *Drosophila persimilis* [22]. The data involve 23 different alleles scattered over 60 genes at this locus. One allele has 32 representatives. The rejection of Hardy-Weinberg equilibrium due to excessive homozygosity reported in the output

```
ANALYSIS RESULTS FOR LOCUS XDH:

ALLELE       ESTIMATED    STANDARD
 NAME        FREQUENCY     ERROR

1             0.0167       0.0166
10            0.0167       0.0166
11            0.0167       0.0166
12            0.0167       0.0166
13            0.0167       0.0166
14            0.0167       0.0166
15            0.0167       0.0166
16            0.0167       0.0166
17            0.0167       0.0166
18            0.0167       0.0166
```

```
19              0.0333       0.0232
2               0.0167       0.0166
20              0.0333       0.0232
21              0.0333       0.0232
22              0.0667       0.0322
23              0.5333       0.0645
3               0.0167       0.0166
4               0.0167       0.0166
5               0.0167       0.0166
6               0.0167       0.0166
7               0.0167       0.0166
8               0.0167       0.0166
9               0.0167       0.0166

  WATTERSON       WATTERSON      MCMC
UPPER P-VALUE   LOWER P-VALUE   SAMPLES

  1.000000        0.000000        10000

 EXPECTED      EXPECTED     HARDY-WEINBERG    ESTIMATED
HOMOZYGOTES  HETEROZYGOTES     P-VALUE          GAMMA

   8.9168       21.0832      0.47615E-10       0.0414
```

from the bottom of Summary6f.out confirms a strong departure from neutrality. Note the very low estimate of γ. If you are concerned about the accuracy of the Markov chain Monte Carlo approximation to the lower and upper p-values of the Watterson statistic, then you can increase the value of the keyword SAMPLES in the control file and rerun Mendel.

6.5 Germane Keywords

```
ANALYSIS_OPTION = Allele_Frequencies
MODEL
NEW_DEFINITION_FILE
POPULATION_FACTOR
POPULATIONS
PSEUDO_ALLELES
READ_PEDIGREE_COPIES
SAMPLES
```

Random Quotes

I don't make jokes. I just watch the Government and report the facts.

Will Rogers

7 Analysis Option 7: Risk Prediction (Genetic Counseling)

7.1 Background

Analysis Option 7 computes risks to individuals in pedigrees segregating Mendelian diseases. As a conditional probability, a genetic risk involves two likelihoods, a numerator likelihood with the riskee having an disease genotype or phenotype at the disease locus and a denominator likelihood with the riskee having an unknown or non-specific phenotype at the disease locus. Depending on the problem, complicating features such as age of onset, mutation, linked markers, and biochemical tests come into play [15, 83].

7.2 Appropriate Problems and Data Sets

Option 7 will handle fairly complicated problems involving mutation, reduced penetrance, and linked markers. Many such risk prediction problems are intractable by hand on all but the smallest pedigrees. Option 7 does not provide empiric risks or theoretical risks under models for genetic heterogeneity or polygenic inheritance.

7.3 Input Files

Option 7 requires definition, map, and pedigree files. If age of onset is an issue or biochemical tests are employed, then a penetrance file is also required. As just mentioned, Mendel expects two separate pedigrees, a numerator pedigree and a denominator pedigree. To take into account mutation at the disease locus, you must specify the disease locus and the corresponding female and male mutation rates by inserting commands such as

```
FEMALE_MUTATION_RATE = 0.0001 :: DMD
MALE_MUTATION_RATE = 0.0001 :: DMD
```

in the control file. In this example, we have specified that the normal allele at locus DMD (Duchenne muscular dystrophy) mutates to the disease allele with female and male mutation rates of 10^{-4}. Mendel implicitly assumes that the first DMD allele is the normal allele and the second DMD allele is the disease allele. Back-mutation from the disease allele to the normal allele is ignored. All loci carry default mutation rates of 0.0. It is possible to include one or more mutable loci in other options, but any such assumption will slow Mendel down.

7.4 Examples

In our first example, the disease locus is cystic fibrosis. Examination of the pedigree file, Ped7a.in, shows that the denominator pedigree does not even list the riskee individual,

P15. Equivalently, one could list her with blank phenotype at the CF locus. Because cystic fibrosis is a recessive disease, individuals have one of two phenotypes, NORMAL or AFFECTED. In Def7a.in, the genotypes +/+ and +/– correspond to the NORMAL phenotype, and the genotype –/– corresponds to the AFFECTED phenotype. A novel feature of this textbook problem [83] is the presence of two inbreeding loops in the pedigree. Mendel needs no special input beyond the information in the pedigree file to do the calculations correctly. The risk of 0.037 listed in Summary7a.out differs slightly from the risk computed in the reference [83] because the latter ignores possibilities for multiple entries of the cystic fibrosis gene into the pedigree.

As a footnote to this exercise, we note that the cystic fibrosis gene was mapped and cloned long ago [111]. DNA testing is now available, so this example is mainly of historical interest. The same can be said for our next example on Duchenne muscular dystrophy [81]. In defense of the kind of risk evaluation embodied in Option 7, not all disease genes are cloned, and even when they are, not all disease mutations are known with certainty. Furthermore, genotyping of flanking markers and biochemical tests may be cheaper in genetic risk prediction than direct DNA mutation testing.

Our example on Duchenne muscular dystrophy incorporates mutation, linked flanking markers, and a carrier detection test based on creatine phosphokinase (CPK). Mutation is handled by designating the DMD locus as mutable and supplying female and male mutation rates in the control file as indicated above. These mutation assumptions are repeated for emphasis near both the top and bottom of the output file, Mendel7b.out. The frequency of the DMD disease allele – is taken to be twice the common female and male mutation rate, in accord with Haldane's theory of the balance between selection and mutation [15, 39, 83]. Equilibrium theory for unequal female and male mutation rates is covered in Holloway and Smith (1973) [45]. In X-linked examples, Mendel identifies the male hemizygous genotypes + and – with the female homozygous genotypes +/+ and –/–, respectively.

In Summary7b.out, Option 7 reports a relatively low risk of 0.037 to the boy labeled P10. This low risk is a consequence of two favorable circumstances. First, the boy has completely different alleles from his affected brother, person P9, at the two flanking markers RC8 and L128. Second, his brother is more likely to be a new mutation, and his mother, person P6, is less likely to be a carrier in view of the facts that the uncle, person P5, is normal and the aunt, person P4, has a low natural-log CPK value. The quantitative variable LN_CPK, named in the definition file Def7b.in, is assumed to be normally distributed with genotype specific means and a common standard deviation. See Option 14 for the construction of the penetrance file Pen7b.in, which is echoed in Mendel7b.out.

Finally, we observe that we can compute the probability that the mother is a carrier by slightly modifying the pedigree file. In the numerator pedigree of the pedigree file, we return the riskee to unknown disease status at the DMD locus and change his mother's

DMD normal phenotype + to the carrier genotype +/–. The denominator pedigree is left unchanged. Rerunning Option 7 produces a conditional probability of 0.504 that the mother is a carrier. The same strategy applies to other carrier assessment problems.

Our last example involves age of onset considerations. Mutations in the DFNB1 locus account for half of the autosomal recessive cases of non-syndromic deafness [38]. Onset of deafness occurs anywhere between 0 and 2 years of age in homozygotes at risk. Both branches of the hypothetical pedigree displayed in file Ped7c.in contain affected people. The couple seeking counseling, individuals 6 and 7, have one normal child of age 1 year and are planning on having a second child. This child, individual 10, is at risk because it has an affected grandmother, individual 2, in one branch of the pedigree and an affected first cousin, individual 12, in the other branch.

To compute the risk to individual 10, we construct a definition and map file containing the locus dfn with just two alleles. Two different mutations occur in the dfn locus, but little error is committed in simplifying the inheritance model to a normal allele (D) and a single disease allele (d). No phenotypes are listed in either the locus or pedigree files. Age-dependent penetrances are collected in the penetrance file Pen7c.in. The numerator pedigree of this file

```
TOP,      2,          dfn,          D-D,  0.000562
TOP,      2,          dfn,          D-d,  0.000562
TOP,      2,          dfn,          d-d,  0.95
TOP,     11,          dfn,          D-D,  0.9989
TOP,     11,          dfn,          D-d,  0.9989
TOP,     11,          dfn,          d-d,  0.02
TOP,     12,          dfn,          D-D,  0.000473
TOP,     12,          dfn,          D-d,  0.000473
TOP,     12,          dfn,          d-d,  0.02
```

focuses on the three most important people in the pedigree, the affected grandmother, the normal sibling, and the affected first cousin. The penetrances for the grandmother 2 are the conditional probabilities that a female succumbs to deafness at 0 years of age. The penetrances of the sibling 11 are the conditional probabilities that a female is disease free at age 1. Finally, the penetrances of the first cousin 12 are the conditional probabilities that a male succumbs at age 2. Two of these penetrances are discrete densities, and one is a discrete survivorship function, that is, 1 minus a discrete distribution function. Disease allele frequencies and age-dependent penetrances are ethnic specific; in this case the values are pertinent to Spain [38].

The computed risk 0.006360 is quite small. In this case the parents probably would not elect to undergo mutation testing. Although application of discrete densities and survivorship functions is problematic when only a few intervals cover the relevant ages of onset, no choice of penetrance values is apt to change the conclusion of low risk to the unborn child.

7.5 Germane Keywords

```
ANALYSIS_OPTION = Risk_Prediction
DEFAULT_PENETRANCE
FEMALE_MUTATION_RATE
MALE_MUTATION_RATE
PENETRANCE_FILE
```

Random Quotes

The question, of course, is how well this experiment has succeeded. My own point of view — which, however, does not seem to be shared by most of the people who worked with the students — is pessimistic. I don't think I did very well by the students. When I look at the way the majority of students handled the problems on the examination, I think the system is a failure. Of course, my friends point out to me that there were one or two dozen students who — very surprisingly — understood almost everything in the lectures, and who were quite active in working with the material and worrying about the many fine points in an excited and interested way.

Richard Feynman in the Introduction to *The Feynman Lectures on Physics*

The most important questions of life, are, for the most part, really only problems of probability.

Pierre Simon de Laplace

He uses statistics as a drunken man uses lamp-posts — for support rather than illumination.

Andrew Lang

The value of marriage is not that adults produce children, but that children produce adults.

Peter de Vries

We shall not flag or fail. We shall go on to the end. We shall fight in France, we shall fight on the seas and oceans, we shall fight with growing confidence and strength in the air, we shall defend our island, whatever the cost may be, we shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills; we shall never surrender.

Winston Churchill in a speech to the House of Commons

8 Analysis Option 8: Gamete Competition

8.1 Background

The gamete competition model is an application of the Bradley-Terry method of ranking [14]. The Bradley-Terry method was originally applied to problems such as ranking teams in a sports league based on the intra-league win/loss records. In genetics, alleles assume the role of teams, and transmission assumes the role of winning [50, 61]. If allele i is assigned a segregation parameter τ_i, then we can write the probability that a heterozygous parent with genotype i/j transmits allele i as the ratio

$$\Pr(i/j \to i) \quad = \quad \frac{\tau_i}{\tau_i + \tau_j}. \tag{1}$$

Because this ratio is invariant when τ_i and τ_j are multiplied by the same constant c, we impose the constraint that the most frequent allele k has $\tau_k = 1$. Mendelian segregation corresponds to the choice of $\tau_i = 1$ for all i. To test whether Mendelian segregation is true, we estimate the τ_i by maximum likelihood from pedigree data and conduct a likelihood ratio test.

Although this procedure provides an interesting method for performing segregation analysis on a genetic marker, its real promise lies in generalizing the TDT test described under Analysis Option 13 [93, 94, 95]. The gamete competition model considers both qualitative and quantitative outcomes. It also uses full pedigree data and gives an estimate of the strength of transmission distortion to affected children, allele by allele. The crux of the matter is to use the Mendelian ratio of $1/2$ for transmission to normal children and the gamete competition ratio defined by equation (1) for transmission to affected children. Option 8 makes it possible to substitute the complementary ratio

$$\Pr(i/j \to i) \quad = \quad 1 - \frac{\tau_i}{\tau_i + \tau_j} \quad = \quad \frac{\tau_j}{\tau_i + \tau_j} \tag{2}$$

for transmission to normal children. This is ill advised unless we are certain that children beyond a given age are free of the disease. For a quantitative trait, we set $\tau_i = e^{\omega_i x_k}$, where x_k is a severity index for child k, and ω_i is a parameter with value 0 under the null hypothesis of Mendelian segregation. For a quantitative trait, we constrain the most frequent allele k to have $\omega_k = 0$.

The gamete competition model is better adapted to missing data than the TDT. It is also more prone to the problems presented by ethnic stratification and infrequent alleles. The permutation version of the TDT employed in Analysis Option 13 does not rely on large sample approximations and consequently requires less vigilance about allele frequencies. As safeguards in the gamete competition model, it is usually prudent to estimate allele frequencies from the data being analyzed and to consolidate rare alleles. Poorly specified

allele frequencies can lead to spurious rejection of the null hypothesis. In examples where the null hypothesis is rejected, the most influential pedigrees can be identified by computing deviances as instructed in Section 0.9.4. It also possible to assess in Option 8 whether maternal or paternal transmission is driving rejection of the null hypothesis.

Because it is a parametric model, the gamete competition model allows more nuanced statistical analysis than the TDT. For example, with a quantitative trait we might want to assess the influence of covariates of the child k on transmission. This can be achieved by replacing his or her severity index x_k in the formula $\tau_i = e^{\omega_i x_k}$ by $x_k - \sum_l \beta_l y_{kl}$, where y_{kl} is a predictor variable such as sex or body mass index and β_l is a corresponding parameter mediating the strength of the predictor.

Finally, it is worth emphasizing that the null hypotheses differ slightly between the TDT with independent parent-offspring trios and the gamete competition model with full pedigrees. The null hypothesis for the TDT with independent trios is no association or no linkage; the null hypothesis for the gamete competition model with full pedigrees is no association and no linkage.

8.2 Appropriate Problems and Data Sets

The gamete competition model applies to pedigrees, even those with missing marker data. With too many marker alleles in Option 8, computational efficiency suffers and, as just noted, large sample statistical assumptions become suspect. We recommend consolidating alleles via Option 16 until at most eight alleles remain and each has a frequency of 0.05 or greater. If the fraction of missing data is large, ethnic stratification may come into play. One remedy is to limit analysis to a single ethnic group; another is to use ethnic-specific allele frequencies as suggested in Section 0.5.4.5. If you opt for the latter strategy, then you cannot simultaneously estimate allele frequencies and transmission parameters.

If the number of independent pedigrees is small, the gamete competition model confounds linkage and association. In this situation, Option 8 has little to offer over traditional linkage analysis. Beware of assigning over-transmitted marker alleles causal effects. They may, in fact, be merely the marker alleles linked to the disease allele in a few relevant founders. Because the gamete competition model limits analysis to a single marker at a time, it may be seriously under-powered relative to multipoint location scores. With a large number of independent pedigrees, linkage analysis is inherently incapable of detecting associated marker alleles. At the same time, the gamete competition model is unlikely to detect linkage in the absence of association. Thus, the two methods give increasingly different perspectives on disease transmission as the number of independent pedigrees increases.

8.3 Input Files

The gamete competition model operates locus by locus over all loci common to the definition and map files. For this reason, the values of all recombination fractions are irrelevant. If allele frequencies are estimated from the data, then the input allele frequencies are irrelevant as well. To accommodate super-loci constructed by combining loci, Option 8 is flexible enough to handle dominant and recessive alleles. Of course, any departure from codominant alleles must be carefully specified in the definition file. For a qualitative trait, all affected children must be labeled as such. For a quantitative trait, a severity index must be included as a quantitative variable in the variable and pedigree files. Analysis Option 16 can be used to consolidate rare alleles prior to analysis. Analysis Option 8 includes the 6 sub-options (models) summarized in Table 8.1.

Table 8.1: Models for Gamete Competition Analysis

Model Number	Allele Frequencies	Trait Type	Complementary Transmission	Predictors Allowed
1	Preset	Qualitative	Possible	No
2	Estimated	Qualitative	Possible	No
3	Preset	Quantitative	Impossible	No
4	Estimated	Quantitative	Impossible	No
5	Preset	Quantitative	Impossible	Yes
6	Estimated	Quantitative	Impossible	Yes

In implementing the models of Table 8.1, the control file deserves special comment. In models 1 and 2, some version of the commands

```
MODEL = 2
AFFECTED = ILL
AFFECTED_LOCUS_OR_FACTOR = HEALTH
```

must be inserted in the control file. The value of the keyword `AFFECTED` can be omitted if one of the default values, AFFECTED and 2, is used to label affected people. In checking for Mendelian segregation at a new marker, you may designate everyone as affected or simply set

```
AFFECTED = Everyone
```

in the control file. If you set AFFECTED equal to Everyone, then you may dispense with the superfluous factor, such as HEALTH, conveying disease status. This convention is

employed in our first example. Coding disease status as a phenotype at a locus rather than as a category of a factor can lead to some confusion. If you insist on conveying disease status in this way, then omit that locus from the map file. Otherwise, the locus will be among those analyzed.

If you set

```
GENDER_NEUTRAL = False
```

in the control file, then Mendel will test maternal and paternal distortion separately at each model locus. Transmission from members of the ignored sex is assumed to follow the usual Mendelian rules. When you put

```
COMPLEMENTARY_TRANSMISSION = True
```

in the control file under model 1 or 2, the complementary transmission probability (2) will be invoked for transmission to normal offspring. Mendel treats a person as normal if he/she has a non-blank disease status differing from the affected label. Mendelian transmission is retained for offspring of unknown (blank) disease status.

In models 3 and 4, we must indicate the trait variable. Therefore, some variation on the commands

```
MODEL = 4
QUANTITATIVE_TRAIT = ACE
```

should appear in the control file, informing Mendel which quantitative variable is the severity index. The severity index can be standardized to have mean 0 and standard deviation 1 via the command

```
TRANSFORM = Standardize :: ACE
```

in the control file. Standardization is almost always advisable.

Models 5 and 6 assess the impact of the child's covariates on transmission. The syntax for specifying these predictors is straightforward. For instance, the command

```
PREDICTOR = SEX :: ACE
```

in the control file requests Mendel to incorporate a sex effect as a predictor of the severity index ACE. This might help in deciding whether distorted transmission is limited to female or male children.

8.4 Examples

Our first example considers the marker data of Lewis et al. [74]. Summary8a.out contains the model 2 output

```
MARKER                  LEAST TRANSMITTED              MOST TRANSMITTED
NAME      P-VALUE       TAU      FREQ     ALLELE        TAU      FREQ     ALLELE

RADIN     0.35126      1.0000    0.8959    -           1.2556    0.1041    +
PGM1      0.57821      0.7788    0.1847    1-          1.2390    0.0815    2-
RH        0.62573      0.8173    0.1840    R2          1.6009    0.0203    R0
```

giving the least and most frequently transmitted allele and their estimated τ's and population frequencies. None of the gamete competition tests is significant, but the low frequency of the R0 allele at the Rh locus suggests that it might be wise to consolidate this allele with the next most infrequent allele and redo the analysis. Estimates and their standard errors for all parameters can be found in Mendel8a.out. Note that each marker requires two maximum likelihood searches.

Our second example illustrates the use of a quantitative trait. The data involve a 287 base pair insertion/deletion polymorphism in the angiotensin-1 converting enzyme (ACE) gene. The deletion allele appears to be associated with high plasma ACE activity [51]. In the current sample, ACE activity has been determined in 404 people in 69 families. Summary8b.out contains the model 4 output

```
MARKER                  LEAST TRANSMITTED              MOST TRANSMITTED
NAME      P-VALUE       OMEGA    FREQ     ALLELE        OMEGA    FREQ     ALLELE

INS     0.9946E-19    -1.3052    0.4791    1           0.0000    0.5209    2
```

confirming strong over-transmission of the deletion allele (2) to children with high ACE activity levels. Again, fuller output appears in Mendel8b.out.

Our third example builds on the second example by switching from model 4 to model 6. In the process, we introduce the predictor SEX as suggested earlier. The output

```
MARKER                  LEAST TRANSMITTED              MOST TRANSMITTED
NAME      P-VALUE       OMEGA    FREQ     ALLELE        OMEGA    FREQ     ALLELE

INS     0.9946E-19    -1.3052    0.4791    1           0.0000    0.5209    2

PREDICTOR     ESTIMATE      STD ERR      P-VALUE

FEMALE          0.08292      0.10411      0.4242980
MALE           -0.08292      0.10411      0.4242980
```

from Summary8c.out summarizes all three rounds of estimation and two separate likelihood ratio tests. The first round simply estimates allele frequencies as dictated by model 6. The second round estimates the ω_i in addition to allele frequencies. Note that in this example the impressive p-value (0.9946×10^{-19}) rejecting the null hypothesis (all $\omega_i = 0$) is the same as under model 4. The third round estimates the ω_i, allele frequencies, and the female and male regression coefficients. The p-value 0.42430 for the interaction likelihood ratio test is non-significant. A single p-value is quoted for both predictors (β's) because of the omnibus character of the test. Much fuller output, including all parameter estimates, standard errors, parameter constraints, and p-values, appears in the output file Mendel8c.out.

8.5 Germane Keywords

```
ANALYSIS_OPTION = Gamete_Competition
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
COMPLEMENTARY_TRANSMISSION
DEVIANCES
GENDER_NEUTRAL
MODEL
PREDICTOR
QUANTITATIVE_TRAIT
TRANSFORM
```

Random Quotes

> Perhaps they were of that prudent sort who are equally alarmed by virtues and vices and always preach the doctrine that perfection lies in the golden mean; they fix the golden mean as the point which they themselves happen to have reached and settle down there very comfortably.
>
> *Alessandro Manzoni* in *The Betrothed*

> Indeed, if the press were to hang a sign out like every other trade, it would have to read: "Here men are demoralized in the shortest possible time on the largest possible scale for the smallest possible price."
>
> *Soren Kierkegaard* in his journals

> Remember me when I am gone away,
> Gone far away into the silent land;
> When you can no more hold me by the hand,
> Nor I half turn to go yet turning stay.
>
> *Christina Rosetti*

9 Analysis Option 9: Pedigree Selection

9.1 Background

In the pedigree selection problem, the true pedigree connecting several people is unknown. On the basis of marker genotypes, one must decide between two or more candidate pedigrees. Traditionally, this problem has been treated as one of correctly specifying the relationship between pairs of individuals. Even in the case of paternity testing, such a perspective is too restrictive. It is the comparison of the putative father's genotypes against genotypes of both the mother and the child that is most informative. If a genetic inconsistency is found, then in the absence of typing error or mutation, the putative father can be eliminated from consideration. On the other hand, if the trio is consistent at all loci typed, then either a rare event has occurred or the putative father is the actual father. The rarity of a match can be quantified by computing either a paternity index or a non-exclusion probability.

There are obvious analogies between paternity testing and determining twin zygosity. In both cases one looks for exclusions on the basis of genotyping a large number of markers. If there are no inconsistencies, then a measure of how likely the twins are to be identical should be computed. Similar considerations apply in determining whether sibs are half sibs or full sibs in a genome scan. Plane crashes and other disasters where bodies are damaged beyond recognition provide yet another example. If the remains and a handful of each victim's relatives are genotyped, then one can assign remains to surviving families.

These special cases suggest that it is worth solving the pedigree selection problem by Bayes' rule [100]. This necessitates assigning a prior probability to each candidate pedigree. Sometimes assignment is easy. For example in zygosity testing, the prior probabilities that same-sexed twins are identical are well known for many different populations [17]. In the plane wreck problem, a uniform prior over all victims is usually indicated. Because prior probabilities of paternity are more problematic, Mendel reports a paternity index (PI). The formula

$$\frac{\text{Post(Pat)}}{\text{Post(nonPat)}} = \frac{\text{Prior(Pat)}}{\text{Prior(nonPat)}}\text{PI}$$

then converts the ratio of prior probabilities of paternity into the ratio of posterior probabilities of paternity. It is up to the judge or jury to supply the prior probabilities based on the non-genetic evidence.

In computing posterior probabilities of candidate pedigrees in paternity testing and other applications, it is useful to include the possibility of genotyping error. This permits incomplete matches to be quantified. Mendel is capable of dealing with genotyping error in this analysis option as explained in Sobel, Papp, and Lange (2002) [100].

9.2 Appropriate Problems and Data Sets

Analysis Option 9 is appropriate whenever you want to determine the correct pedigree connecting a group of people and the number of realistic choices for that pedigree is small. In order for Mendel to compute posterior probabilities, you must supply prior probabilities for each of the alternative pedigrees. The more markers typed on the pedigrees, the better the discrimination will be among the pedigrees. Option 10 comparing theoretical and kinship coefficients can help you spot suspect pedigrees. For instance, it might suggest that two people labeled siblings are really half siblings.

Analysis Option 9 has four models. Model 1, the default model, covers paternity testing assuming no genotyping error. The remaining three models are relevant to the general pedigree selection problem, not just to paternity testing. Model 2 performs pedigree selection with unlinked markers. Models 3 and 4 permit linked markers and genotyping errors. In model 3, genotyping errors are assumed to be uniformly distributed over available genotypes. In model 4, genotyping errors are assumed to occur in proportion to genotype frequencies. Models 2 through 4 need prior probabilities for each pedigree. These are passed to Mendel via pedigree copy numbers.

9.3 Input Files

Preparation of the pedigree file is crucial in Option 9. The sample pedigree file

```
PEDIGREE NUMBER 1 HAS 3 MEMBERS AND NAME .

     ID        PARENT    IDS       SEX       TWIN      ABO       ADA
  1  ACCUSED                       M                   B         1-2
  2  MOTHER                        F                   AB        1-1
  3  CHILD     ACCUSED   MOTHER    F                   B         1-2

PEDIGREE NUMBER 2 HAS 4 MEMBERS AND NAME .

     ID        PARENT    IDS       SEX       TWIN      ABO       ADA
  1  RANDOM                        M
  2  MOTHER                        F                   AB        1-1
  3  CHILD     RANDOM    MOTHER    F                   B         1-2
  4  ACCUSED                       M                   B         1-2
```

echoed in Mendel9a.out shows two alternative pedigrees connecting the same set of three individuals. In this paternity testing problem, the first pedigree contains the putative father as the actual father. The second pedigree contains a random male with all phenotypes unknown as the actual father. This random male must occur as the first person of the

second pedigree. The putative father appears as an isolated individual in the second pedigree. Input the pedigrees for your own model 1 examples in the same order. There is no need for a penetrance file or for pedigree prior probabilities. Given the nature and limited extent of the pedigree file, we recommend against letting Mendel estimate allele frequencies. Check that the allele frequencies entered in the definition file are the correct ones for the ethnic group considered.

To alert Mendel that prior probabilities are being used in models 2 to 4, the command

```
READ_PEDIGREE_COPIES = True
```

must appear in the control file. If you omit this statement or leave all pedigree copy numbers blank, then Mendel will take each pedigree to have copy number 1. Prior probabilities are assumed proportional to copy numbers. Hence, the default, where all copy numbers equal 1, entails a uniformly distributed prior across all pedigrees.

The default genotyping error rate for models 3 and 4 is 0.025. This rate can be reset to the level 0.01, say, using the command

```
GENOTYPING_ERROR_RATE = 0.01
```

in the control file.

9.4 Examples

Continuing the above paternity testing problem, the analysis output near the bottom of the output file Mendel9a.out reads:

```
LOCUS NAME    PATERNITY INDEX    EXCLUSION PROBABILITY

 ABO                1.389             0.0784000
 ADA                7.576             0.8723560

CUMULATIVE   PATERNITY INDEX     EXCLUSION PROBABILITY

                   10.522             0.8823633
```

Here we see reported a paternity index and an exclusion probability. The exclusion probability is simply the probability that a random male would be excluded by at least one of the marker tests given the genotypes of the mother and child. Both the paternity index and the exclusion probability depend on the marker allele frequencies of the relevant ethnic group. By reporting these two statistics for each locus separately, one can see which loci are critically important in confirming paternity. The cumulative measures, however, are the ones that should be cited.

In the output file Mendel9b.out, describing a twin zygosity problem, the prior probability 0.46154 that like-sexed twins are identical is passed as the pedigree copy number 46154 of the first pedigree, where the twins are identical. The second pedigree, where the twins are fraternal, has a pedigree copy number of 53846 and a prior probability of 0.53846. These priors are taken from the reference [17]. The analysis output at the bottom of Mendel9b.out reads:

```
                CUMULATIVE RESULTS OVER ALL LOCI

PEDIGREE   PRIOR  LOGLIKELIHOOD   LIKELIHOOD   RATIO    POSTERIOR
PED#1      0.462    -12.443        0.000004   1.000000   0.908
PED#2      0.538    -14.889        0.000000   0.086645   0.092
```

Here we have selected model 3 and an error rate of 0.01. In this model, the test results from different markers are not independent, and Mendel reports only the cumulative results. The posterior probability 0.908 that the twins are monozygous is relatively low because of the low polymorphism of the two markers. The ratio column in the above output will be explained shortly.

The final two examples for Option 9 involve computation of forensic probabilities. DNA left by the perpetrator at a crime scene shows the genotypes 10/11, 8/8, 6/7, and 16/16 at four unlinked, highly polymorphic markers. Our third example duplicates the analysis of Charles Brenner (web site http://dna-view.com) on this problem. Given the allele frequencies in Def9c.in, the probability that a unrelated suspect matches at all four loci is 0.000138. This number appears in the likelihood column of the output

```
                CUMULATIVE RESULTS OVER ALL LOCI

PEDIGREE   PRIOR  LOGLIKELIHOOD   LIKELIHOOD   RATIO    POSTERIOR
PED#1      1.000    -8.886        0.000138   1.000000   1.000
```

from Mendel9c.out. In this situation, the quoted prior and posterior probabilities are not meaningful.

In our fourth example, we take the same data and compute the likelihood ratio of two different scenarios. The pedigree file Ped9d.in shows two pedigrees with the suspect labeled S and the perpetrator P in each. In the first pedigree, S and P are unrelated. In the second pedigree, they are siblings. The most important fact to be gleaned from the output

```
                CUMULATIVE RESULTS OVER ALL LOCI

PEDIGREE   PRIOR  LOGLIKELIHOOD   LIKELIHOOD   RATIO    POSTERIOR
PED#1      0.500    -17.547        0.000000   0.005385   0.005
PED#2      0.500    -12.323        0.000004   1.000000   0.995
```

from Mendel9d.out is that the likelihood ratio is 0.005385. In general, the ratio column gives the likelihood of each pedigree standardized by the largest likelihood encountered. Inspection of the control file Control9d.in shows that we have elected model 3 with a genotyping error rate of 0.01. The posterior probabilities quoted in the above output are not typically used in American courts. Users interested in pursuing the theory behind these computations can consult the reference [110].

9.5 Germane Keywords

```
ANALYSIS_OPTION = Pedigree_Selection
GENOTYPING_ERROR_RATE
MODEL
READ_PEDIGREE_COPIES
```

Random Quotes

Probability and Statistics used to be married; then they separated; then they got divorced; now they hardly see each other.

David Williams in *Weighing the Odds: A Course in Probability and Statistics*

Tiny differences in input could quickly become overwhelming differences in output . . . In weather, for example, this translates into what is only half-jokingly know as the Butterfly Effect — the notion that a butterfly stirring the air today in Peking can transform storm systems next month in New York.

James Gleick in *Chaos*

I lied and I lied — and then I lied some more. I lied about where I had been, I lied about where I had found information, I lied about how I wrote the story. And these were not everyday little white lies — they were complete fantasies, embellished down to the tiniest made-up detail.

I lied about a plane flight I never took, about sleeping in a car I never rented, about a landmark on a highway I had never been on. I lied about a guy who helped me at a gas station that I found on the Internet and about crossing railroad tracks I only knew existed because of aerial photographs in my private collection. I lied about a house I had never been to and decorations and furniture in a living room I had only seen in photographs in an archive maintained by Times photo editors.

Jayson Blair, former reporter for the *New York Times*

One upon a time there were four little Rabbits, and their names were — Flopsy, Mopsy, Cottontail, and Peter.

Beatrix Potter in *The Tale of Peter Rabbit*

10 Analysis Option 10: Kinship

10.1 Background

Kinship coefficients quantify the degree of relationship between two relatives. Analysis Option 10 computes global kinship coefficients and local kinship coefficients. (Global kinship coefficients are also known as theoretical; local are also known as empirical or conditional.) A global kinship coefficient Φ_{ij} between two individuals i and j is the probability that a randomly sampled gene from i is identical by descent to a randomly sampled gene from the same arbitrary locus of j.

We have two overall strategies to calculate global kinship coefficients. In our first strategy, Φ_{ij} depends only on the relationship between i and j as given in the provided pedigree structure, and ignores all marker data on them. For example, the kinship coefficient of two siblings is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. The first factor of $\frac{1}{2}$ is the probability that the sampled genes are both maternal or both paternal in origin. The second factor of $\frac{1}{2}$ is the conditional probability that the two sampled genes from the mother or from the father coincide. If i and j are the same person, then the sampling is done with replacement. Thus, a non-inbred person has kinship coefficient $\frac{1}{2}$ with himself or herself.

Our second strategy provides an estimate for the global kinship coefficients based on dense, genome-wide SNP genotypes. Here the user has several options as to how to move from SNP genotypes to global kinship coefficients. If one has confidence that the given pedigrees are distinct, i.e., that members of different pedigrees are at most distantly related, then one can quickly estimate global kinship for all pairs of individuals within the provided pedigrees. On the other hand, if pedigree information is dubious or absent, then the user can choose to estimate global kinship coefficients for all pairs of individuals. Another user-specified option is the number of SNPs to use in the kinship estimation. Finally, the user can specify which of two formulae is used to convert SNP genotypes to kinship coefficients. Each of these three decisions is of course specified by setting a keyword in the control file, as explained in the following sections.

The SNP-based global kinship estimates are accurate given a reasonable amount of genome-wide data [26], and still fast to calculate, since it can be completed in a single pass through the data. Even ignoring all prior relationship information, we can use these estimates to cluster individuals into groups of relatives. These groups accurately mimic true pedigree groupings although they lack the detailed relationship structure, since, for example, both parent-offspring and full-siblings have global kinship coefficients of $\frac{1}{4}$.

In contrast to global values, a local kinship coefficient $\hat{\Phi}_{ij}$ at a locus L must depend on genotyping outcomes at or near L. Typing results may suggest that i and j bear identical chromosome segments in some regions and non-identical segments in other regions. Local kinship coefficients are useful in QTL (quantitative trait loci) mapping as performed

by Option 19. We again have two strategies for calculating the local kinship coefficients. Our first method is likelihood-based and designed for sparser data. The second method is designed for dense genome-wide SNPs. It uses imputation to reduce the variability of the estimates and a dynamic programming technique for speed [26].

It is much harder to compute local kinship coefficients than global kinship coefficients. Global kinship coefficients based solely on pedigree structure yield readily to the classical algorithm sketched in Chapter 5 of Lange (2002) [59], which is almost trivial to compute. The SNP-based global kinship estimates are also reasonably quick even with millions of SNPs. Both these algorithms extend immediately to X-linked inheritance. In contrast, Mendel will only compute the likelihood-based local kinship coefficients estimates on small to medium-sized pedigrees. Section 0.10.1 explains the limitations of the Lander-Green-Kruglyak algorithm in this regard. SimWalk is capable of approximating local kinship coefficients on large pedigrees for this type of data assuming autosomal inheritance [69, 99, 101], although its Markov chain Monte Carlo (MCMC) technique is still computationally intensive. Our SNP-based local kinship coefficient estimates are much more efficient, but there are still significant practical limitations in the size of the data sets it can handle.

The comparison of pedigree-based global kinship coefficients to SNP-based global or any local kinship coefficients is often helpful in detecting relationship miss-specifications. If the number of alternative pedigrees is small, then Option 9 is the preferred method for deciding among the pedigrees. However, even in this situation, Option 10 gives insight into the extent of over or under-sharing of genes identical by descent.

10.2 Appropriate Problems and Data Sets

First, some general notes. Global kinship coefficients based solely on pedigree structure can be quickly computed on pedigrees of all sizes, since marker data are ignored. For any local or SNP-based kinship analysis, avoid mixing autosomal and X-linked loci. Both sets of loci can be analyzed, but they must be run separately.

For standard, text-based input files, local kinship analyses will be likelihood-based. For such analyses, including comparing the local to global kinship coefficients, large pedigrees will be bypassed. For binary data files, the SNP-based global and local methods will be used. The option to cluster individuals by their global kinships is only available for these binary files. A summary of the available suboptions is in Table 10.1.

10.3 Input Files

We will first consider data sets presented in standard, text-based files. Option 10 decides that all loci are autosomal or X-linked based on the first locus in the map file shared with

Table 10.1: Analyses available in the Kinship Option

Sub-Option	Text-based Input Files	Binary Input Files
Model 1	Pedigree-based global kinship coefficients	Both Pedigree-based and SNP-based global kinship coefficients
Model 2	Likelihood-based local kinship coefficients	SNP-based local kinship coefficients
Model 3	Compares the Model 1 and Model 2 values	Compares the two values in the Model 1 analysis
Model 4	Not available	Clusters individuals based on SNP-based global kinship coefficients

the definition file. Models 2 and 3 analyze all loci common to the map and definition files. There must be at least two such loci. If you only have data on a single marker, then add a second dummy marker with a single allele.

Models 2 and 3 will consider points between markers if the number of such points is specified in the map file or determined by a command such as `INTERIOR_POINTS = 2` inserted in the control file. If you want the number of interior points between adjacent markers to vary, then follow the instructions in Section 0.5.5.

Model 3 reports in the summary file the relative pairs with the biggest discrepancies between their pedigree-based global and likelihood-based local kinship coefficients. By default, the 25 most deviant relative pairs over all pedigrees are recorded. To change the default, enter a command such as `MAX_KINSHIP_PAIRS = 50` in the control file.

We now consider data sets that include binary files. Here, Mendel will read from the SNP definition file if either all analyzed SNPs are autosomal or all X-linked. Again, both types of loci can be analyzed, but they must be run separately. The SNP-based kinship calculations work best with dense genome-wide data sets.

As mentioned above, the user has several options for how SNP-based kinship estimation is carried out. The keyword `KINSHIP_SOURCE` (with case insensitive values) determines whether kinship coefficients are estimated only for pairs of individuals within the provided pedigrees or for all pairs of individuals. The default, which can be selected via the command `KINSHIP_SOURCE = SNPs_within_pedigrees`, only analyzes pairs of individuals within each pedigree listed in the input files. This method implies that all kinship coefficients are zero for members of different pedigrees. In practice, this method is fast because most stated pedigrees will be of moderate size. Alternatively, using the command `KINSHIP_SOURCE = SNPs_using_everyone` analyzes all pairs of individuals, regardless of

any pedigree structures in the input files. Clearly, this is the strategy that uses the least assumptions about the validity of the listed pedigree structures. Of course this comes as some cost, namely, longer computation times for large data sets. When using SNP genotypes in global kinship estimation, the run time is proportional to mn^2 where m is the number of SNPs used and n is the number of individuals in the largest pedigree. By ignoring any input pedigree structures, one is effectively lumping all individuals into one overall pedigree. In practice, this strategy still has acceptable run times even for large data sets. It should be used whenever relatedness between members of different pedigrees is a reasonable possibility.

Next, the user can set how many SNPs to use during kinship estimation. In practice, with common large data sets, we have found using 20% of the SNPs gives fine estimates for the global kinship coefficients. Thus, the default setting for the keyword `SNP_SAMPLING_INCREMENT` is 5, i.e., sample every fifth SNP, under models 1, 3, and 4. Under model 2, which estimates local kinship and is a slower analysis, the default value is 100, i.e., sample every hundredth SNP. With reasonably dense SNPs, and since recombination is a relatively rare event, this coarse analysis of only 1% of the SNPs still gives a good picture of common ancestry locally along the genome. Alternatively, setting this keyword to 1 forces 100% of the SNPs to be used, which results in somewhat more accurate estimates but longer run times. The value applied to this keyword can be any positive integer. As an example, using the commands

```
KINSHIP_SOURCE = SNPs_using_everyone
SNP_SAMPLING_INCREMENT = 1
```

in a control file will cause Mendel to use all SNP genotype data to estimate the kinship coefficients of all pairs of individuals. Compared to Mendel's default settings, this will result in more accurate results although with longer run times. If the pedigree structures listed in the input files are of decent quality, then the increase in accuracy for global kinship estimates will be small, since members of different pedigrees will probably be at most distantly related.

Of course, for smaller data sets one needs to use a greater fraction of the SNPs to accurately recover the global kinship coefficients. This is enforced using the keyword `MIN_SAMPLED_SNPS`, which has default value 5000. If the SNP sampling increment results in fewer SNPs being sampled than the value chosen for `MIN_SAMPLED_SNPS`, then the SNP sampling increment is decreased just enough to bring the number of SNPs sampled above the minimum value.

The last option for global kinship estimation is a choice in the actual formula that converts from the selected SNP genotypes to a global kinship coefficient. The case insensitive value of the keyword `KINSHIP_METHOD` determines which of two formulae is chosen. The command `KINSHIP_METHOD = GRM`, which is the default, tells Mendel to use the Genetic

Relationship Matrix method. Under GRM, the estimate of the global kinship coefficient of individuals i and j is

$$\hat{\Phi}^*_{ij} = \frac{1}{2S}\sum_{k=1}^{S}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$$

where k ranges over the selected S SNPs, p_k is the minor allele frequency of SNP k, and x_{ik} is the number of minor alleles in individual i's genotype at SNP k. Alternatively, the command `KINSHIP_METHOD = MoM` tells Mendel to use the Method of Moments [26]. Under MoM, the estimate of the global kinship coefficient of individuals i and j is

$$\hat{\Phi}^*_{ij} = \frac{e_{ij}-\sum_{k=1}^{S}\left[p_k^2+(1-p_k)^2\right]}{S-\sum_{k=1}^{S}\left[p_k^2+(1-p_k)^2\right]},$$

where

$$e_{ij} = \frac{1}{4}\sum_{k=1}^{S}\left[x_{ik}x_{jk}+(2-x_{ik})(2-x_{jk})\right]$$

is the observed fraction of alleles identical by state (IBS) between i and j. Both methods are fairly quick to calculate and provide good estimates given reasonably dense genome-wide data. When using the GRM method, very rare SNPs (those with minor allele counts less than 3) are not used since they become over weighted. In general, one can think of the GRM method centering and scaling each genotype, while the MoM method uses the raw genotypes and then centers and scales the final result.

The output of Option 10 are kinship coefficients for pairs of individuals. For SNP genotype data sets input through binary files, the output includes results either for only pairs contained within the provided pedigree structures or for all pairs, regardless of input pedigrees. As described above, this choice is determined by the keyword `KINSHIP_SOURCE`. For example, Model 1 output for SNP binary files lists, for the specified set of pairs of individuals, two estimates of their global kinship coefficients: one based solely on pedigree structure and one based on the selected SNPs' genotypes.

Similarly, Model 2 outputs estimates of the local kinship at the selected SNPs for each of the selected pairs of individuals.

Model 3 reports in the summary file the pairs of individuals with the biggest discrepancies between their pedigree-based and SNP-based global kinship coefficients. By default, the 25 most deviant relative pairs over all pedigrees are recorded. To change the default, enter a command such as `MAX_KINSHIP_PAIRS = 50` in the control file.

Finally, Model 4 clusters the individuals into groups using the set of all pairwise SNP-based global kinship coefficients. Two individuals will be put into the same group if their

SNP-based global kinship coefficient is at or above a threshold value, which by default is 0.05. To change this threshold, enter a command such as `KINSHIP_THRESHOLD = 0.125` in the control file. Larger values for this threshold result in smaller groups since only closely related individuals are grouped. Alternatively, smaller values result in bigger groups as more distantly related individuals are brought together.

10.4 Examples

Most of the relevant output is written to the summary file. In addition, the standard output file reports the points at which local coefficients are computed and any pedigrees that were too complex to process. In models 1 and 2, kinship coefficients appear in the summary file. The model 1 output

```
          PEDIGREE-BASED GLOBAL KINSHIP ESTIMATES
       FOR ALL PAIRS OF INDIVIDUALS IN EACH PEDIGREE.

    PEDIGREE          INDIVIDUALS      GLOBAL KINSHIP
 NUMBER   NAME      FIRST     SECOND     COEFFICIENT

      1  INBRED     P1        P1           0.50000
      1  INBRED     P2        P1           0.00000
      1  INBRED     P2        P2           0.50000
      1  INBRED     P4        P1           0.25000
      1  INBRED     P4        P2           0.25000
      1  INBRED     P4        P4           0.50000
      1  INBRED     P3        P1           0.25000
      1  INBRED     P3        P2           0.25000
      1  INBRED     P3        P4           0.25000
      1  INBRED     P3        P3           0.50000
      1  INBRED     P6        P1           0.25000
      1  INBRED     P6        P2           0.25000
      1  INBRED     P6        P4           0.37500
      1  INBRED     P6        P3           0.37500
      1  INBRED     P6        P6           0.62500
      1  INBRED     P5        P1           0.25000
      1  INBRED     P5        P2           0.25000
      1  INBRED     P5        P4           0.37500
      1  INBRED     P5        P3           0.37500
      1  INBRED     P5        P6           0.37500
      1  INBRED     P5        P5           0.62500

PEDIGREE INBRED HAS AVERAGE INBREEDING COEFFICIENT 0.08333.
```

from Summary10a.out is fairly typical. These pedigree-based global kinship coefficients are the ones cited for the brother-sister mating pedigree in Chapter 5 of Lange (2002) [59]. To give an idea of the overall level of inbreeding, the average inbreeding coefficient is appended to the bottom of each pedigree's list of global kinship coefficients [23]. Since example data set 10a is presented in text-based files, model 2 output would be a similar list of likelihood-based local kinship coefficients for each locus or interior point.

Model 3 reports the most deviant pairs for three statistics. When using text-based input files, suppose two relatives i and j have pedigree-based global kinship coefficient Φ_{ij} and likelihood-based local kinship coefficient $\hat{\Phi}_{ijp}$ at point p along the genetic map connecting the first and last loci shared by the definition and map files. If there are n such points in all, then the statistics are

$$\begin{aligned} S_{1ij} &= \frac{1}{n}\sum_{p=1}^{n}\left(\hat{\Phi}_{ijp}-\Phi_{ij}\right) \\ S_{2ij} &= \frac{1}{n}\sum_{p=1}^{n}\left(\frac{\hat{\Phi}_{ijp}-\Phi_{ij}}{\sqrt{\Phi_{ij}}}\right) \\ S_{3ij} &= \frac{1}{n}\sum_{p=1}^{n}\left(\frac{\hat{\Phi}_{ijp}-\Phi_{ij}}{\Phi_{ij}}\right). \end{aligned}$$

These give progressively more weight to distantly related pairs. Model 3 for data sets using binary files performs the same comparisons except with $n = 1$ and $\hat{\Phi}_{ijp}$ replaced with $\hat{\Phi}^{*}_{ij}$, the SNP-based global kinship coefficient estimate.

In the model 3 output

```
                    MOST DEVIANT RELATIVE PAIRS:

 STATISTIC NUMBER 1

     PEDIGREE          INDIVIDUALS         STATISTIC
  NUMBER    NAME      FIRST     SECOND        VALUE

     153   430         3         4            0.2420
     181   579         4         5            0.2418
      12   30          4         5            0.2415
     126   334         3         4            0.2408
      49   132         3         4            0.2389
      42   115         4         5            0.2381
      80   201         3         4            0.2378
     123   325         4         5            0.2376
     146   386         3         4            0.2372
```

89	222	3	4	0.2370
76	190	3	4	0.2369
90	227	4	5	0.2359
133	355	4	5	0.2335
9	27	3	4	0.2299
108	265	3	4	0.2278
15	41	4	5	0.2242
141	379	3	5	0.1850
100	250	3	4	-0.1816
139	371	3	4	0.1785
168	509	6	8	-0.1701
22	56	6	7	0.1683
92	233	3	4	-0.1653
164	469	3	4	0.1618
93	235	3	6	-0.1601
82	206	3	5	0.1583

from Summary10b.out, we see the 25 most deviant pairs displayed for the first statistic S_1. The other two statistics produce the same list with different statistic values. In these data from the AGRE database [36], the pedigree file Ped10b.in fails to indicate that 16 siblings are identical twins. (Section 0.5.6 describes how to specify identical twins.) Treating the identical twins as fraternal twins gives pedigree-based global kinship coefficients that are consistently too low, and this fact is evident in the first 16 entries above.

We next constructed an example data set, 10c, that uses dense SNP data in binary files. To simulate data with realistic linkage disequilibrium (LD) structure, we took advantage of phased sequence data from chromosome 19 on 85 individuals of northern and western European ancestry (originally from the CEPH sample) made publicly available in the 1000 Genomes Project [52]. After we removed markers that were mono-allelic in this set of individuals, 253,141 SNPs remained. Almost half of the SNPs have minor allele frequencies (MAF) below 5%. The haplotype pairs attributed to the 85 CEPH members were reassigned to the 85 founders of 27 pedigree structures selected from the Framingham Heart Study (FHS, http://www.framinghamheartstudy.org). The selected Framingham pedigrees were chosen to reflect the kind of pedigrees commonly collected in family-based genetic studies. The 27 pedigrees encompass 212 people, range in size from 1 to 36 people and from 1 to 5 generations, and contain sibships of 1 to 5 children. The genotypes of non-founders were simulated, using Option 17, conditional on the haplotypes imposed on the founders. All genotypes were recorded as unordered for subsequent analyses.

The model 1 output

```
SNP-BASED GLOBAL KINSHIP COEFFICIENTS ARE ESTIMATED
          WITHIN EACH INPUT PEDIGREE
```

```
                    AND WITH 50628 EVENLY SPACED SNPS
                 (USING A SNP SAMPLING INCREMENT OF 5 SNPS).
              ESTIMATION IS VIA THE GENETIC RELATIONSHIP MATRIX METHOD.


      FIRST                SECOND              GLOBAL KINSHIPS
PEDIGREE INDIVIDUAL   PEDIGREE INDIVIDUAL     PED-BASED  SNP-BASED

1         16          1         16             0.50000    0.50186
1         8228        1         16             0.00000    0.01186
1         8228        1         8228           0.50000    0.50000
1         17008       1         16             0.00000    0.01514
1         17008       1         8228           0.00000    0.00000
1         17008       1         17008          0.50000    0.50000
1         9218        1         16             0.25000    0.25187
1         9218        1         8228           0.00000    0.00000
1         9218        1         17008          0.25000    0.28057
1         9218        1         9218           0.50000    0.52051
1         3226        1         16             0.12500    0.12367
1         3226        1         8228           0.25000    0.25014
1         3226        1         17008          0.12500    0.12968
1         3226        1         9218           0.25000    0.25207
1         3226        1         3226           0.50000    0.50000

PEDIGREE 1 HAS AVERAGE INBREEDING COEFFICIENT 0.00895.
```

at the top of Summary10c.out shows the pedigree-based and SNP-based global kinship coefficients for all pairs of individuals in pedigree 1. The agreement is reasonably good considering this data set only contains SNPs on chromosome 19, hardly genome-wide. After three seconds to read in the data and perform standard quality control procedures, the entire kinship analysis takes well under one second on a standard laptop computer.

If we run this data set under model 2, the top of the summary file

```
                SNP-BASED LOCAL KINSHIP COEFFICIENT ESTIMATES
                       WITHIN EACH INPUT PEDIGREE
                    AND WITH 2532 EVENLY SPACED SNPS
                (USING A SNP SAMPLING INCREMENT OF 100 SNPS).


  SNP             FIRST                SECOND          LOCAL KINSHIP
  NUMBER    PEDIGREE INDIVIDUAL   PEDIGREE INDIVIDUAL    COEFFICIENT

    8       1         16          1         16             0.50000
```

```
8       1       8228       1       16          0.00000
8       1       8228       1       8228        0.50000
8       1       17008      1       16          0.00000
8       1       17008      1       8228        0.00000
8       1       17008      1       17008       0.50000
8       1       9218       1       16          0.50000
8       1       9218       1       8228        0.00000
8       1       9218       1       17008       0.25000
8       1       9218       1       9218        0.50000
8       1       3226       1       16          0.00000
8       1       3226       1       8228        0.25000
8       1       3226       1       17008       0.25000
8       1       3226       1       9218        0.25000
8       1       3226       1       3226        0.50000
```

lists the SNP-based local kinship coefficient at SNP number 8 for the same pairs of individuals. The file goes on to list the estimates for all pairs of individuals in each pedigree at 2532 SNPs evenly spaced throughout the data set, that is at every 100^{th} SNP. The standard output file lists the name and position for each of the 2532 SNPs cited in the summary file.

If we also run this data set under model 3, the top of the summary file

```
SNP-BASED GLOBAL KINSHIP COEFFICIENTS ARE ESTIMATED
                WITHIN EACH INPUT PEDIGREE
              AND WITH 50628 EVENLY SPACED SNPS
           (USING A SNP SAMPLING INCREMENT OF 5 SNPS).
      ESTIMATION IS VIA THE GENETIC RELATIONSHIP MATRIX METHOD.


                    MOST DEVIANT RELATIVE PAIRS:

 STATISTIC NUMBER 1

       FIRST                  SECOND              GLOBAL KINSHIPS      STATISTIC
 PEDIGREE INDIVIDUAL    PEDIGREE INDIVIDUAL     PED-BASED  SNP-BASED      VALUE

 31         15884       31         19770         0.25000    0.13384     -0.1162
 17         908         17         10418         0.25000    0.14687     -0.1031
 10040      234         10040      234           0.50000    0.59459      0.0946
 24         19621       24         19621         0.50000    0.59052      0.0905
 8          13234       8          5226          0.25000    0.33390      0.0839
 31         14705       31         4257          0.25000    0.16776     -0.0822
 23         9943        23         392           0.12500    0.04510     -0.0799
 19         22375       19         720           0.12500    0.04693     -0.0781
```

31	152	31	19770	0.12500	0.04701	-0.0780
11	7670	11	11280	0.25000	0.17287	-0.0771
14	26732	14	264	0.00000	0.07659	0.0766
31	14705	31	6885	0.12500	0.04950	-0.0755
19	26591	19	22375	0.12500	0.05097	-0.0740
19	14432	19	22375	0.25000	0.17617	-0.0738
25	11822	25	24192	0.25000	0.17634	-0.0737
31	14705	31	10439	0.12500	0.05649	-0.0685
25	3012	25	3016	0.12500	0.19101	0.0660
31	16785	31	17673	0.12500	0.05918	-0.0658
31	1365	31	10472	0.12500	0.05923	-0.0658
11	10579	11	15898	0.25000	0.18631	-0.0637
31	15884	31	24905	0.12500	0.06134	-0.0637
25	24192	25	398	0.12500	0.06192	-0.0631
25	3012	25	398	0.12500	0.06259	-0.0624
31	14705	31	24905	0.12500	0.06273	-0.0623
23	9943	23	743	0.12500	0.18643	0.0614

lists the 25 pairs of related individuals with the largest discrepancy between their pedigree-based and SNP-based global kinship coefficients, as measured by the first statistic S_1. Again, the discrepancies are reasonably small considering the data set contains only chromosome 19 SNPs.

Finally, to demonstrate the clustering capabilities of Option 10, we use the command `KINSHIP_SOURCE = SNPs_using_everyone` in Control10d.in to force Mendel to ignore all pedigree structures in the input files during its analysis. (Of course one can strip out all pedigree information manually by putting everyone in the same pedigree, removing all parental information, and making sure no two individuals have the same name.) This example uses the same input data files as example 10c. We thus mimic genetic data collected from a small community by researchers who did not collect any relationship information. With dense SNP data, all necessary pedigree groupings can be reconstructed. The Control10d.in file

```
!  Input Files
!
DEFINITION_FILE = Def10c.in
PEDIGREE_FILE = Ped10c.in
SNP_DATA_FILE = SNP_data10c.bin
SNP_DEFINITION_FILE = SNP_def10c.in
!
!  Output Files
!
Output_file = Mendel10d.out
```

```
Summary_file = Summary10d.out
NEW_PEDIGREE_FILE = Ped10d.out
NEW_DEFINITION_FILE = Def10d.out
NEW_SNP_DATA_FILE = SNP_data10d_out.bin
NEW_SNP_DEFINITION_FILE = SNP_def10d.out
!
!  Analysis Parameters
!
ANALYSIS_OPTION = Kinship
MODEL = 4
KINSHIP_SOURCE = SNPs_using_everyone
KINSHIP_THRESHOLD = 0.125
```

shows that we have put the clustering threshold at 0.125, a bit larger than the 0.05 default. With only chromosome 19 represented, the SNP-based global kinship estimates are a bit looser, so a more stringent threshold is reasonable. We have also defined the names of new files in which to place the clustered individuals and their genotypes. Although this procedure does not alter any genotypes, a new SNP data file is required to be generated because the order of the individuals in the new pedigree file will almost always be different than in the input files.

Comparing Ped10d.out with Ped10c.in shows that Mendel can indeed reconstruct the true pedigree groupings. There are a couple of points to make about Ped10d.out. First, when Mendel was told to ignore input pedigree structures, it combined everyone into one pedigree, named COMBINED. Each individual was also renamed in the format $p : i$ where p is the number of the indivual's pedigree and i is the number of the individual in the original input pedigree file. Second, if an individual's parents are listed in the original input file, and those parents end up in the same pedigree grouping after the cluster analysis, then the parents are listed in the output file. However, parental information is not used during the analysis, only SNP-based global kinship estimates are used to form the clusters. This can be demonstrated by manually stripping out all pedigree structure information as described above and rerunning the analysis. The results are identical, except missing parental values.

10.5 Germane Keywords

```
ANALYSIS_OPTION = Kinship
INTERIOR_POINTS
MAX_KINSHIP_PAIRS
MIN_SAMPLED_SNPS
MODEL
KINSHIP_METHOD
```

```
KINSHIP_SOURCE
KINSHIP_THRESHOLD
SNP_SAMPLING_INCREMENT
```

Random Quotes

Eventually, all things merge into one, and a river runs through it. The river was cut by the world's great flood and runs over rocks from the basement of time. On some of the rocks are timeless raindrops. Under the rocks are the words, and some of the words are theirs. I am haunted by waters.

Norman Maclean in *A River Runs Through It*

Writing came easy; it would only get hard when I got better at it.

Gary Wills

No passion in the world is equal to the passion to alter someone else's draft.

H.G. Wells

Sleep is an excellent way of listening to an opera.

James Stephens

I was angry with my friend;
I told my wrath, my wrath did end.
I was angry with my foe;
I told it not, my wrath did grow.

William Blake in *A Poison Tree*

Let me not to the marriage of true minds
Admit impediments. Love is not love
Which alters when it alteration finds,
Or bends with the remover to remove:
O no! it is an ever-fixed mark
That looks on tempests and is never shaken;
It is the star to every wandering bark,
Whose worth's unknown, although his height be taken.
Love's not Time's fool, though rosy lips and cheeks
Within his bending sickle's compass come:
Love alters not with his brief hours and weeks,
But bears it out even to the edge of doom.
If this be error and upon me proved,
I never writ, nor no man ever loved.

William Shakespeare, Sonnet 116

11 Analysis Option 11: Genetic Equilibrium

11.1 Background

The notions of Hardy-Weinberg equilibrium and linkage (or gametic phase) equilibrium play central roles in population genetics theory. Analysis Option 11 is designed to test both assumptions. Hardy-Weinberg equilibrium relates genotype and allele frequencies. If allele i has population frequency p_i, then the homozygous genotype i/i has frequency p_i^2, and the heterozygous genotype i/j has frequency $2p_i p_j$. Hardy-Weinberg equilibrium is attained in a single generation under the ideal condition of random union of gametes. Linkage equilibrium involves two or more linked loci. Suppose allele i at locus 1 has frequency p_i and allele j at locus 2 has frequency q_j. Under linkage equilibrium, the haplotype ij has frequency $p_i q_j$. This multiplicative (or independence) rule extends naturally to multiple linked loci. In contrast to Hardy-Weinberg equilibrium, linkage equilibrium may be reached very slowly even under ideal conditions. For this reason, linkage equilibrium is useful in disease association tests such as the TDT featured in Analysis Option 13. Ignoring this important application, Hardy-Weinberg equilibrium and linkage equilibrium generally simplify statistical analysis and are assumed in many options of Mendel. Hence, it is a good idea to test for genetic equilibrium as a prelude to more thorough data analysis.

11.2 Appropriate Problems and Data Sets

Mendel relies on random samples of individuals to test Hardy-Weinberg and linkage equilibrium. Pedigrees are generally too complicated for this purpose. (If you insist on using pedigrees, see model 3 of Option 6.) In Option 11, each person comprises a different pedigree. As described in Section 0.5.6 and illustrated by some of the Option 11 example input files, these pedigrees may have copy numbers. In testing genetic equilibrium, Mendel will accept either haplotype data, multilocus genotype data without phase information, or multilocus phenotype data. Although you must not mix these data types, missing genotypes and phenotypes are allowed.

If you have haplotype data, then we suggest that you represent haplotypes in the pedigree file by ordered genotypes. The application of an ordered allele separator for this purpose is described in Section 0.5.4.1 and illustrated in the sample pedigree file Ped11c.in. Analysis Options 3 and 23 can help you form haplotypes. Except for males at X-linked loci, every person in the pedigree file contributes two haplotypes. If haplotypes are unavailable, then you can use codominant unordered genotypes in the pedigree file. Mendel then tests for independence of genotypes across a sliding window of markers. Finally, for non-codominant markers, you can enter phenotypes in the pedigree file and test for independence of phenotypes across a sliding window of markers.

It is possible to use phenotypes in the pedigree file and preserve either the haplotype or genotype analysis paradigm. However, each phenotype listed in the definition file must then be compatible with just a single ordered or unordered genotype, respectively. Mendel will deduce what form of data you intend for analysis. Beware, however, that a single phase ambiguity in the pedigree file will force Mendel to assume genotype data rather than haplotype data. A single phenotype in the pedigree file corresponding to more than one unordered genotype will force Mendel to assume phenotype data rather than genotype data. Thus, the hierarchy is haplotype data, genotype data, and finally phenotype data. Males at X-linked loci are ignored if the data are viewed as genotype or phenotype data rather than haplotype data.

The current version of Mendel no longer supports the old method of inputting haplotypes, which asked the user to create one completely homozygous individual for each observed haplotype. If you have legacy pedigree files prepared according to this convention, then it is possible to salvage them. Just change all loci to X-linked loci, and declare all individuals to be males. This maneuver tricks Mendel into interpreting a sequence of homozygous genotypes as a single haplotype.

One can also use Analysis Option 6 to test for Hardy-Weinberg equilibrium at a single locus at a time. Option 6 relies on a likelihood ratio test and estimation of allele frequencies. P-values may be misleading if some genotypes are rare, for example with microsatellite markers having large numbers of alleles. In such cases, Option 11 is preferred because it conditions on observed allele counts and handles sparse data more faithfully. When we pass to haplotypes formed from many loci, sparse data are the rule rather than the exception.

11.3 Input Files

Option 11 implements two models for examining groups of loci. In model 1, the default model, Mendel conducts Fisher's exact test of independence [59] among the loci in a sliding window. Each window encompasses a contiguous block of loci common to the definition and map files. The width of the sliding window is determined by the keyword `NUMBER_OF_MARKERS_INCLUDED`. This keyword has default value 1, implying that all loci will be tested for Hardy-Weinberg equilibrium. If the width of the window is set to a value greater than 1 with a command in the control file such as

```
NUMBER_OF_MARKERS_INCLUDED = 2
```

then Mendel will test for linkage equilibrium in the haplotype case and a combination of Hardy-Weinberg and linkage equilibrium in the multilocus genotype and phenotype cases. Mendel accepts the value "ALL" for the keyword `NUMBER_OF_MARKERS_INCLUDED`.

Option 11 uses a permutation method to approximate p-values. The accuracy of such approximation depends on the number s of data permutations undertaken. The value of s is determined by the keyword SAMPLES. A command such as SAMPLES = 100000 in the control file overrides the default value of 10,000 for s. If none of the results from the data permutations are as or more extreme than the value from the observed data, then the p-value is reported as less than $1/s$. Each reported p-value $\hat{p}$ has an attached range of plus or minus twice $\sqrt{\hat{p}(1-\hat{p})/s}$, the approximate standard error of $\hat{p}$. If $\hat{p}$ is reported as 1.0 or less than $1/s$, then the range is omitted. The references [59, 73] explain the conditional inference framework of Option 11.

Model 2 is intended for genetic equilibrium testing with pairs of loci. Under model 2, Mendel visits all pairs of loci common to the definition and map files. Most of the visited pairs will be nonadjacent. For each pair of loci, Mendel conducts Fisher's exact test and, for haplotype data, computes two measures of linkage disequilibrium. Suppose n_{ij} haplotypes are observed with allele i at the first locus of a pair and allele j at the second locus of the pair. The marginal sums

$$n_{i\cdot} = \sum_j n_{ij}, \quad n_{\cdot j} = \sum_i n_{ij}, \quad n = \sum_i n_{i\cdot} = \sum_j n_{\cdot j}$$

allow estimation of the frequency of allele i, the frequency of allele j, and the joint frequency of the two alleles as

$$q_i = \frac{n_{i\cdot}}{n}, \quad r_j = \frac{n_{\cdot j}}{n}, \quad p_{ij} = \frac{n_{ij}}{n},$$

respectively. Mendel reports the adjusted χ^2 statistic

$$\frac{\chi^2}{n} = \frac{1}{n}\sum_{ij}\frac{(n_{ij}-nq_ir_j)^2}{nq_ir_j} = \sum_{ij}\frac{(p_{ij}-q_ir_j)^2}{q_ir_j}$$

for every pair of loci and Lewontin's D' statistic

$$D' = \begin{cases} \frac{p_{11}-q_1r_1}{\min\{q_1r_2,q_2r_1\}} & \text{when } p_{11}-q_1r_1 \geq 0 \\ \frac{p_{11}-q_1r_1}{\min\{q_1r_1,q_2r_2\}} & \text{when } p_{11}-q_1r_1 < 0 \end{cases}$$

for every pair of biallelic loci.

To summarize, model 1 tests for Hardy-Weinberg equilibrium whenever the window width is 1. In the presence of haplotype data, it tests for linkage equilibrium whenever the window width exceeds 1. In the presence of multilocus genotype or phenotype data, it tests for a combination of Hardy-Weinberg and linkage equilibrium. Model 2 tests all pairs of loci for genetic equilibrium. For haplotype data, it also computes two measures of linkage disequilibrium. If you are on a fishing expedition for linkage disequilibrium, it is probably wise to start with model 2. If you detect evidence favoring linkage disequilibrium, try model 1 with an increasing sequence of window widths until you discover haplotype block boundaries.

11.4 Examples

Three examples illustrate Option 11. The recombination fractions appearing in the map file are irrelevant. In model 1 good estimates of allele frequencies are necessary. These can be left blank and filled in by Mendel. In the first example, the output

```
        FISHER'S EXACT TEST FOR HARDY-WEINBERG EQUILIBRIUM

LOCUS     ESTIMATED                POPULATION   RANDOM     EXPECTED      OBSERVED
NAME       P-VALUE         RANGE     SIZE      SAMPLES   HOMOZYGOTES   HOMOZYGOTES

MN        0.8515000 +/- 0.0071119    208       10000      131.01         132
```

appears in the file Summary11a.out. This particular example tests Hardy-Weinberg equilibrium at the MN locus in a Syrian population. Obviously, there is no evidence for departure from equilibrium. If there had been further loci common to the definition and map files, then each of these would rate a summary line similar to that for the MN locus. The population size, 208 in this example, represents the number of genotypes reported at the MN locus in the input file Ped11a.in. This number incorporates pedigree copy numbers. The excellent match between the expected and observed number of heterozygotes is consistent with the nonsignificance of Fisher's exact test.

The second example produces the output

```
                DATA TREATED AS MULTILOCUS GENOTYPES

             FISHER'S EXACT TEST FOR GENETIC EQUILIBRIUM

FIRST      LAST       ESTIMATED                  POPULATION   RANDOM
LOCUS      LOCUS       P-VALUE          RANGE      SIZE      SAMPLES

MARKER1    MARKER2    0.0003000  +/-  0.0003464     98        10000
```

in Summary11b.out. In this case, the data are two-locus genotypes. Note the reminder that the data are treated as multilocus genotypes rather than haplotypes. Always check this reminder to see if the data conform to your analysis expectations. Because the keyword NUMBER_OF_MARKERS_INCLUDED is set to 2 in Control11b.in, each window considers two loci at a time. If there had been three loci common to the definition and map files rather than two, then there would be a second window spanning MARKER2 and MARKER3 and a second line of output. A combination of Hardy-Weinberg and linkage equilibrium is tested in this example. Clearly, the data are significant. Before reporting this discovery in a scientific paper, it might be prudent to redo the analysis with a larger value for the keyword SAMPLES. Because the first pedigree has a missing genotype at MARKER2, the population

size is 98 rather than 99, the actual number of pedigrees counting copies. Only complete multilocus genotypes contribute to each p-value.

The third example illustrates model 2. Consider the amended, partial output from Summary11c.out:

DATA TREATED AS HAPLOTYPES

PAIRWISE TESTS AND STATISTICS FOR LINKAGE EQUILIBRIUM

FIRST LOCUS	SECOND LOCUS	ESTIMATED P-VALUE		RANGE	POPULATION SIZE	ADJUSTED CHI-SQ	DPRIME
Marker1	Marker2	< 0.0001000			194	0.80014	0.94360
Marker1	Marker3	< 0.0001000			194	0.40252	0.87799
Marker1	Marker4	< 0.0001000			194	0.55622	0.94085
Marker1	Marker5	1.0000000			194	0.00028	-0.19167
Marker1	Marker6	< 0.0001000			194	0.29705	0.87067
Marker1	Marker7	0.6110000	+/-	0.0097505	194	0.00691	-1.00000
Marker1	Marker8	< 0.0001000			194	0.37464	0.87643
Marker1	Marker9	0.0007000	+/-	0.0005290	194	0.05079	-0.86528
Marker1	Marker10	0.0500000	+/-	0.0043589	194	0.02136	-0.78444
Marker2	Marker3	< 0.0001000			194	0.51840	0.94454
Marker2	Marker4	< 0.0001000			194	0.62605	0.94623
Marker2	Marker5	1.0000000			194	0.00059	-0.26515
Marker2	Marker6	< 0.0001000			194	0.38628	0.94121
Marker2	Marker7	0.6245000	+/-	0.0096850	194	0.00769	-1.00000

This output shows strong evidence for linkage disequilibrium between most marker pairs. P-values for Fisher's exact test were generated using 10,000 random samples. The two linkage disequilibrium statistics give a better idea of the strength of linkage disequilibrium than the p-values alone.

If you revert to model 1, the default model, in this third example, and set the keyword NUMBER_OF_MARKERS_INCLUDED = 10 in the control file, then Summary11c.out will read:

DATA TREATED AS HAPLOTYPES

FISHER'S EXACT TEST AND W_1 TEST FOR LINKAGE EQUILIBRIUM

FIRST LOCUS	LAST LOCUS	FISHER P-VALUE	RANGE	POPULATION SIZE	RANDOM SAMPLES	W_1 P-VALUE
Marker1	Marker10	< 0.0001000		194	10000	0.0000001

Here we see that the p-value for Fisher's exact test is reported as less than the reciprocal of the number of samples because none of the permuted data sets resulted in a value as or more extreme than the observed data. When this is the case Mendel also omits the range for the p-value estimate. More importantly, the p-value of the new statistic W_1 now appears. This statistic equals the number of different haplotypes observed in the data. Although Mendel does not output W_1, in the current example it equals 28. Evidently many of the 194 observed haplotypes coincide. With 10 biallelic markers, there are 1024 possible haplotypes. The p-value quoted is the probability under the null hypothesis of linkage equilibrium that the random variable $W_1 \leq 28$. Mendel computes this p-value by the algorithm sketched in Section 4.5 of the reference [59]. The algorithm requires accurate estimates of the allele frequencies at the various markers.

11.5 Germane Keywords

```
ANALYSIS_OPTION = Genetic_Equilibrium
MODEL
NUMBER_OF_MARKERS_INCLUDED
READ_PEDIGREE_COPIES
SAMPLES
```

Random Quote

Bore, n. a person who talks when you wish him to listen.
Cynic, n. a blackguard whose faulty vision sees things as they are, not as they ought to be.
Edible, adj. good to eat, and wholesome to digest, as a worm to a toad, a toad to a snake, a snake to a pig, a pig to a man, and a man to a worm.
Labor, n. one of the processes by which A acquires property for B.
Prejudice, n. a vagrant opinion without any means of support.
Saint, n. a dead sinner revised and edited.

Ambrose Bierce in *The Devil's Dictionary*

We must be careful not to confuse data with the abstractions we use to analyze them.

William James

The more it snows (Tiddley pom), The more it goes (Tiddley pom), The more it goes (Tiddley pom), on snowing, and nobody knows (Tiddley pom), How cold my toes (Tiddley pom), How cold my toes (Tiddley pom), are growing.

A.A. Milne in *The World of Pooh*

12 Analysis Option 12: Association by Permutation (Cases and Controls)

12.1 Background

Analysis Option 12 tests for genetic differences between two kinds of people, termed loosely cases and controls. Differences can be assessed at either the haplotype level, the multilocus genotype level, or the multilocus phenotype level. In population isolates such as Finland, Iceland, and Costa Rica's central valley, the carriers of a disease gene typically descend from a handful of founders. Unless the isolate is very old, recombination usually has had insufficient time to scramble founder haplotypes. Option 12 reveals conserved haplotypes by comparing haplotype, genotype, or phenotype frequencies between cases and controls. No assumptions about genetic equilibrium — either Hardy-Weinberg or linkage equilibrium — are made in the process. Case/control data from different populations can even be combined. Analysis Option 11 is designed to test for genetic equilibrium.

Option 12 presents two well-known contingency table tests of case/control homogeneity. Fisher's exact test is a good omnibus test across many cells. The $Z_{\max}$ test has greater power when one or two cells deviate strongly between cases and controls. The complicated theory behind these tests is presented in [59] and will not be pursued here except for some brief comments on permutation procedures. Analysis Option 24 handles case-control data from dense genome scans. It also has some advantages in dealing with covariates and performing model selection.

12.2 Appropriate Problems and Data Sets

Option 12 evaluates p-values by permutation of case/control labels. Unless instructed otherwise, Option 12 permutes these labels across an entire random sample of people, where each pedigree contains exactly one person. As described in Section 0.5.6 and illustrated by the Option 12 example input files, these pedigrees may have copy numbers. In testing for homogeneity, Mendel will accept either haplotype data, multilocus genotype data without phase information, or multilocus phenotype data. Unfortunately, you must not mix these. Missing phenotypes and genotypes are allowed in the pedigree file.

If you have haplotype data, then ordinarily each case or control haplotype should be input as a single-person pedigree. We suggest that you represent haplotypes in the pedigree file by ordered genotypes. The application of an ordered allele separator for this purpose is described in Section 0.5.4.1 and illustrated in the sample pedigree file Ped12a.in. Analysis Options 3 and 23 can help you form haplotypes. Except for males at X-linked loci, every person in the pedigree file contributes two haplotypes. If haplotypes are unavailable, then you can use codominant unordered genotypes in the pedigree file. Finally, for

non-codominant markers, you can enter phenotypes in the pedigree file.

It is possible to use phenotypes in the pedigree file and preserve either the haplotype or genotype analysis paradigm. However, each phenotype listed in the definition file must then be compatible with just a single ordered or unordered genotype, respectively. Mendel will deduce what form of data you intend for analysis. Beware, however, that a single phase ambiguity in the pedigree file will force Mendel to assume genotype data rather than haplotype data. A single phenotype in the pedigree file corresponding to more than one unordered genotype will force Mendel to assume phenotype data rather than genotype data. Thus, the hierarchy is haplotype data, genotype data, and finally phenotype data. Males at X-linked loci are ignored if the data are viewed as genotype or phenotype data rather than haplotype data.

The current version of Mendel no longer supports the old method of inputting haplotypes, which asked the user to create one completely homozygous individual for each observed haplotype. If you have legacy pedigree files prepared according to this convention, then it is possible to salvage them. Just change all loci to X-linked loci, and declare all individuals to be males. This maneuver tricks Mendel into interpreting a sequence of homozygous genotypes as a single haplotype.

Regardless of what form your data takes, all allele frequencies appearing in the definition file and all recombination fractions appearing in the map file are irrelevant. Option 12 operates by sliding a window along the marker map. Only those markers within a given window enter into the current statistical tests. If the width of the window is 1, then case/control homogeneity is tested one marker at a time. Option 6 can also test for homogeneity in this setting, but it relies on a likelihood ratio test and estimation of allele frequencies. P-values may be misleading if some genotypes are rare, for example with microsatellite markers having large numbers of alleles. In such cases, Option 12 is preferred because it conditions on observed allele counts and handles sparse data more faithfully. With haplotypes formed from many loci, sparse data are the rule rather than the exception. Multilocus phenotypes engender even greater spareness.

As an alternative to permutation across the entire sample, Option 12 can also limit permutations to certain defined units. This more subtle form of permutation facilitates matched case/control testing. For the purposes of Option 12, a permutation unit is simply a group of exchangeable people. A unit need not correspond to a population or ethnic group in the ordinary sense. Thus, you can consider a sibship a separate unit or an ethnically matched husband and wife a separate unit. Matching protects against ethnic stratification and permits the use of related cases. For example, when a linkage study is carried out, there may be several affecteds per pedigree. In a subsequent association study with the same data, defining permutation units based on sibships salvages many of the affecteds from these pedigrees, who would ordinarily be discarded in an unmatched design that

relies on independent cases. Of course, if you start with random samples of cases and controls, it would be foolish to impose matching.

12.3 Input Files

Mendel conducts Fisher's exact test and the $Z_{\max}$ test along a sliding window drawn from the loci common to the definition and map files. The width of this window is determined by the keyword NUMBER_OF_MARKERS_INCLUDED. This keyword has default value 1, implying that loci will be tested one by one for differences between cases and controls. If the width of the window is set to a value greater than 1 by a command such as

```
NUMBER_OF_MARKERS_INCLUDED = 2
```

in the control file, then Mendel will test for case/control homogeneity simultaneously across all of the loci within each window. As a special case, Mendel accepts the value "ALL" for the keyword NUMBER_OF_MARKERS_INCLUDED. No explicit correction for multiple tests is offered by Option 12. Tests from overlapping windows will be correlated, so the standard Bonferroni correction is apt to be too conservative.

Information on case/control status must be communicated to Mendel. This is done through a statement such as

```
AFFECTED_LOCUS_OR_FACTOR = DISEASE
AFFECTED = CASE
```

inserted in the control file. This instructs Mendel to look in the definition file for the factor DISEASE naming the case (affected) and control (not affected) categories. Once this factor is found, all people can be classified as being in one of the two categories. People with missing case/control labels are ignored. Do not invoke the default value of blank for the keyword AFFECTED_LOCUS_OR_FACTOR. Finally at the risk of confusing matters, if you want to test for genetic differences between two populations, say France and China, rather than between disease cases and controls, you might write the commands

```
AFFECTED_LOCUS_OR_FACTOR = COUNTRY
AFFECTED = FRANCE
```

in the control file. Now the frequencies of French and Chinese haplotypes, genotypes, or phenotypes can be compared.

As mentioned earlier, Option 12 uses a permutation procedure to approximate p-values. The accuracy of such approximation depends on the number s of data permutations undertaken. The value of s is determined by the keyword SAMPLES. A command such as SAMPLES = 100000 in the control file overrides the default value of 10,000 for s. If none

of the permuted data sets have a more extreme statistic than the observed data, then the p-value is reported as less than $1/s$. Each p-value $\hat{p}$ reported has an attached range of plus or minus twice $\sqrt{\hat{p}(1-\hat{p})/s}$, the approximate standard error of $\hat{p}$. If $\hat{p}$ is reported as 1.0 or less than $1/s$, then the range is omitted.

Note that the permutations undertaken in Options 11 and 12 are quite different. In Option 11, Mendel permutes alleles, genotypes, or phenotypes at each locus. In Option 12, Mendel preserves these genetic configurations and permutes case and control labels. The sizes of the case and control populations are kept constant during the permutation procedure. Inference in Option 12 is conditional on the haplotypes, multilocus genotypes, or phenotypes actually present in the data.

If you want to limit permutation of case/control labels to defined permutation units or groups, then you must insert a command such as `GROUP_FACTOR = SIBSHIP` in the control file. Only those individuals assigned a non-blank value at this factor participate in testing. Any sibship lacking either cases or controls is uninformative and should be omitted. Otherwise, each member of a sibship should be assigned a common group value unique to the sibship. With a group factor defined, it is no longer necessary to constrain pedigrees to contain exactly one person. If you use pedigree repeat numbers, then bear in mind that some of your permutation units may be larger than you might intend.

12.4 Examples

Three examples illustrate Option 12. In the first example, the output

```
                    DATA TREATED AS HAPLOTYPES

FIRST LOCUS NAME          : Marker1
LAST LOCUS NAME           : Marker9
FISHER P-VALUE            : 0.0021000 +/- 0.0009156
ZMAX   P-VALUE            : 0.0068000 +/- 0.0016436
MOST ABERRANT TYPE        : 1,1,2,1,2,2,2,2,2
CASE     SAMPLE SIZE      :    52
CONTROL  SAMPLE SIZE      :   194

FIRST LOCUS NAME          : Marker2
LAST LOCUS NAME           : Marker10
FISHER P-VALUE            : 0.0018000 +/- 0.0008478
ZMAX   P-VALUE            : 0.0051000 +/- 0.0014246
MOST ABERRANT TYPE        : 1,2,1,2,2,2,2,2,2
CASE     SAMPLE SIZE      :    50
CONTROL  SAMPLE SIZE      :   194
```

appears in the file Summary12a.out. This particular example tests homogeneity between case and control haplotypes at 10 different markers. For didactic purposes, the width of the window is 9 so that 2 windows span the 10 markers. The first window extends from Marker1 to Marker9, and the second window from Marker2 to Marker10. The Fisher exact test and $Z_{\max}$ test are significant for both windows. The most aberrant haplotype for each window is listed. The sample sizes refer to the number of complete case and control haplotypes. These differ between the two windows because the second pedigree is missing a genotype at Marker10. Only complete haplotypes contribute to each p-value.

The second example produces the output

```
                DATA TREATED AS PHENOTYPES

LOCUS NAME                : ABO
FISHER P-VALUE            : 0.0295000 +/- 0.0033841
ZMAX   P-VALUE            : 0.0177000 +/- 0.0026372
MOST ABERRANT TYPE        : O
CASE     SAMPLE SIZE      :   521
CONTROL  SAMPLE SIZE      :   680
```

in Summary12b.out. This is the same ABO ulcer data analyzed in Option 6. Now the data are phenotypes rather haplotypes. Clearly, the data deviate significantly from homogeneous phenotypic frequencies. Phenotype O is the most aberrant. Before reporting these discoveries in a scientific paper, it might be prudent to redo the analysis with a larger value for the keyword SAMPLES. In this particular problem, the default value of 1 applies to the keyword NUMBER_OF_MARKERS_INCLUDED.

The third example uses genotype data from Colombia and Japan to test for an association between Alzheimer's disease and the alleles of the APOE polymorphism [49, 117]. The output

```
       DATA TREATED AS GENOTYPES

LOCUS NAME                : APOE
FISHER P-VALUE            : LESS THAN 0.0001000
ZMAX   P-VALUE            : LESS THAN 0.0001000
MOST ABERRANT TYPE        : 3/4
CASE     SAMPLE SIZE      :   246
CONTROL  SAMPLE SIZE      :   242
```

from Summary12c.out shows the well-known association and the increased prevalence of the 3/4 genotype. The entries in the control file

```
AFFECTED_LOCUS_OR_FACTOR = DISEASE
```

```
AFFECTED = CASE
GROUP_FACTOR = COUNTRY
```

clearly defines who is affected and the two permutation units (Colombia and Japan).

12.5 Germane Keywords

```
ANALYSIS_OPTION = Association_by_Permutation
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
GROUP_FACTOR
NUMBER_OF_MARKERS_INCLUDED
READ_PEDIGREE_COPIES
SAMPLES
```

Random Quotes

Whenever two people meet, there are six people present. There is each man as he sees himself, each man as the other person sees him, and each man as he really is.

William James

Either he's dead or my watch has stopped.

Groucho Marx in *A Day at the Races*

Between two evils, I always pick the one I never tried before.

Mae West

When pygmies cast such long shadows, it must be very late in the day.

Karl Kraus

All the great things we know have come to us through neurotics. It is they and only they who have founded religions and created great works of art.

Marcel Proust in *The Guermantes Way. My Grandmother's Illness*

I succeeded in gaining the foot of the cliff on the eastern extremity of the glacier, and there discovered the mouth of a narrow avalanche gully, through which I began to climb, intending to follow it as far as possible, and at least obtain some fine wild views for my pains . . . After gaining a point about halfway to the top, I was suddenly brought to a dead stop, with arms outspread, unable to move hand or foot either up or down. My doom appeared fixed. I must fall. There would be a moment of bewilderment, and then a lifeless rumble down the one general precipice to the glacier below.

John Muir in *The Mountains of California*

13 Analysis Option 13: TDT

13.1 Background

The transmission/disequilibrium test (TDT) introduced by Spielman et al. [102] and Terwilliger and Ott [105] permits testing for linkage and/or association between a disease trait and the alleles of a codominant marker. The TDT is designed to protect against spurious associations due to population stratification. The original TDT is a McNemar's test [79] involving just two alleles and an asymptotic χ^2 approximation. Analysis Option 13 has the advantages of applying to multiple alleles and avoiding χ^2 approximations [73].

A marker locus is in transmission equilibrium with a disease if the parental alleles transmitted to a child at the marker locus are independent of the child's disease status. In other words, transmission disequilibrium describes associations with disease status within nuclear families conditional on the genotypes of the parents. In contrast, linkage disequilibrium describes associations within populations ignoring parental genotypes. Transmission disequilibrium obviously occurs when one of the marker alleles plays a direct role in the disease process. Transmission disequilibrium also occurs when the marker locus is in linkage disequilibrium with a disease-predisposing locus. A strongly positive TDT result suggests that the tested marker is a disease-predisposing locus or closely linked to such a locus. In the latter case, one or more marker alleles are in linkage disequilibrium with a disease allele.

Mendel extracts from the presented pedigrees all genotyped trios containing an affected person together with his or her parents. Because there may be multiple such trios in any given pedigree, Mendel will confound linkage and association owing to the dependencies of the trios. This is not necessarily a fatal flaw, but users should not expect new insight when the data consists of one or two large disease pedigrees already showing linkage. With many small unrelated pedigrees, the chance of confusing linkage with association becomes less of an issue, and Option 13 can help in identifying associated marker alleles.

Conditioning on observed parental genotypes eliminates nuisance parameters such as allele frequencies in addition to guarding against spurious associations due to ethnic stratification. In Mendel, p-values are assessed by a permutation procedure that shuffles the alleles passed and not passed to each affected child by its parents. Permutation tests are called exact because they avoid dubious large sample approximations. These approximations are particularly suspect for markers with multiple alleles because the contingency tables encountered in the TDT often contain many cells with small expected counts. In practice, permutation p-values must be estimated by Monte Carlo simulation. If the necessary simulation runs long enough, computed p-values will be sufficiently accurate for valid scientific inference.

Mendel computes p-values for two multiallelic test statistics. These reduce to the original biallelic TDT in case of a biallelic marker. If $c_{i/j\to i}$ counts the number of times a heterozygous parent of genotype i/j transmits allele i to an affected child, then the ordinary TDT statistics for a marker with alleles 1 and 2 is defined as

$$\text{TDT} = \frac{(c_{1/2\to 1} - c_{1/2\to 2})^2}{c_{1/2\to 1} + c_{1/2\to 2}}.$$

This McNemar's statistic ignores transmissions from homozygous parents and has an approximate χ^2 distribution with one degree-of-freedom. In extending the TDT to $k > 2$ alleles, the permutation test implemented in Mendel contrasts the number of alleles t_i of type i transmitted to affected children by heterozygous parents with the number of alleles n_i of type i that are not transmitted to affected children by heterozygous parents. In the latter case, each counted parent has heterozygous genotype i/j for some $j \neq i$. Mendel computes the permutation distribution of the test statistics

$$\begin{aligned}
\text{TDT}_1 &= \sum_{i=1}^{k} \frac{(t_i - n_i)^2}{t_i + n_i} \\
\text{TDT}_2 &= \max_{1\le i\le k} \frac{(t_i - n_i)^2}{t_i + n_i}
\end{aligned}$$

as proposed by Morris et al. [82] and others. In the output of Mendel, TDT_1 is referred to as the "chi-square TDT statistic" and TDT_2 as the "largest standardized residual TDT statistic."

13.2 Appropriate Problems and Data Sets

Option 13 is appropriate for dichotomous disease traits and codominant markers. If multiple relevant trios exist in some of the pedigrees, then the TDT test will confound linkage and association. Option 13 operates one marker at a time. If you want to use haplotypes spanning several markers, then try Analysis Option 8. Option 8 also handles noncodominant markers and disease severity indicators. Both options can be employed to check for Mendelian segregation of the alleles of a new marker.

13.3 Input Files

Although most input features are standard, it is important to note that allele frequencies are ignored in Option 13. These frequencies can be omitted in the definition file or filled in with arbitrary values. Recombination fractions are also irrelevant because each marker common to the definition and map files is analyzed separately. All loci in the definition

and pedigree files are considered markers; there is no trait locus. Some version of the commands

```
AFFECTED = 2
AFFECTED_LOCUS_OR_FACTOR = HEALTH
```

must be inserted in the control file to distinguish which children are affected. The first keyword can be omitted if one of the default values, AFFECTED and 2, is used to label affected individuals. Coding disease status as a phenotype at a locus rather than as a category of a factor can lead to some confusion. If you insist on conveying disease status in this way, then omit that locus from the map file. Otherwise, the locus will be among those analyzed.

Option 13 locates affected children and their parents throughout each pedigree. If all members of a trio are genotyped, then the trio is accepted. If one parent of the trio is untyped, but the remaining parent and affected child are different heterozygotes, then the typed duo also contributes unbiased information and is accepted. Observe that if no acceptable trios or duos are found, then p-values will be 1. If you obtain a p-value of 1 in an analysis, then it is worth looking for typographical errors and untyped markers. In checking for Mendelian segregation at a new marker, you may designate everyone as affected or simply set

```
AFFECTED = Everyone
```

in the control file. If you set `AFFECTED` equal to Everyone, then you may dispense with the superfluous factor, such as HEALTH, conveying disease status.

The default number of Monte Carlo samples is 10,000. To estimate small p-values well, it is prudent to do more simulation before publication. The command `SAMPLES = 100000` in the control file overrides the default and increases the number of samples to 100,000. If none of the s sampled statistics have a more extreme value than the observed data, then the p-value is reported as less than $1/s$. If you increase the size of `SAMPLES`, you should see smaller error bounds on your p-values. These error bounds extend two standard deviations from the mean.

If you set

```
GENDER_NEUTRAL = False
```

in the control file, then Mendel will test maternal and paternal transmission distortion separately at each model locus. There is one catch in doing so. When both parents and the affected child share the same heterozygous genotype, say 1/2, it is impossible to discern which allele the mother contributes and which allele the father contributes. In the ordinary TDT, this does not make a difference because one parent must transmit the 1 allele

and the other parent the 2 allele. In testing maternal and paternal transmission distortion separately, Mendel discards a parent if it is impossible to discern which allele the parent contributes. This action diminishes the power of the sex specific tests beyond what they already experience in omitting half the parents.

13.4 Example

In the current example, we examine autism data from the AGRE database [36] found at www.agre.org. As the chromosome 7 output

MARKER NAME	TDT 1 P-VALUE	TDT 2 P-VALUE	EXTREME ALLELE NAME	EXTREME ALLELE NUMBER
D7S2201	0.57310	0.48040	1	1
D7S3047	0.00170	0.00150	5	3
D7S1802	0.87990	0.86290	5	4
D7S1808	0.20680	0.21050	6	5
D7S817	0.81490	0.67050	6	5
D7S2846	0.36010	0.19530	5	3
D7S1818	0.93860	0.91100	6	4
D7S1830	0.68720	0.62510	5	4
D7S2204	0.88570	0.89980	13	8
D7S2212	0.42040	0.45490	3	1
GATA156	0.91770	0.85890	6	4
D7S821	0.38300	0.54820	7	5
D7S1799	0.25870	0.23700	2	1
D7S2847	0.25930	0.08120	2	1
D7S1804	0.73230	0.71470	3	2
D7S1824	0.30370	0.10350	4	3
D7S2195	0.49740	0.60700	5	3
D7S3058	0.01280	0.07600	4	4

from Summary13.out shows, there is reason to believe that marker D7S3047 is associated with autism. Allele 5 gives the largest value of the ratio $(t_i - n_i)^2/(t_i + n_i)$.

Examination of the TDT table

RESULTS FOR LOCUS D7S3047:

ALLELE NUMBER	1	2	3	4
PASSED	96	114	64	23
NOT PASSED	68	90	112	27

```
THE CHI-SQUARE TDT STATISTIC HAS APPROXIMATE P-VALUE
 0.1700E-02 PLUS OR MINUS  0.8239E-03 BASED ON    10000 RESAMPLES.

THE LARGEST STANDARDIZED RESIDUAL TDT STATISTIC HAS APPROXIMATE P-VALUE
 0.1500E-02 PLUS OR MINUS  0.7740E-03 BASED ON    10000 RESAMPLES.

THE MOST UNDER-TRANSMITTED ALLELE IS ALLELE NUMBER   3, NAMED 5.
THE MOST OVER-TRANSMITTED ALLELE IS ALLELE NUMBER   1, NAMED 3.
OF THESE TWO, THE MORE EXTREME IS ALLELE NUMBER   3, NAMED 5.
```

in Mendel13.out documents that not only is allele number 3 (the allele named 5) under-transmitted to affecteds, but that allele number 1 (named 3) is over-transmitted. Mendel uses allele numbers rather than allele names in the TDT table to save space. As usual, allele names appear near the top of the standard output file, Mendel13.out. The definition file, Def13.in, shows evidence of previous allele lumping via Analysis Option 16.

13.5 Germane Keywords

```
ANALYSIS_OPTION = TDT
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
GENDER_NEUTRAL
MAP_DISTANCE_UNITS
SAMPLES
```

Random Quotes

> It's not pining. It's passed on. This parrot is no more. It has ceased to be. It's expired and gone to meet its maker. This is a late parrot. It's a stiff. Bereft of life, it rests in peace. If you hadn't nailed it to its perch, it would be pushing up daisies. Its metabolic processes are now history. It's off the twig. It's kicked the bucket, shuffled off this mortal coil, rung down the curtain and joined the choir invisible. This is an ex-parrot.
>
> *John Cleese* in *Monty Python's Flying Circus, episode 8*

> Vigorous writing is concise. A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, for the same reason that a drawing should have no unnecessary lines and a machine no unnecessary parts. This requires not that the writer should make all sentences short, or that he avoid all detail and treat his subjects only in outline, but that every word should tell.
>
> *William Strunk Jr.* in *The Elements of Style*

14 Analysis Option 14: Penetrance Estimation

14.1 Background

Penetrance is the relationship of phenotype to genotype. Technically, a penetrance function defines the likelihood or probability of a particular phenotype given a particular genotype. We get in the habit of neglecting penetrances because they are ordinarily assumed to be 0 or 1 for marker loci. If we take typing error into account or consider disease loci rather than marker loci, then penetrances become an issue. Penetrance estimation has two purposes. First as a prelude to linkage analysis, it promotes better statistical modeling of a trait and addresses the question of whether a major gene model explains trait variation better than a strictly environmental model. Genotypes at the trait locus are unobserved. In this sense penetrance estimation forms part of classical segregation analysis. Segregation analysis is broader in scope because it involves estimation of transmission parameters as well. Thus, the gamete competition model of Option 8 falls under the umbrella of segregation analysis.

Another purpose of penetrance estimation is to provide a framework for testing association between trait values and the alleles at a candidate gene. In this setting the underlying alleles are at least partially observed, and one seeks to forge a connection between the trait and particular alleles. If alleles are fully observed and everyone is typed, then penetrance estimation can be done with standard commercial software that treats all cases as independent. For example, logistic regression can be done in this fashion with case-control data. However, if for any reason alleles are obscure, then one must either impute allele counts or include all possible genotypic combinations in likelihood evaluation. Analysis Option 20 takes the former approach for normally distributed traits. Option 20 has the further advantage of accommodating correlated environment and polygenic effects. For discrete traits, you will have to fall back on Option 14 if you want to assess simultaneously the role of genotype and predictors such as age and sex. Both Options 14 and 20 take into account ascertainment through designated probands. At this time neither option handles complex gene-by-gene or gene-by-environment interactions. Analysis Option 24 does handle interactions.

Analysis Option 14 replaces two earlier penetrance options that dealt separately with dichotomous traits and quantitative traits. In both of these options, the penetrance functions were special cases of the generalized linear models that pervade classical statistics [28, 78]. It is not terribly difficult to implement a wider range of generalized linear penetrance models [68]. Option 14 is our attempt to realize this agenda. In extending Mendel's capabilities for penetrance estimation, we have also revisited the question of how best to correct for ascertainment. Analysis Option 14 now allows multiple probands per pedigree and conditions on specially appended proband pedigrees during parameter estimation. Construction of the proband pedigrees is achieved by a hidden call to Analysis Option 21,

whose documentation you may want to read as background. Sections 0.5.6 and 0.9.2 also contain pertinent discussions of ascertainment, one of the most subtle issues in genetic epidemiology.

Model 2 of Option 14 allows you to generate penetrance files relevant to other Mendel options. Model 2 evaluates penetrances at a single point in parameter space and outputs the results locus by locus and person by person to a penetrance file. See Section 0.5.7 for a description of these files. The mechanics of invoking model 2 are not hard to master and make generating penetrance files easy. Of course, you can import penetrance files from other computing environments if you wish.

14.2 Appropriate Problems and Data Sets

This option is intended for pedigree data, with or without designated probands. Penetrance estimation is carried out separately for each locus common to both the definition and map files. If you are conducting classical segregation analysis with a hypothetical major locus, you should define two or more alleles at the locus in the definition file and blank phenotypes in the corresponding field of the pedigree file. Recall that only those trait genotypes consistent with recorded phenotypes in the corresponding pedigree field enter into Mendel's likelihood computations. In association analysis, it is appropriate to retain observed genotypes. In either case, you must include a separate quantitative variable to encode the trait phenotype. This is true even for dichotomous traits. If trait values are subject to censoring, then you must include a censoring indicator variable as well. If good allele frequencies are unavailable, then you can estimate allele frequencies jointly with penetrance parameters.

14.3 Generalized Linear Penetrance Models

Rather than dwell on the technical definition of a generalized linear model, we begin by referring the reader to Table 14.1. Here we see various distributional families displayed. The mean of each distribution is expressed in terms of one or two parameters. The primary parameter μ in each case is the focus of generalized linear modeling. Some of the distributions also require a scale parameter σ. For the binomial family, σ is replaced by the number of trials n; for the negative binomial family, n represents the required number of successes. Although the list of distributions in Table 14.1 is hardly exhaustive, it is wide enough to cover many interesting examples. Note that the displayed logistic distribution is continuous, not discrete. Logistic regression falls under the binomial distribution with one trial.

To preserve the parallel with linear regression, we could represent μ as an inner product $x^t p = \sum_i x_i p_i$ of a vector x of predictors and a vector of parameters p. This representation suffers from two defects. First, it is insufficiently general, and, second, it may violate natural

Table 14.1: Available Distributional Families

Name	Density	Mean	Variance	Link
Binomial	$\binom{n}{x}\mu^x(1-\mu)^{n-x}$	$n\mu$	$n\mu(1-\mu)$	Logit
Negative_Binomial	$\binom{n+x-1}{x}(1-\mu)^x\mu^n$	$\frac{n(1-\mu)}{\mu}$	$\frac{n(1-\mu)}{\mu^2}$	Logit
Poisson	$\frac{\mu^x}{x!}e^{-\mu}$	μ	μ	Log
Exponential	$\frac{1}{\mu}e^{-x/\mu}$	μ	μ^2	Log
Gamma	$\left(\frac{\sigma}{\mu}\right)^\sigma \frac{x^{\sigma-1}e^{-\sigma x/\mu}}{\Gamma(\sigma)}$	μ	$\frac{\mu^2}{\sigma}$	Log
Inverse_Gaussian	$\frac{e^{-(x-\mu)^2/(2x\mu^2\sigma^2)}}{\sqrt{2\pi x^3\sigma^2}}$	μ	$\mu^3\sigma^2$	Identity
Logistic	$\frac{1}{\sigma}\frac{e^{(x-\mu)/\sigma}}{[1+e^{(x-\mu)/\sigma}]^2}$	μ	$\frac{\pi^2\sigma^2}{3}$	Identity
Lognormal	$\frac{e^{-(\ln x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi x^2\sigma^2}}$	$e^{\mu+\sigma^2/2}$	$e^{2(\mu+\sigma^2)}-\mu^2$	Identity
Normal	$\frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$	μ	σ^2	Identity

bounds on μ. The remedy is to define a link function $f(\mu)$ connecting μ to $x^t p$. In practice, it is more convenient to consider the inverse link $f^{[-1]}(x^t p) = \mu$. Table 14.2 defines some link functions in common use. Mendel's default link function for each distributional family appears in the last column of Table 14.1.

The syntax for specifying a trait and its predictors is straightforward. Consider the sample commands

```
ANALYSIS_OPTION = Penetrances
QUANTITATIVE_TRAIT = LN_CHOL
PENETRANCE_MODEL = Normal :: Distribution
PENETRANCE_MODEL = Include :: Scale
PENETRANCE_MODEL = Identity :: Link_function
PREDICTOR = GRAND :: LN_CHOL
PREDICTOR = SEX :: LN_CHOL
PREDICTOR = SMOKING :: LN_CHOL
PREDICTOR = AGE :: LN_CHOL
PREDICTOR = DOMINANT :: LN_CHOL
```

from a hypothetical control file for Option 14. The second of these commands names the trait LN_CHOL to be analyzed. The third and fourth commands tell Mendel to model the trait as normally distributed and to estimate its common standard deviation (scale) as well

Table 14.2: Link and Inverse Link Functions

Name	Link	Inverse Link
Identity	$f(y) = y$	$f^{[-1]}(y) = y$
Logit	$f(y) = \ln\left(\frac{y}{1-y}\right)$	$f^{[-1]}(y) = \frac{e^y}{1+e^y}$
Log	$f(y) = \ln y$	$f^{[-1]}(y) = e^y$
Reciprocal	$f(y) = y^{-1}$	$f^{[-1]}(y) = y^{-1}$
Power	$f(y) = y^p$	$f^{[-1]}(y) = y^{1/p}$
Probit	$f(y) = \Phi^{[-1]}(y)$	$f^{[-1]}(y) = \Phi(y)$
Loglog	$f(y) = -\ln(-\ln y)$	$f^{[-1]}(y) = e^{-e^{-y}}$
Cloglog	$f(y) = \ln[-\ln(1-y)]$	$f^{[-1]}(y) = 1 - e^{-e^y}$

as its mean parameters. The fifth command sets the link function as the identity. Because the identity is the default for the normal family, the fifth command is, in fact, redundant. The remaining commands name predictors. Note that the grand mean (also known as the intercept) predictor is defined by default for each analyzed quantitative trait. Thus, the sixth command is also redundant. Categorical predictors such as SEX and SMOKING involve factors with several levels. Mendel reserves a parameter for each level. To ensure identifiability, the levels are tied together by a constraint setting the sum of the participating parameters equal to 0. For example with sex, the female and male parameters sum to 0. For a quantitative variable such as age, Mendel reserves one parameter with no implied constraints. Phenotypes defined at a secondary locus can also serve as trait predictors. If you want to use genotypes at the secondary locus as trait predictors, recode the genotypes as phenotypes. If you want to generate allele counts at the secondary locus, see the documentation of Analysis Option 20.

For example, if we denote the trait value by X and the regression coefficients by p_1 (grand mean), p_2 (female offset), p_3 (male offset), p_4 (nonsmoker), p_5 (light smoker), p_6 (heavy smoker), and p_7 (age), then the constraints $p_2 + p_3 = 0$ and $p_4 + p_5 + p_6 = 0$ are always enforced. Furthermore, for a female, nonsmoker, aged 50, we have the mean representation $f[\mathrm{E}(X)] = p_1 + p_2 + p_4 + 50\, p_7$ for link function $f(x)$.

You can impose one of five different genetic models at each locus analyzed. These models are named “allelic”, “dominant”, “recessive”, “genotypic”, and “imprinting”. An arbi-

trary number of alleles is possible for an allelic model. The dominant, recessive, genotypic, and imprinting models assume two alleles per locus and involve two, two, three, and four parameters, respectively. The dominant and recessive models assume that the first allele is the wild-type allele and the second allele is the disease or trait allele. Each allele of the allelic model contributes to the inner product $x^t p$ determining μ in proportion to the number of alleles present in a genotype. The genotypic model involves unordered genotypes and the imprinting model ordered genotypes. Depending on the link function selected, the allelic model can entail multiplicative action of the alleles. As with other grouped predictors, these parameters are tied together by a linear constraint forcing their sum to be 0.

There are a few other commands that come in handy. For example, the command

```
PENETRANCE_MODEL = Include :: Allele_frequencies
```

instructs Mendel to estimate allele frequencies along with penetrance parameters. It is also possible to estimate the required number of successes for a negative binomial model through the command

```
PENETRANCE_MODEL = Include :: Required_successes
```

Estimating the number of trials for a binomial model is forbidden because the set of possible values is discrete. If the required number of successes, the scale, or the allele frequencies are not estimated, then they are fixed. The default allele frequencies are taken from the definition file. The default number of trials for the binomial is 1. The default scale is chosen by computing the method of moments estimates of μ and σ from the sample moments of the trait via the theoretical mean and variance entries in Table 14.1. These estimates assume all pedigree members are unrelated and no ascertainment occurs. The default required number of successes for the negative binomial is determined in the same way and therefore can be fractional. Mendel automatically selects initial parameter values in its maximum likelihood search using these and other defaults. You can override the defaults for the scale, power of the power link function, number of trials, and required number of successes by issuing commands such as

```
PENETRANCE_MODEL = 0.5 :: Scale
PENETRANCE_MODEL = -2.0 :: Power
PENETRANCE_MODEL = 3 :: Trials
PENETRANCE_MODEL = 4 :: Required_successes
```

in the control file.

The likelihood surfaces encountered in penetrance estimation can be bumpy with multiple local maxima. Consequently, you may want to try several different starting points. You can override Mendel's parameter defaults through commands such as

```
PARAMETER_INITIAL_VALUE = -0.2 :: +
PARAMETER_MAX = 1.0 :: -
PARAMETER_MIN = 0.0 :: -
```

These three commands set the initial value of the parameter named "+" to -0.2 and the maximum and minimum values of the parameter named "–" to 1.0 and 0.0, respectively. In the commands just shown, we have selected an allelic contribution to μ for a locus with alleles named "+" and "–". Mendel names the parameters pertaining to the two alleles by the corresponding allele names. In defining new initial values, keep in mind that these should be consistent with any relevant parameter constraints.

Mendel names parameters based on the penetrance model selected. For example, as mentioned above, the command

```
PENETRANCE_MODEL = Include :: Allele_frequencies
```

causes Mendel to estimate allelic frequencies along with penetrances. When you employ this command, the parameter named `FREQ-1` holds the frequency of the first allele listed in the definition file at the relevant locus; the parameter `FREQ-2` holds the frequency of the second allele; and so forth. (If alleles at the locus are not listed in the definition file, then they are considered in the lexicographical order of their names, for example, 0, 1, 2, A, C, G, T.) Note that when you estimate allele frequencies, a command such as `PARAMETER_MIN = 0.5 :: FREQ-1` forces the first allele to be the most frequent allele. If the locus is biallelic, then entering the two commands

```
PARAMETER_INITIAL_VALUE = 0.9 :: FREQ-1
PARAMETER_EQUATION = FREQ-2 :: 0.01
```

causes an error message since it violates the implicit constraint that the two allele frequencies sum to 1. As another naming convention, a command such as

```
PREDICTOR = DOMINANT :: Health
```

creates two parameters named `1/1` and `1/2&2/2` that are constrained to sum to 0. Again, 1 refers to the first allele and 2 to the second allele as listed in the definition file at the relevant locus. These parameters hold the coefficients for the genotype groupings relevant to the dominance model. On the other hand, recessive models use two parameters named `1/1&1/2` and `2/2`; genotypic models use three parameters named `1/1`, `1/2`, and `2/2`; and imprinting models use four parameters named `1|1`, `1|2`, `2|1`, and `2|2`. In general, you can deduce Mendel's parameter naming conventions by running Option 14 and examining the output files.

As explained in Section 0.10.2, it is possible to control the search process in other ways. One example of such fine tuning is to reset the maximum step length. In Option 14

the default maximum step length is 1. You can override the default by issuing a command such as

```
MAX_STEP_LENGTH = 0.5
```

in the control file.

As a matter of convenience in Option 14, Mendel conducts for each model locus a likelihood ratio test of the specified genetic effect. Test results are deposited in the summary file. You may need to adjust the degrees-of-freedom and the asymptotic p-value for a given test if parameter estimates fall on lower or upper boundaries. Deviances will be included as part of the standard output file provided the command

```
DEVIANCES = True
```

is inserted in the control file. In this option, deviances are corrected for ascertainment via probands.

14.4 Survival Analysis Models

Survival analysis deals with nonnegative random variables T modeling random lifetimes. Let such a random variable T have density function $f(t)$ and distribution function $F(t)$. The hazard function

$$h(t) \;=\; \lim_{s \downarrow 0} \frac{\Pr(t < T \leq t+s \mid T > t)}{s} \;=\; \frac{f(t)}{1-F(t)}$$

represents the instantaneous rate of death at time t. Statisticians call the right-tail probability $1 - F(t) = S(t)$ the survival function and view $h(t)$ as the derivative

$$h(t) \;=\; -\frac{d}{dt} \ln S(t).$$

The cumulative hazard function $H(t) = \int_0^t h(s)ds$ obviously satisfies the identity

$$S(t) \;=\; e^{-H(t)}.$$

Many clinical trials involve right censoring. In other words, instead of observing a lifetime $T = t$, we observe $T > t$. Censored and ordinary data can be mixed in the same study. Generally, each observation T comes with a censoring indicator W. Mendel sets the penetrance function equal to: (a) the density of T at time t when W is missing or 0, (b) the distribution $\Pr(T \leq t)$ when W equals -1, and (c) the right-tail probability $S(t)$ when W equals 1. Interval censoring is not allowed.

It is simple to set up a parametric survival model in Mendel. For example, the commands

```
QUANTITATIVE_TRAIT = Onset
CENSORING_VARIABLE = Censor
PENETRANCE_MODEL = Exponential :: Distribution
```

in the control file tell Mendel that the trait is Onset, the censoring variable is Censor, and the survival times are exponentially distributed. The canonical Log link is the default link function.

Mendel also implements Cox's proportional hazards model. To allow longevity to depend on covariates as well as time, we now write the hazard function as

$$h(t) \quad = \quad \lambda(t)e^{x^t p},$$

where x and p are column vectors of predictors and regression coefficients, respectively. For instance, x might be $(1, d)^t$, where d indicates dosage of a life-prolonging drug. The function $\lambda(t)$ is known as the baseline hazard function. Straightforward integration shows that the cumulative hazard function amounts to

$$H(t) \quad = \quad \Lambda(t)e^{x^t p},$$

where

$$\Lambda(t) \quad = \quad \int_0^t \lambda(s)ds$$

is the baseline cumulative hazard. The baseline and baseline cumulative hazard functions must be input as separate variables in the pedigree file. In practice, the control file commands

```
BASELINE_HAZARD = Hazard
BASELINE_CUMULATIVE_HAZARD = Months
PENETRANCE_MODEL = Proportional_hazards :: Distribution
```

convey the choices for these two variables in the proportional hazards model.

14.5 Segregation Analysis Examples

Our first example features a binary disease trait. The fragment

```
ANALYSIS_OPTION = Penetrances
QUANTITATIVE_TRAIT = Health
PENETRANCE_MODEL = Binomial :: Distribution
PENETRANCE_MODEL = Include :: Allele_frequencies
PREDICTOR = Genotypic :: Health
```

```
PARAMETER_INITIAL_VALUE = -2.0 :: 1/1
PARAMETER_INITIAL_VALUE = 2.0 :: 2/2
PROBAND = 1
PROBAND_FACTOR = PROBAND
```

of the sample control file Control14a.in sets up a binomial model for the quantitative trait Health scored as 0 (normal) and 1 (sick) in the pedigree file. (Section 0.5.4.6 explains how to transform affection status into a 0/1 variable. In the current example, the dichotomous trait values at the locus DISEASE were previously transformed into the 0/1 quantitative variable Health.) The default number of trials is 1, and the default link function is the logit, so this example can be viewed as logistic regression over the hidden genotypes possible at the trait locus DISEASE. The above commands instruct Mendel to estimate allele frequencies, a grand trait mean, and an offset for each of the three genotypes. Anyone with a 1 at the PROBAND factor is considered a proband. The initial genotype offsets are -2, 0, and 2 for the three genotypes 1/1, 1/2, and 2/2. Here alleles 1 and 2 refer to the first and second alleles listed for locus DISEASE. These alleles have the names + and – in the definition file.

The pedigree number 25 output

```
PEDIGREE NUMBER 25 HAS 4 MEMBERS AND NAME A11.

      ID     PARENT    IDS     SEX   TWIN  DISEASE PROBAND Health

  1   100                      M                            1
  2   101                      F                            0
  3   200      100      101     M                    1       1
  4   201      100      101     M                            0
```

from Mendel14a.out illustrates how the various conventions appear in the pedigree file. Here genotypes at the disease locus are blank. This corresponds to the assumption in classical segregation analysis that genotypes are unobserved. Disease status is conveyed via the Health field and proband status through the PROBAND field. Note that the PROBAND field comes before the Health field because the former is a factor and the latter is a quantitative variable.

The summary output file Summary14a.out

```
LOCUS NAME:     DISEASE

PARAMETER      ESTIMATE     STANDARD ERROR
 GRAND         -2.4895        1.0297
 1/1           -2.8260        1.3088
 1/2            1.4121        1.1909
```

```
 2/2             1.4138         2.2305
 FREQ-1         0.95240        0.35968E-01
 FREQ-2         0.47596E-01    0.35968E-01

LNLIKELIHOOD = -124.4368
LIKELIHOOD RATIO TEST INAPPROPRIATE
```

records the maximum likelihood estimates for our model with genotype specific penetrances and variable allele frequencies. (The exponential notation appearing in Summary14a.out is changed here for ease of interpretation.) The estimated effects $\widehat{1/2} = 1.412$ and $\widehat{2/2} = 1.414$ of genotypes 1/2 and 2/2 agree surprising well with the assumption of a dominant model with reduced penetrance. Of course with a low frequency for the disease allele, the attached standard errors of 1.19 and 2.23 of these two estimates differ sharply, reflecting the far greater information available on heterozygotes compared to disease-allele homozygotes. Our initial guess of a disease-allele frequency of 0.010 is now superseded by the maximum likelihood estimate 0.048, with the wide standard error 0.036. Looking in the more detailed output file Mendel14a.out, we see parameter estimates that are fairly correlated. Rerunning this problem under a dominant model leads to virtually the same results.

Computation times for markers with many alleles become prohibitive, so you may want to combine alleles. P-values are suspect in the classical segregation analysis setting, where genotypes are unavailable. Thus, allele frequencies become irrelevant to disease status when allele or genotype parameters are omitted under the null model. For this reason, the file Summary14a.out explicitly says testing is inappropriate in this example. However, using the maximum likelihoods found in Mendel14a.out, the likelihood ratio statistic $2\times(129.917-124.436) = 10.962$ suggests that the genetic effects have substantial explanatory power. It is interesting that rerunning this example with allele frequencies fixed at 0.99 and 0.01 produces a likelihood ratio statistic of 3.022 with the unimpressive asymptotic p-value of 0.221.

Our second example considers a single large pedigree segregating hypercholesterolemia [91]. All cholesterol values have been log transformed to eliminate skewness. The lower half

```
ANALYSIS_OPTION = Penetrances
QUANTITATIVE_TRAIT = LN_CHOL
PENETRANCE_MODEL = Normal :: Distribution
!PENETRANCE_MODEL = Include :: Allele_frequencies
PENETRANCE_MODEL = Include :: Scale
PREDICTOR = DOMINANT :: LN_CHOL
PARAMETER_INITIAL_VALUE = -0.2 :: 1/1
PARAMETER_INITIAL_VALUE = +0.2 :: 1/2&2/2
```

```
PROBAND = P
PROBAND_FACTOR = PROBAND
```

of the sample control file Control14b.in sets up a Gaussian model for the quantitative trait LN_CHOL. The default link function is the identity, so this example can be viewed as ordinary regression over the hidden genotypes possible at the trait locus HC. The above commands instruct Mendel to estimate a grand trait mean and dominant genetic effects. Note that we have commented out the command to estimate allele frequencies. This allows likelihood ratio testing. Anyone with a P at the PROBAND factor is considered a proband. The initial genotype offsets are -0.2 for genotype 1/1 and 0.2 for the two genotypes 1/2 and 2/2. Genotypes at the HC locus are unobservable, so everyone has a blank phenotype at the HC locus. Nevertheless, you must include the HC locus in the definition file. The possible predictor AGE is ignored.

The summary output file Summary14b.out

```
LOCUS NAME:     HC

PARAMETER       ESTIMATE     STANDARD ERROR
 GRAND           5.6203       0.18906E-01
 1/1            -0.28127      0.16585E-01
 1/2&2/2         0.28127      0.16585E-01
 SCALE           0.15299      0.11750E-01

LNLIKELIHOOD =   4.288781
LIKELIHOOD RATIO STATISTIC = 56.16831
DEGREES OF FREEDOM =    1
ASYMPTOTIC P-VALUE =  0.66525E-13
```

records the maximum likelihood estimates. (The exponential notation in Summary14b.out is changed here for ease of interpretation.) The estimated genotypic offsets $\widehat{1/1} = -0.281$ and $\widehat{1/2} = \widehat{2/2} = 0.281$ under the dominant model agree well with the initial values given in the control file. To achieve the quick, smooth convergence evident in the standard output file Mendel14b.out, initial estimates should be as good as possible. The likelihood ratio statistic $2 \times (4.290 + 23.795) = 56.170$ is drawn from a χ^2 distribution with one degree-of-freedom under the null hypothesis. This is extremely significant. The estimated scale (standard deviation) of 0.153 demonstrates good separation between the normal and (dominant) disease phenotypes. The difference in estimated means $2 \times (0.281) = 0.562$ is almost $4 \times 0.153 = 0.612$, that is, four standard deviations. In Mendel14b.out one can see that parameter estimates tend to be more weakly correlated than in our first sample problem for this option.

14.6 Penetrance File Construction Examples

Our first model 2 example prepares a penetrance file for the sample problem of Analysis Option 22. Note that for the simple uniform penetrance function used here, one could assign penetrance values directly from the control file. See the sample control file Control22.in. This example illustrates the model 2 alternative of letting Mendel construct an explicit penetrance file. The operative part of the control file Control14c.in

```
NEW_PENETRANCE_FILE = Pen22.out
ANALYSIS_OPTION = Penetrances
MODEL = 2
AFFECTED_LOCUS_OR_FACTOR = TRAIT
AFFECTED = 2
TRANSFORM = Indicator :: 0-1Var
QUANTITATIVE_TRAIT = 0-1Var
PENETRANCE_MODEL = Binomial :: Distribution
PENETRANCE_MODEL = Identity :: Link_function
PREDICTOR = Dominant :: 0-1Var
PARAMETER_INITIAL_VALUE =  0.52 :: GRAND
PARAMETER_INITIAL_VALUE = -0.47 :: 1\1
PARAMETER_INITIAL_VALUE =  0.47 :: 1\2&2\2
```

transforms the variable named 0-1Var so that it codes for affection status as determined by the locus named TRAIT. Affecteds (phenotype 2) are assigned the value 1, normal individuals are assigned the value 0, and missing values are preserved. These maneuvers are necessary because Option 14 requires a quantitative trait variable. The variable 0-1Var enters with all values missing and is defined during execution.

Choosing the binomial distribution with one trial (the default) and identity link function leaves only specification of the success probability μ. Subsequent commands in Control14c.in instruct Mendel to write μ as the sum of a grand mean plus a dominant offset. Thus, genotype 1\1 has penetrance constant $\mu = 0.52 - 0.47 = 0.05$ and genotypes 1\2 and 2\2 have common penetrance constant $\mu = 0.52 + 0.47 = 0.99$. In this dominant model, the sum of the offsets is again 0.

It is worth checking the generated penetrance file to see whether penetrances have been computed correctly. In this case the file reads

```
108,     109,          TRAIT,          d\d,  1.0000000     ,
108,     109,          TRAIT,          d\D,  1.0000000     ,
108,     109,          TRAIT,          D\D,  1.0000000     ,
108,     110,          TRAIT,          d\d,  1.0000000     ,
108,     110,          TRAIT,          d\D,  1.0000000     ,
108,     110,          TRAIT,          D\D,  1.0000000     ,
```

```
108,     108,          TRAIT,          d|d,  0.0500000     ,        2
108,     108,          TRAIT,          d|D,  0.9900000     ,        2
108,     108,          TRAIT,          D|d,  0.9900000     ,        2
108,     108,          TRAIT,          D|D,  0.9900000     ,        2
```

for pedigree 108. Section 0.5.7.1 describes the different fields in this file. The recorded penetrances are correct even though some of them are listed for ordered genotypes.

Our second example prepares the penetrance file for sample problem 2b under Analysis Option 2. The control file Control14d.in tells most of the story. The fragment

```
NEW_PENETRANCE_FILE = Pen2b.out
ANALYSIS_OPTION = Penetrances
MODEL = 2
QUANTITATIVE_TRAIT = ACE
PENETRANCE_MODEL = Normal :: Distribution
PENETRANCE_MODEL = Include :: Scale
PREDICTOR = GENOTYPIC :: ACE
PARAMETER_INITIAL_VALUE =  1.7010 :: GRAND
PARAMETER_INITIAL_VALUE =  0.7148 :: SCALE
PARAMETER_INITIAL_VALUE = -2.031  :: 1\1
PARAMETER_INITIAL_VALUE = -0.151  :: 1\2
PARAMETER_INITIAL_VALUE =  2.182  :: 2\2
```

from Control14d.in names the new penetrance file, declares that model 2 of Option 14 is pertinent, and sets up a normal mixture model for the quantitative trait ACE. Subsequent commands in the control file inform Mendel that the mean of each of the three genotypes is written as the sum of a grand mean plus a genotypic offset. Thus, genotype 1\1 has trait mean $1.701 - 2.031 = -0.330$, genotype 1\2 has mean $1.701 - 0.151 = 1.550$, and genotype 2\2 has mean $1.701 + 2.182 = 3.883$. Again the offsets sum to 0. The three genotypic distributions share the common standard deviation 0.7148. There is no need to name the link function here since the default identity link is appropriate.

Penetrances are computed for locus ACE_LOC, the sole model locus defined by the intersection of the loci in the definition and map files. All phenotypes at the locus ACE_LOC are left blank in the pedigree file Ped14c.in to indicate that no genotypes are disallowed for any person. Mendel evaluates penetrances at the given parameter values and deposits the results in the new penetrance file. The part of this file relevant to pedigree 4

```
4,       1,        ACE_LOC,          1\1,  1.0000000     ,
4,       1,        ACE_LOC,          1\2,  1.0000000     ,
4,       1,        ACE_LOC,          2\2,  1.0000000     ,
4,       2,        ACE_LOC,          1\1,  1.0000000     ,
4,       2,        ACE_LOC,          1\2,  1.0000000     ,
```

```
4,      2,      ACE_LOC,        2\2,  1.0000000     ,
4,      4,      ACE_LOC,        1\1,  0.2661005     ,
4,      4,      ACE_LOC,        1\2,  0.0003409     ,
4,      4,      ACE_LOC,        2\2,  0.5836350E-11 ,
4,      3,      ACE_LOC,        1\1,  0.4798276     ,
4,      3,      ACE_LOC,        1\2,  0.0035561     ,
4,      3,      ACE_LOC,        2\2,  0.5374364E-09 ,
```

shows that individuals 1 and 2 lack observed trait values and are assigned unit penetrances. Note that penetrances maybe written in Fortran exponential notation. For example, $0.5374364\text{E-}09 = 0.0000000005374364$.

Our third example prepares the penetrance file for example 7b of Option 7. In this genetic risk prediction example, the variable LN_CPK measured on a single crucial female provides information on who is a carrier for the X-linked recessive mutation causing Duchenne muscular dystrophy (DMD). Most of the features in the control file Control14e.in have already been discussed in the last two examples. The abbreviated version

```
NEW_PENETRANCE_FILE = Pen7b.out
ANALYSIS_OPTION = Penetrances
MODEL = 2
QUANTITATIVE_TRAIT = LN_CPK
PENETRANCE_MODEL = Normal :: Distribution
PENETRANCE_MODEL = Include :: Scale
PREDICTOR = DOMINANT :: LN_CPK
PARAMETER_INITIAL_VALUE =  1.835  :: GRAND
PARAMETER_INITIAL_VALUE =  0.3255 :: SCALE
PARAMETER_INITIAL_VALUE = -0.265  :: 1\1
PARAMETER_INITIAL_VALUE =  0.265  :: 1\2&2\2
```

tells us that LN_CPK is normally distributed with common standard deviation 0.3255 and mean $\mu = 1.835 - 0.265 = 1.570$ for the recessive genotype 1\1 and common mean $\mu = 1.835 + 0.265 = 2.100$ for the dominant genotypes 1\2 and 2\2.

14.7 Survival Analysis Example

Although artificial, our survival analysis example illustrates how to set up a problem. The partial list of commands

```
ANALYSIS_OPTION = Penetrances
QUANTITATIVE_TRAIT = Months
CENSORING_VARIABLE = Censor
BASELINE_HAZARD = Hazard
```

```
BASELINE_CUMULATIVE_HAZARD = Months
PENETRANCE_MODEL = Proportional_hazards :: Distribution
PREDICTOR = ALLELIC :: Months
```

from Control14f.in identifies the trait variable Months, the censoring variable Censor, the baseline hazard variable Hazard, and the baseline cumulative hazard variable Months. In this problem, the hazard function is constant, and the trait and the baseline cumulative hazard coincide. In other words, survival times are exponentially distributed. Analyzing the data with an exponential model yields the same estimates. The predictors include a grand mean (always present) and additive allelic effects. In a more complicated and realistic problem, the user must enter in the pedigree file values for the baseline hazard and baseline cumulative hazard specific to each person. Only complete cases are included in penetrance evaluation. Pedigree data and ascertainment via probands are allowed.

The output

```
LOCUS NAME:     Dummy

PARAMETER      ESTIMATE     STANDARD ERROR
 GRAND        -4.9129        0.21648
 1            -0.27890       0.10825
 2             0.27890       0.10825

LNLIKELIHOOD =  -191.4082
LIKELIHOOD RATIO STATISTIC =  5.67076
DEGREES OF FREEDOM =    1
ASYMPTOTIC P-VALUE =  0.01725
```

from the summary file Summary14f.out shows that genotype (coded as 1-1 or 2-2 in Ped14f.in) has a significant impact on survival time. It is possible to check for outliers by printing the deviance of each patient.

14.8 Germane Keywords

```
ANALYSIS_OPTION = Penetrances
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
ALLELE_SEPARATOR
BASELINE_CUMULATIVE_HAZARD
BASELINE_HAZARD
CENSORING_VARIABLE
DEVIANCES
FEMALE_MUTATION_RATE
```

```
MALE_MUTATION_RATE
MODEL
NEW_PENETRANCE_FILE
PARAMETER_INITIAL_VALUE
PARAMETER_MAX
PARAMETER_MIN
PENETRANCE_MODEL
PREDICTOR
PROBAND
PROBAND_FACTOR
QUANTITATIVE_TRAIT
TRANSFORM
```

Random Quote

Population, when unchecked, increases in a geometric ratio. Subsistence increases only in an arithmetic ratio. A slight acquaintance with numbers will show the immensity of the first power in comparison with the second.

Thomas Malthus in *An Essay on the Principle of Population*

Whenever the literary German dives into a sentence, that is the last you are going to see of him till he emerges on the other side of his Atlantic with his verb in his mouth.

Mark Twain in *A Connecticut Yankee in King Arthur's Court*

. . . Dons and parsons live by presenting the sufferings of others, and that is regarded as religious, uncommonly deep religion even; for the religion of the congregation is nothing but hearing this presented. As a religion, *charmante*, just about as genuine as tea made from a bit of paper which once lay in a drawer beside another bit of paper which had once been used to wrap up a few dried tea-leaves from which tea had already been made three times.

Soren Kierkegaard in his journals

Use the active voice. Put statements in positive form. Use definite, specific, concrete language. Omit needless words.

William Strunk Jr. in *The Elements of Style*

When I am in the company of scientists, I feel like a shabby cleric who has strayed by mistake into a drawing room of dukes.

W.H. Auden

If you can't say anything good about someone, sit right here by me.

Alice Roosevelt Longworth

15 Analysis Option 15: Ethnic Admixture

15.1 Background

People of mixed ethnicity enter many genetic studies. The potential for confounding ethnicity with disease risk is well known and has led to the development of widely used statistical methods, such as the TDT. The current analysis option is predicated on the view that ethnicity is simply another predictor (covariate) that affects phenotypes. To adjust for ethnicity, we need genome-wide estimates of fractional ancestry. Option 15 is designed to provide these estimates. The estimates are applicable as linear predictors in Options 14, 19, and 20.

Particularly useful for dense genome-wide data, Option 15 can also perform a principal components analysis (PCA) using the top ancestry informative markers (AIMs) in the data set. If the user specifies to use N SNPs and calculate p principal components, Mendel first finds the N SNPs that are most informative at distinguishing the subpopulations [89] and then uses them to perform the PCA. These principal components can be used as predictors in GWAS analysis to avoid population stratification issues.

There are also interesting anthropological applications of Option 15. The classical work is well summarized in the book [17] and the papers [30, 88]. Readers should note that the recently introduced technique of admixture mapping [80] exploits the variation in fractional ancestries across the genome of each subject in a gene mapping study. Our more modest goal is simply to adjust for ethnic admixture in a global fashion. The program Admixture [2] extends the functionality of this analysis option.

15.2 Appropriate Problems and Data Sets

Mendel requires standard, text-based files for estimation of fractional ancestry. However, for finding the best AIMs and for PCA, Mendel can use either text-based data sets or the binary data files described in Section 0.6.

When `MODEL = 1`, which is always the default, Option 15 estimates fractional ancestry. This model can be used with full pedigrees, nuclear families, random individuals, or any combination thereof [97]. The usual text-based definition, map, and pedigree files come into play except that all markers should be unlinked. This assumption promotes computational efficiency and allows large pedigrees to be processed. Markers should be chosen to discriminate between the contributing ethnic groups. One highly polymorphic marker near each chromosome end is ideal. The definition file should list reliable estimates of the allele frequencies for each marker for each ethnic group.

Mendel operates pedigree by pedigree and estimates the fractional ancestry of each pedigree founder unless the founder is already assigned to a particular ethnic group in

the pedigree file. One can steer these maximum likelihood estimates toward reasonable average values by defining independent and identically distributed Dirichlet priors for the unassigned founders. Once the fractional ancestries are estimated for the founders, these are propagated to the entire pedigree by recursively setting each child's fractional ancestries to the average of his or her parent's fractional ancestries.

Model 2 of Option 15 again estimates fractional ancestry and uses the standard text-based files. However, model 2 treats all individuals as unrelated and relies on an EM (expectation-maximization) algorithm [104, 114] to estimate fractional ancestries. It ignores untyped people, requires codominant markers, and delivers less precise estimates than model 1 on pedigree data. In compensation it is much faster and can exploit linked markers in linkage equilibrium.

Model 3 uses an extension of an algorithm described by Rosenberg et al. [89] to quickly find the N most informative AIMs within the data set. Our extension allows for non-uniform population sampling. This model can use text-based or binary data sets. Some small fraction of the typed individuals in the data set must have their ethnicity specified. These individuals are used to determine the top AIMs. Mendel will issue a warning if there are too few typed individuals in any subpopulation.

Finally, model 4 performs PCA. Model 4 has the same restrictions as model 3. Indeed, model 4 first carries out the same analysis as model 3 and then uses the top N AIMs found to perform the PCA. The user specifies the p principal components output as variables for each individual in a new data set. This new data set can then be used for a GWAS conditioned on ethnicity.

15.3 Input Files

The control file, Control15a.in,

```
DEFINITION_FILE = Def15a.in
MAP_FILE = Map15a.in
PEDIGREE_FILE = Ped15a.in
POPULATION_FACTOR = RACE
POPULATIONS = 2
AFFECTED_LOCUS_OR_FACTOR = TRAIT
AFFECTED = 2
!
! Output Files
!
OUTPUT_FILE = Mendel15a.out
SUMMARY_FILE = Summary15a.out
NEW_DEFINITION_FILE = Def15a.out
NEW_PEDIGREE_FILE = Ped15a.out
```

```
!
! Analysis Parameters
!
ANALYSIS_OPTION = Ethnic_Admixture
MODEL = 1
!POPULATION_PRIOR_COUNT = 2.0 :: SPANISH
!POPULATION_PRIOR_COUNT = 3.0 :: AMERIND
```

illustrates how easy it is to set up a problem. The most important thing to notice is the required definition of the number of populations and the factor in the definition file defining the populations. Some of the other commands are optional. Because the fractional ancestries are deposited in the summary file, there is no need to define new definition and pedigree files unless you want ready-to-run definition and pedigree files with the fractional ancestries appended to each person record as additional variables. Population prior counts can be omitted if you prefer pure maximum likelihood estimates. The prior counts are used to construct the Dirichlet priors for the various unassigned pedigree founders. Section 0.5.4.5 describes the default format for inserting multiple population-specific frequencies for each defined allele.

The control file, Control15d.in,

```
! Input Files
!
DEFINITION_FILE = Def15c.in
PEDIGREE_FILE = Ped15c.in
SNP_DEFINITION_FILE = SNP_def15c.in
SNP_DATA_FILE = SNP_data15c.bin
POPULATION_FACTOR = RACE
POPULATIONS = 2
!
! Output Files
!
OUTPUT_FILE = Mendel15d.out
SUMMARY_FILE = Summary15d.out
NEW_PEDIGREE_FILE = Ped15d.out
NEW_DEFINITION_FILE = Def15d.out
NEW_SNP_DEFINITION_FILE = SNP_def15d.out
NEW_SNP_DATA_FILE = SNP_data15d_bin.out
!
! Analysis Parameters
!
ANALYSIS_OPTION = Ethnic_Admixture
MODEL = 4
DESIRED_PREDICTORS = 1000 :: AIMs
```

```
PRINCIPAL_COMPONENTS = 2
MIN_SUCCESS_RATE_PER_SNP = 0.0
MIN_SUCCESS_RATE_PER_INDIVIDUAL = 0.0
```

illustrates the few additional commands used to set the number of AIMs and PCs. First, note that this data set uses binary files as input. A population factor is again defined. As mentioned, a small fraction of the individuals must have their population specified. These individuals are used to find which AIMs are best. The command

```
DESIRED_PREDICTORS = 1000 :: AIMs
```

tells Mendel to find the top 1000 AIMs, which is the default. (The word "AIMs" is not case sensitive.) These top AIMs are then used in the PCA. Although Mendel's PCA analysis is very efficient, using multiple CPU cores simultaneously if they are available, selecting a huge number of AIMs will cause the PCA to take considerable time. Less than 10,000 AIMs should be sufficient even for genome-wide data and several principal components. Finally, the value assigned to the keyword `PRINCIPAL_COMPONENTS` sets the number of principal components Mendel will output as variables in the `NEW_PEDIGREE_FILE`.

Since the output pedigree file may list the individuals in a different order than the input pedigree file, a new SNP data file must also be generated. The new definition file includes the definitions for the new variables containing the principal components. Note that it is usually important to filter out any SNPs and individuals that genotype poorly by setting the keywords `MIN_SUCCESS_RATE_PER_SNP` and `MIN_SUCCESS_RATE_PER_INDIVIDUAL` to values close to 1; their default values are 0.98. However, in the tiny, demonstration data set used here, we set these keywords to 0.0 to prevent any SNPs or individuals from being removed.

15.4 Example

The summary file, Summary15a.out,

```
 ADMIXTURE COEFFICIENTS PEDIGREE BY PEDIGREE

 PEDIGREE     PERSON      POPULATION     ESTIMATED    STD ERROR OF
   NAME        NAME          NAME        PROPORTION     ESTIMATE

 1            1            AMERIND        0.5213        0.2195
 1            1            SPANISH        0.4787        0.2195

 1            9            AMERIND        0.8497        0.1932
 1            9            SPANISH        0.1503        0.1932

 1            10           AMERIND        0.4508        0.2700
```

```
1           10          SPANISH       0.5492       0.2700

1           12          AMERIND       0.5082       0.2162
1           12          SPANISH       0.4918       0.2162

1           15          AMERIND       0.3434       0.2345
1           15          SPANISH       0.6566       0.2345

1           17          AMERIND       0.3899       0.2279
1           17          SPANISH       0.6101       0.2279
```

is mostly self-explanatory. The model 1 results are listed for each pedigree member for the postulated ancestral populations Spanish and Amerindian. For some people, standard errors of the estimated proportions are left blank. This is the case, for instance, with a pedigree founder who is preassigned to a specific ethnic group or whose estimated proportions suggest no admixture. Reported standard errors may differ slightly on different computer platforms due to the delicate numerical operations employed in finding them. Average fractional ancestries are given for each pedigree and for the sample overall. It is possible for exceptionally complex pedigrees to be skipped in analysis. If this happens, then the omitted pedigrees are mentioned at the bottom of the standard output file along with their computational complexities. See Section 0.10.1 for suggestions on how to deal with omitted pedigrees.

On the same data, model 2 gives the estimates

```
ADMIXTURE COEFFICIENTS PEDIGREE BY PEDIGREE

PEDIGREE     PERSON      POPULATION     ESTIMATED
  NAME        NAME          NAME        PROPORTION

1            1            AMERIND        0.3384
1            1            SPANISH        0.6616

1            9            AMERIND        0.8496
1            9            SPANISH        0.1504

1            10           AMERIND        0.3803
1            10           SPANISH        0.6197

1            12           AMERIND        0.5082
1            12           SPANISH        0.4918

1            15           AMERIND        0.3435
1            15           SPANISH        0.6565
```

```
1           17          AMERIND       0.3899
1           17          SPANISH       0.6101
```

Note that model 2 does not deliver standard errors. In this example, the people listed here are all founders. Models 1 and 2 give essentially the same answers for the fully typed founders 9, 12, 15, and 17. Founders 1 and 10 have missing genotypes and therefore different estimates under the two models. Because the model 1 estimates take into account the genotypes of descendants, they are better. Untyped people with no designated ancestral population are assigned uniform ancestries in model 2.

Model 3 finds the top AIMs in a data sets and outputs them to the summary file. For example, Summary15c.out

```
MARKER        INFORMATION GAIN

rs2816             5.38209
rs1800498          2.43301
rs2763             1.89671
rs4884             0.51594
rs2065160         -0.42728
rs2695            -0.42728
rs4646            -1.17309
rs17203           -1.31714
rs3309            -1.81230
```

lists the markers in descending order of their information gain [89].

Model 4 outputs the requested principal components in a new pedigree file. For example, the last few entries in Ped15d.out are

```
6 , 106 ,     ,     , M , ,   , 0 , ,   -0.0000,   -0.0000,
6 , 117 ,     ,     , F , , 1 , 1 , ,   -0.1038,    0.0537,
6 , 121 ,     ,     , F , , 1 , 1 , ,   -0.1930,    0.1804,
6 , 118 ,     ,     , M , , 1 , 1 , ,   -0.1316,    0.1833,
6 , 108 , 106 , 117 , F , , 2 , 1 , ,   -0.0881,    0.0780,
6 , 109 , 106 , 117 , F , , 2 , 1 , ,   -0.1282,    0.0199,
6 , 111 , 106 , 117 , M , , 1 , 1 , ,   -0.1136,   -0.0205,
6 , 112 , 106 , 117 , F , , 1 , 1 , ,   -0.1469,    0.0681,
6 , 119 , 118 , 112 , M , ,   , 1 , ,   -0.0889,    0.1710,
6 , 120 , 118 , 112 , M , ,   , 1 , ,   -0.1946,    0.0843,
6 , 113 , 106 , 117 , M , , 2 , 1 , ,   -0.1196,    0.1033,
6 , 114 , 106 , 117 , M , , 1 , 1 , ,   -0.0608,    0.2570,
6 , 115 , 106 , 117 , M , ,   , 0 , ,   -0.0000,   -0.0000,
6 , 123 , 115 , 121 , F , ,   , 1 , ,   -0.1576,    0.0002,
6 , 122 , 115 , 121 , M , , 1 , 1 , ,   -0.1081,    0.1037,
```

The last two columns are the two requested principal component values for these individuals. Since this is part of a binary data set, all genotypes are in the SNP data file in a binary representation.

15.5 Germane Keywords

```
ANALYSIS_OPTION = Ethnic_Admixture
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
DESIRED_PREDICTORS
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
MODEL
NEW_DEFINITION_FILE
NEW_PEDIGREE_FILE
NEW_SNP_DATA_FILE
NEW_SNP_DEFINITION_FILE
POPULATION_FACTOR
POPULATION_PRIOR_COUNT
POPULATIONS
PRINCIPAL_COMPONENTS
```

Random Quotes

If (the educated person) engages in controversy of any kind, his disciplined intellect preserves him from the blundering discourtesy of ... less educated minds, who, like blunt weapons, tear and hack, instead of cutting clean, who mistake the point of an argument, waste their strength on trifles, misconceive their adversary, and leave the question more involved than they find it.

John Henry Newman in *The Idea of a University*

In order to be intellectually honest, you have to have an intellect in the first place.

Flannery O'Connor

For children are innocent and love justice, while most of us are wicked and naturally prefer mercy.

G.K. Chesterton

I have endured a great deal of ridicule without much malice; and have received a great deal of kindness, not quite free from ridicule. I am used to it.

Abraham Lincoln

16 Analysis Option 16: Combining Alleles

16.1 Background

Rare marker alleles are a fact of life in most genetic studies and often cause statistical and computational headaches. The large sample theory approximations of some linkage and association tests break down in the presence of rare alleles. Computation times can also balloon out of control. For these reasons, it is often prudent to amalgamate rare alleles. Analysis Option 16 facilitates this. No statistical analysis is undertaken.

16.2 Appropriate Problems and Data Sets

Any data set with rare alleles is fair game for Option 16, which creates new definition and pedigree files with consolidated alleles. Model 1, the default model of Option 16, operates recursively at each locus common to both the definition and map files. If the frequency of the currently rarest allele falls below a fixed threshold, or if the number of alleles exceeds a fixed maximum, then the two currently rarest alleles are combined. Codominant genotypes in the pedigree file and genotype sets attached to particular phenotypes in the definition file are adjusted accordingly. The allele frequencies input in the definition file determine which alleles are consolidated. Consolidation proceeds until the rarest allele frequency falls above the threshold, and the number of alleles no longer exceeds the maximum. Note that this is one analysis option where Mendel does not estimate allele frequencies, so it is crucial to include frequencies in the definition file. If you want finer and more predictable control of the allele combining process, you might wish to use model 2 rather than model 1 of Option 16. Model 2 simply combines alleles that have identical frequencies in the input definition file. No recursion is involved. Of course, you will have to set allele frequencies accordingly. If you list the same average frequency for each allele in a group of alleles, then it is possible to preserve the correct summed frequency for their combination.

16.3 Input Files

The map file limits what loci are acted on. Recombination fractions can be left blank. In the control file, some version of the following commands

```
NEW_DEFINITION_FILE = Def16.out
NEW_PEDIGREE_FILE = Ped16.out
```

are necessary in running Option 16. These two output files contain the redefined alleles, phenotypes, and genotypes in correct Mendel format. The names of the new definition and pedigree files should not coincide with the names of the old definition and pedigree files.

If you desire to change either the frequency threshold or the maximum number of alleles from their default values of 0.05 and 10, then commands such as

```
ALLELE_COMBINING_THRESHOLD = 0.1
MAX_COMBINED_ALLELES = 8
```

should be inserted in the control file.

16.4 Examples

In our Option 16 example, the map file, Map16.in, contains all of the loci named in the definition file. Hence, all loci are subjected to allele consolidation. Consider the locus information

```
MARKER1          ,AUTOSOME, 3
  1        , 0.1800000
  2        , 0.5500000
  3        , 0.2700000
MARKER2          ,AUTOSOME, 3,  5
  2        , 0.4600000,     ! ORIGINAL ALLELE NUMBERS:  2
  3        , 0.4600000,     ! ORIGINAL ALLELE NUMBERS:  3
  4        , 0.0800000,     ! ORIGINAL ALLELE NUMBERS:  1  4
 A-A        ,  1
   4|4
 A-C        ,  2
   3|4
   4|3
 B-B        ,  1
   2|2
 B-D        ,  2
   2|4
   4|2
 C-D        ,  2
   3|4
   4|3
MARKER3          ,AUTOSOME, 3
  p        , 0.3200000
  q        , 0.6100000
  r        , 0.0700000
MARKER4          ,AUTOSOME, 3
  1        , 0.1400000,     ! ORIGINAL ALLELE NUMBERS:  1
  3        , 0.7900000,     ! ORIGINAL ALLELE NUMBERS:  3
  5        , 0.0700000,     ! ORIGINAL ALLELE NUMBERS:  2  4  5
```

```
DISEASE         ,AUTOSOME, 2,  3
  +       , 0.9999000
  -       , 0.0001000
 NORMAL    ,  3
   +|+
   +|-
   -|+
 AFFECTED  ,  2
   +|-
   -|+
 CARRIER   ,  2
   +|-
   -|+
```

contained in the output file Def16.out.

Comparison with the original definition file, Def16.in, demonstrates that some loci show consolidation and some do not. When alleles are consolidated, the new allele names are positive integers. In spite of the fact that the disease allele is rare, the disease locus undergoes no consolidation because it possesses just two alleles. The first and third markers also show no consolidation. The second and fourth markers do undergo consolidation. At the second marker, allele A and allele D, the first and fourth alleles, are consolidated. Note that Mendel interprets the string A-A as a phenotype rather than as a genotype because the allele separator has been set to \ rather than -. For this reason, the phenotype A-A suffers no change when the new files are constructed. However, the allele contributing to A-A changes from A to 4. Genotypes such as A-C that were formerly represented by one unordered genotype are now represented by two ordered genotypes. This substitution in the definition file affects all loci, not just those undergoing allele consolidation. Substitution of ordered genotypes for unordered genotypes does no harm and reflects how Mendel operates internally. At the fourth marker, alleles b, d, and e, the second, fourth, and fifth alleles, are consolidated. Here Mendel first combines alleles b and d and then combines their union with allele e.

A consistent story

```
      5,FAMILY1
FATHER  ,       ,       ,M     ,      ,      ,      ,      ,      ,CARRIER
MOTHER  ,       ,       ,F     ,      ,      ,      ,      ,      ,CARRIER
CHILD1  ,FATHER ,MOTHER ,F     ,      ,1\2   ,A-A   ,p\r   ,3\5   ,NORMAL
CHILD2  ,FATHER ,MOTHER ,M     ,      ,3\3   ,A-C   ,q\q   ,1\5   ,AFFECTED
CHILD3  ,FATHER ,MOTHER ,M     ,      ,3\3   ,A-C   ,q\q   ,1\5   ,NORMAL
     4,FAMILY2
DAD     ,       ,       ,M     ,      ,      ,      ,      ,      ,CARRIER
MOM     ,       ,       ,F     ,      ,      ,      ,      ,      ,CARRIER
```

```
KID1    ,DAD    ,MOM    ,M     ,      ,1\1   ,B-B   ,p\p   ,5\5   ,NORMAL
KID2    ,DAD    ,MOM    ,F     ,      ,2\3   ,A-C   ,q\r   ,1\5   ,AFFECTED
     4,FAMILY3
PA      ,       ,       ,M     ,      ,      ,      ,      ,      ,NORMAL
MA      ,       ,       ,F     ,      ,      ,      ,      ,      ,NORMAL
BROTHER ,PA     ,MA     ,M     ,      ,2\2   ,C-D   ,p\r   ,5\3   ,NORMAL
SISTER  ,PA     ,MA     ,F     ,      ,1\2   ,B-D   ,q\q   ,1\5   ,NORMAL
```

is told by the pedigree data output in Ped16.out. Here phenotypes such as A-C at the second locus are preserved, while genotypes such as d\d at the fourth locus get renamed using positive integers. In the later case, allele d constitutes part of allele 5, so the new genotype is 5\5.

The same outcomes can be achieved if we use model 2 and manipulate the frequencies in the definition file. For instance, to direct Mendel to combine alleles A and D of MARKER2, we could set their allele frequencies to the average value 0.04. Since allele A has frequency 0.03 and allele D frequency 0.05, the average value 0.04 preserves the correct frequency for the combined allele.

16.5 Germane Keywords

```
ANALYSIS_OPTION = Combining_Alleles
ALLELE_COMBINING_THRESHOLD
MAX_COMBINED_ALLELES
NEW_DEFINITION_FILE
NEW_PEDIGREE_FILE
```

Random Quotes

> “But I can assure you,” she added, “that Lizzy does not lose much by not suiting his fancy; for he is a most disagreeable, horrid man, not at all worth pleasing. So high and so conceited that there was no enduring him! He walked here and he walked there, fancying himself so very great! Not handsome enough to dance with! I wish you had been there, my dear, to have given him one of your set downs. I quite detest the man.”
>
> *Jane Austen* in *Pride and Prejudice*

> Elinor agreed with it all, for she did not think he deserved the compliment of rational opposition.
>
> *Jane Austen* in *Sense and Sensibility*

> I’m not a snob. Ask anybody. Well, anybody who matters.
>
> *Simon Lebon* of *Duran Duran*

17 Analysis Option 17: Gene Dropping

17.1 Background

In many circumstances, it is useful to simulate genetic data consistent with a postulated map. For example, you might want to generate p-values empirically or to estimate the power of a collection of pedigrees to detect linkage. Gene dropping randomly fills in genotypes subject to prescribed allele frequencies, a given genetic map, and Hardy-Weinberg and linkage equilibrium. Analysis Option 17 performs gene dropping and generates a new pedigree file for further analysis. The missing data pattern in the new pedigrees can mimic the input pedigrees, or the user can specify a new pattern.

17.2 Appropriate Problems and Data Sets

The raw material for gene dropping consists of sets of pedigrees and loci. People within the pedigrees must be assigned either blank phenotypes or Mendelian consistent phenotypes. Gene dropping is carried out independently of observed phenotypes at those loci common to the definition and map files. By varying the content of the map file, you can choose exactly which loci to subject to gene dropping. Phenotypes at the remaining loci of the definition file are left untouched. Simulated genotypes rather than simulated phenotypes are reported. There are no limits on the complexity of the pedigrees or the number of loci. You can use founders from different populations, provided these populations are defined. Each founder should be assigned to a population; any unassigned founders are assumed to come from the first population. Follow the population conventions described in Section 0.5.4.5.

17.3 Input Files

Option 17 requires typical definition, map, and pedigree files. The allele frequencies and recombination fractions listed in the definition and map files determine how gene dropping is performed. If a locus in the definition file lacks one or more allele frequencies, then Mendel will estimate all allele frequencies at this locus by gene counting, blithely assuming that typed people are unrelated. The partial content

```
NEW_PEDIGREE_FILE = Ped17a.out
ANALYSIS_OPTION = Gene_dropping
REPETITIONS = 3
SEED = 21547
MODEL = 2
KEEP_FOUNDER_GENOTYPES = False           ! the default value
```

```
MISSING_DATA_PATTERN = Existing_Data     ! the default value
MISSING_AT_RANDOM = 0.0                  ! the default value
GENE_DROP_OUTPUT = Unordered             ! the default value
```

of Control17a.in displays several important keywords. The first of these determines the pedigree output file, the second the analysis option, and the third the number of independent simulated output pedigrees per each input pedigree. In this analysis option, the default value for REPETITIONS is 1. Each repetition of a pedigree is given a unique name based on the original name of the pedigree. The keyword SEED discussed in Section 0.11 allows you to simulate unique data during each MENDEL run.

The remaining keywords modify the output format. For example, model 1, the default, interleaves simulated pedigrees just as they appear on input; model 2 places all simulated pedigrees corresponding to a given input pedigree in a contiguous block. In analogy to photocopying multiple copies of a large document, model 1 is collated output, model 2 is not. The keyword KEEP_FOUNDER_GENOTYPES indicates whether existing founder genotypes should be left untouched during simulation. If you elect this possibility by setting this keyword to true, then the genotypes of non-founders are simulated conditional on the given founder genotypes. The default value for KEEP_FOUNDER_GENOTYPES is false.

The keywords MISSING_DATA_PATTERN and MISSING_AT_RANDOM determine where missing data is placed in the new output pedigrees. Setting MISSING_DATA_PATTERN to "None" (case insensitive) mandates full genotype data in the output pedigrees. Setting the value to "Existing_Data" (case insensitive and the default value) causes each output pedigree to mimic the missing data pattern of the corresponding input pedigree. Finally, setting MISSING_DATA_PATTERN to the name of a locus or factor in the Definition file causes the missing data pattern at that locus or factor to be replicated at all loci. Superimposed on the overall missing data pattern can be randomly missing genotypes. The rate of randomly missing genotypes is set by the keyword MISSING_AT_RANDOM, which has default value zero, indicating no randomly missing data.

Some examples of how to setup various missing data structures may be helpful. To have the simulated genotypes exactly mimic the missing data pattern in the existing data you can use the commands

```
MISSING_DATA_PATTERN = Existing_Data
MISSING_AT_RANDOM = 0.0
```

or since these are the default values, you could simply leave these keywords out of your control file. To have all simulated genotypes displayed except for a random 2% that will be left blank, use the commands

```
MISSING_DATA_PATTERN = None
MISSING_AT_RANDOM = 0.02
```

Finally, if you want founders to have all missing genotypes, then create a factor that has values only for non-founders. If that factor is named NON-FDR and you use the commands

```
MISSING_DATA_PATTERN = NON-FDR
MISSING_AT_RANDOM = 0.03
```

then all output pedigrees will have genotypes only for non-founders and within the non-founders 3% of the genotypes will be missing at random.

The keyword GENE_DROP_OUTPUT controls the style of the simulated genotypes output to the new pedigree file. The values of this keyword are case insensitive. Its default value of "Unordered" mandates that simulated genotypes are reported as unordered allele pairs. The keyword value "Ordered" tells Mendel to output ordered genotypes for non-founders. When GENE_DROP_OUTPUT has the value "Sourced", each founder gene is assigned a unique label at each locus, and all genotypes are constructed using these labels. Non-founders are assigned ordered sources. The implied full source information conveys the identity-by-descent (IBD) states of all relative pairs at all loci. An example of this style of output is included in the next section. For the Sourced output style, the assigned values of the keywords KEEP_FOUNDER_GENOTYPES, MISSING_AT_RANDOM, and MISSING_DATA_PATTERN are ignored, and Mendel uses the values:

```
KEEP_FOUNDER_GENOTYPES = False
MISSING_DATA_PATTERN = None
MISSING_AT_RANDOM = 0.0
```

Finally, when GENE_DROP_OUTPUT has the value "Population", sourced labels are replaced by numbers indicating the ethnic origins of the corresponding founder genes. This convention allows one to visualize how ethnic sources change along simulated chromosomes. Of course, this necessitates a factor that designates the founders' populations. For a small example data set that uses population labeling, see data set 17b. Note that Control17b.in uses the commands

```
POPULATIONS = 3
POPULATION_FACTOR = Ethnic
```

to define the number of populations and the factor holding the population labels.

Each gene-dropping run that sets an identical initial seed for the random number generator via the keyword SEED (for example, SEED = 12345) will generate the same simulated underlying genotype structure, only differing perhaps in the labels used in the output. For example, if using the same initial seed you run Mendel twice, once with GENE_DROP_OUTPUT = Ordered and once with GENE_DROP_OUTPUT = Sourced, then one pedigree output file will list genotypes and the other will list the founder labels (i.e., IBD

structure) that correspond to those genotypes. If you use `GENE_DROP_OUTPUT = Unordered` instead of Ordered, then the genotypes output will be unordered, but the founder labels will still refer to the ordered genotypes, that is, to specific alleles.

17.4 Example

Consistent with the above content from Control17a.in, the entire output file Ped17a.out becomes:

```
0:Bush  , George  ,         ,         , M  ,  , a/a   , 217/217   , 1946
0:Bush  , Laura   ,         ,         , F  ,  , a/b   , 213/213   , 1946
0:Bush  , Jenna   , George  , Laura   , F  ,  , a/a   ,           , 1981
0:Bush  , Barbara , George  , Laura   , F  ,  , a/a   , 213/217   , 1981
1:Bush  , George  ,         ,         , M  ,  , a/b   , 217/217   , 1946
1:Bush  , Laura   ,         ,         , F  ,  , a/a   , 213/217   , 1946
1:Bush  , Jenna   , George  , Laura   , F  ,  , a/b   ,           , 1981
1:Bush  , Barbara , George  , Laura   , F  ,  , a/b   , 213/217   , 1981
2:Bush  , George  ,         ,         , M  ,  , a/a   , 213/217   , 1946
2:Bush  , Laura   ,         ,         , F  ,  , a/a   , 213/217   , 1946
2:Bush  , Jenna   , George  , Laura   , F  ,  , a/a   ,           , 1981
2:Bush  , Barbara , George  , Laura   , F  ,  , a/a   , 213/217   , 1981
0:Clinto, Bill    ,         ,         , M  ,  , a/b   , 213/217   , 1946
0:Clinto, Hillary ,         ,         , F  ,  , a/a   , 213/217   , 1947
0:Clinto, Chelsea , Bill    , Hillary , F  ,  , a/a   , 213/217   , 1980
1:Clinto, Bill    ,         ,         , M  ,  , a/a   , 213/217   , 1946
1:Clinto, Hillary ,         ,         , F  ,  , a/a   , 213/217   , 1947
1:Clinto, Chelsea , Bill    , Hillary , F  ,  , a/a   , 213/213   , 1980
2:Clinto, Bill    ,         ,         , M  ,  , a/a   , 217/217   , 1946
2:Clinto, Hillary ,         ,         , F  ,  , a/a   , 213/217   , 1947
2:Clinto, Chelsea , Bill    , Hillary , F  ,  , a/a   , 213/217   , 1980
```

In this case, each pedigree is replicated three times. Note that only Jenna Bush is missing marker data and that genotypes rather than phenotypes are reported at the Egomania locus. Quantitative variables are not changed from the input pedigree file.

If GENE_DROP_OUTPUT is set to Sourced, then the Bush section of the pedigree output becomes:

```
0:Bush   , George  ,         ,         , M  ,  , 1/2   , 1/2       , 1946
0:Bush   , Laura   ,         ,         , F  ,  , 3/4   , 3/4       , 1946
0:Bush   , Jenna   , George  , Laura   , F  ,  , 4|1   , 4|1       , 1981
0:Bush   , Barbara , George  , Laura   , F  ,  , 4|1   , 4|1       , 1981
1:Bush   , George  ,         ,         , M  ,  , 1/2   , 1/2       , 1946
1:Bush   , Laura   ,         ,         , F  ,  , 3/4   , 3/4       , 1946
```

```
1:Bush   , Jenna   , George  , Laura   , F  ,  , 4|2   , 4|2       , 1981
1:Bush   , Barbara , George  , Laura   , F  ,  , 4|2   , 4|2       , 1981
2:Bush   , George  ,         ,         , M  ,  , 1/2   , 1/2       , 1946
2:Bush   , Laura   ,         ,         , F  ,  , 3/4   , 3/4       , 1946
2:Bush   , Jenna   , George  , Laura   , F  ,  , 3|2   , 3|2       , 1981
2:Bush   , Barbara , George  , Laura   , F  ,  , 4|2   , 4|2       , 1981
```

The two genes at each locus descending from the founding father George are assigned the unique labels 1 and 2. Laura's gene labels are 3 and 4. All non-founder genotypes are constructed from the founder labels to indicate the ultimate source of each gene segregating in the pedigree. Clearly, Jenna and Barbara share 2 genes IBD at each locus in the first and second replicate pedigree and 1 gene IBD at each locus in the third. It is not too surprising to see the same pattern of inheritance at both loci in these few replicates since the Map17a.in file specifies that these loci are quite tightly linked.

In practice, you can use simulated data to estimate the power of a pedigree sample to detect linkage. If you simulate n pedigrees for every pedigree input, then you can estimate the expected LOD score curve by dividing the LOD score curve of the entire simulated data by n. To estimate empiric p-values, say for significance in the gamete competition model, you must analyze each replicate of the original sample separately and extract from the output files the likelihood ratio statistics. Of course, this requires a lot of computing and special software to find the likelihood ratio statistic amid the clutter of the standard output file.

17.5 Germane Keywords

```
ANALYSIS_OPTION = Gene_Dropping
GENE_DROP_OUTPUT
KEEP_FOUNDER_GENOTYPES
MISSING_AT_RANDOM
MISSING_DATA_PATTERN
MODEL
NEW_PEDIGREE_FILE
REPETITIONS
SEED
```

Random Quotes

Arthur Dent: You know, it's at times like this ... that I really wish I'd listened to what my mother told me when I was young.

Ford Prefect: Why, what did she tell you?

Arthur Dent: I don't know, I didn't listen.

Douglas Adams in *The Hitchhiker's Guide to the Galaxy*

18 Analysis Option 18: Combining Loci

18.1 Background

Despite the limited information content of SNPs (single nucleotide polymorphisms), their sheer abundance throughout the genome and their low mutation rates make them ideal markers. To increase their information content, one can combine neighboring SNPs from a narrow genetic region. Owing to phase uncertainties, the resulting "super-locus" will not have codominant alleles. Of course, one can sometimes assign haplotypes reliably using information on surrounding pedigree members, but there is no guarantee this will work, and assigned haplotypes are often just best guesses. This drawback suggests that it would be preferable to deal directly with super-loci, formed by combining non-recombinant loci, just as we do with ordinary markers with dominant and recessive alleles. To carry out this agenda, it is necessary to prepare definition and pedigree files displaying super-locus phenotypes and their associated genotypes. Analysis Option 18 is designed to achieve this. (This option was previously referred to as Combining_SNPs, and although the old name will continue to work as a keyword value, the new name is preferred.)

18.2 Appropriate Problems and Data Sets

This analysis option is particularly pertinent to association testing and estimation of haplotype frequencies. To a lesser extent it will be useful in linkage mapping. Up to four loci can be combined at a time. Mendel allows loci to be combined with up to nine codominant alleles each. Balanced against this flexibility are the following restrictions: (a) no markers may have dominant or recessive alleles, (b) no ordered genotypes may appear in the pedigree file, and (c) no recombination may occur between marker pairs. If a super-locus later presents Mendelian inconsistencies, the most likely cause is apt to be the failure of condition (c). Once you have combined loci, we recommend applying model 1 of Analysis Option 5 to check for Mendelian errors.

Several analysis options can use the output of Option 18. For instance, Option 6 will estimate haplotype frequencies, Option 8 will look for disease-haplotype association via the gamete competition model, and Option 14 will estimate haplotype-specific penetrances. Combining loci is even valuable in the variance-components Options 19 and 20. Once super-loci are constructed, Option 16 will combine rare haplotypes. This is helpful in reducing the computational complexity of the gamete competition model and penetrance estimation and in guaranteeing that the large sample assumptions of the corresponding options are satisfied.

18.3 Input Files

Loci to be combined are specified either by super-locus entries in the definition file or by a sliding window of loci. Section 0.5.4.4 describes the simple format for super-locus entries. An example can be drawn from the bottom of Def18.in

```
S1+S2, SUPERLOCUS, 2
SNP1
SNP2
3+4+5, SUPERLOCUS, 3
SNP3
SNP4
SNP5
```

When you enter super-locus entries in the definition file, the map file is effectively ignored by Option 18. On the other hand, if you want to define a sliding window of loci, then the map file and definition files jointly define the model loci along which the window is slid. In this case, enter a command such as

```
NUMBER_OF_MARKERS_INCLUDED = 2
```

in the control file to specify the window width. Mendel deposits data on the various super-loci in new definition and pedigree files. These novel output files are named via the instructions

```
NEW_DEFINITION_FILE = Def18.out
NEW_PEDIGREE_FILE = Ped18.out
```

in the control file. The names of the new definition and pedigree files should not coincide with the names of the old definition and pedigree files.

Allele frequencies are omitted in the new definition file unless you insert the command

```
EQUILIBRIUM_FREQUENCIES = True
```

in the control file telling Mendel to deposit linkage equilibrium frequencies. Model 1 of Analysis Option 6 allows you to test whether the hypothesis of linkage equilibrium is justified. From a Bayesian perspective, it is also sometimes useful to steer haplotype frequencies toward linkage equilibrium values. These values must be present in the definition file to achieve either end.

18.4 Example

Our Option 18 example provides an interesting contrast between the first pedigree

```
5, FAMILY_1
1 ,  ,  ,M ,,1\1 ,3   ,1\2 ,    ,1\1
2 ,  ,  ,F ,,1\2 ,2   ,1\2 ,1\1 ,1\2
3 ,1 ,2 ,F ,,1\2 ,2   ,2\2 ,1\1 ,1\1
5 ,1 ,2 ,F ,,1\2 ,2   ,1\2 ,1\2 ,1\1 ,AFFECTED
6 ,1 ,2 ,F ,,1\1 ,3   ,1\1 ,1\2 ,1\2
```

of the old pedigree file, Ped18.in, and its transformation

```
       5,FAMILY_1
1 ,  ,  ,M ,,1\1 ,3   ,1\2 ,    ,1\1 ,12   ,12??11  ,        ,
2 ,  ,  ,F ,,1\2 ,2   ,1\2 ,1\1 ,1\2 ,1212 ,121112  ,        ,
6 ,1 ,2 ,F ,,1\1 ,3   ,1\1 ,1\2 ,1\2 ,1122 ,111212  ,        ,
5 ,1 ,2 ,F ,,1\2 ,2   ,1\2 ,1\2 ,1\1 ,1212 ,121211  ,AFFECTED,
3 ,1 ,2 ,F ,,1\2 ,2   ,2\2 ,1\1 ,1\1 ,1212 ,221111  ,        ,
```

in the new pedigree file, Ped18.out. (Spaces have been removed from the above depiction of the output file for presentation purposes.) Per the instructions in the super-locus entries of the definition file Def18.in, the X-linked SNP1 and SNP2 have been combined, and the autosomal SNP3, SNP4, and SNP5 have been combined. To show the versatility of Mendel, genotypes for SNP2 are coded as 1 (1\1), 2 (1\2), and 3 (2\2). Super-locus phenotypes are constructed by concatenating genotypes at the participating loci. The allele separator is deleted from each genotype before concatenation, and each allele symbol is converted to the numbered position of the allele in the definition file. For clarity in this example, the alleles are already denoted by the appropriate numbers. As an example, the autosomal genotypes 1\1, 1\2, and 1\2 determine the combination phenotype 111212. For X-linked loci, male haplotypes are constructed by concatenating alleles. Thus, the male hemizygous genotypes 1 and 2 (coded as 1\1 and 3 for SNP1 and SNP2 above) determine the haplotype 12 for the father of the displayed nuclear family. In the file Ped18.out, super-locus phenotypes come after the original loci and before the disease status factor.

The snippet

```
SNP5            ,AUTOSOME, 2
  1       , 0.9000000
  2       , 0.1000000
S1+S2           ,X-LINKED, 4,  5
  1+1
  2+1
  1+2
```

```
 2+2
1122      ,  1
    1+2|1+2
12        ,  1
    1+2|1+2
1212      ,  4
    1+1|2+2
    2+1|1+2
    1+2|2+1
    2+2|1+1
21        ,  1
    2+1|2+1
22??      ,  4
    2+1|2+1
    2+1|2+2
    2+2|2+1
    2+2|2+2
```

from the new definition file, Def18.out, shows the pre-existing SNP5 and the new super-locus S1+S2 based on combining SNP1 and SNP2. Each allele at super-locus S1+S2 is a haplotype represented by two allele numbers separated by a + sign. The displayed haplotype frequencies conform to linkage equilibrium. You probably should re-estimate haplotype frequencies via Option 6. The five phenotypes at super-locus S1+S2 are exactly those that appear in the new pedigree file. To list more would be pointless. Each genotype listed for a phenotype is an ordered genotype. Thus, 2+1|1+2 and 1+2|2+1 both appear as genotypes consistent with the phenotype 1212.

You might enjoy rerunning this example with a sliding window rather than a defined collection of loci to combine. To make this change, you will need to remove the super-locus entries in the definition file Def18.in and add (or uncomment) a command such as

```
NUMBER_OF_MARKERS_INCLUDED = 2
```

in Control18.in. If you select a sliding window of size 2 as suggested in the sample control file, then some of the adjacent locus pairs involve one autosomal and one X-linked marker and accordingly will be omitted in the new definition and pedigree files.

18.5 Germane Keywords

```
ANALYSIS_OPTION = Combining_Loci
EQUILIBRIUM_FREQUENCIES
NEW_DEFINITION_FILE
NEW_PEDIGREE_FILE
NUMBER_OF_MARKERS_INCLUDED
```

19 Analysis Option 19: Variance Components (Polygenic and QTL Mapping)

19.1 Background

The method of variance components initiated by R. A. Fisher [34] occupies a curious niche in genetic epidemiology. Until about 1960, variance component analysis was the best available method for understanding the genetics of human traits. It was and still is remarkable for its power to extract useful information from a few scraps of familial correlations. With the advent of human gene mapping and the intense debate surrounding IQ, the limitations of heritability studies became clear. Except for its sanctuaries in plant and animal breeding, heritability analysis almost vanished from the genetics scene. However, its fortunes improved with the push to map quantitative trait loci (QTLs) [3, 4, 11, 37, 47, 71, 90]. Variance component models are successful in this context because they account for both major genes and polygenic background in a surprisingly parsimonious manner. These models are also flexible enough to handle covariates (predictors in our terminology), multivariate traits, correlated environment, gene-by-gene interaction (epistasis), and the combined effects of a single gene on multiple traits (pleiotropy).

Option 19 is designed for both classical heritability studies and QTL mapping. It incorporates factor analytic decompositions that make QTL mapping more parsimonious in capturing pleiotropy. Particular stress is put on robustness, outlier detection, and modeling generality. This does not imply that Option 19 is a panacea. It neglects, for instance, nonlinear mean effects and longitudinal random effects. Furthermore, in QTL mapping you should use Mendel in conjunction with the program SimWalk, which generates for large pedigrees the required conditional kinship coefficients at each putative QTL map position. For the sake of brevity, we make no attempt here to teach the theory behind classical biometrical genetics. For that purpose, consult the references [17, 23, 59] at your leisure.

19.2 Appropriate Problems and Data Sets

Option 19 implements many of the tools of classical biometrical genetics. It estimates mean components, variance components, and heritabilities. It also maps genes for quantitative traits. Your study sample should consist of some combination of nuclear families and extended pedigrees. Computation times and memory requirements per pedigree both increase as the cube of the product of the number of traits times the number of people. Extended pedigrees are particularly useful if you want to explore complicated models with lots of variance components. You define a model for your data by specifying linear predictors of trait means and by choosing from a limited menu of variance decompositions. Two of these decompositions capture additive and dominance polygenic effects. These effects

are invariably combined with random environment effects and often with household effects. Inclusion of identical twins is very helpful in distinguishing between additive polygenic effects and random environment. Household effects tend to be confounded with dominance effects unless the study sample contains extended pedigrees. If you are interested in QTL mapping, it is also possible to specify an additive QTL effect. In QTL mapping, you must supply Mendel with definition and map files and, when large pedigrees are involved, with a SimWalk generated file of conditional kinship coefficients.

Table 19.1: Variance Components

Variance Component	Type of Variation	Source of Coefficients	Large Pedigrees?
Additive	Polygenic	Mendel	Yes
Dominance	Polygenic	Mendel	Yes
QTL	Major Gene	Mendel or SimWalk	With SimWalk
X-Additive	Polygenic	Mendel	Yes
X-QTL	Major Gene	Mendel	No
Household	Non-genetic	Household Field	Yes
Environmental	Non-genetic	Mendel	Yes

Table 19.1 summarizes Mendel's available variance components. Each of these represents a source of random variation contributing additively and independently to the trait or traits under consideration. The X-linked contributions will be new to many users. These are discussed in Lange and Sobel (2006) [70]. Mendel adopts a model in which the covariance between the X-linked contribution u_i to trait u in person i and the X-linked contribution v_j to trait v in person j takes the form

$$\mathrm{Cov}(u_i, v_j) \quad = \quad \Phi_{ij}\sigma_{uv}. \tag{3}$$

Here Φ_{ij} is the X-linked kinship coefficient between i and j, and σ_{uv} is a covariance parameter to be estimated. For the X-QTL component, Φ_{ij} is a conditional kinship coefficient as described in the documentation of Analysis Option 10. If $j = i$ is a male, the covariance formula (3) continues to hold with the understanding that $\Phi_{ij} = 1$. If $j = i$ is a female, then

$$\mathrm{Cov}(u_i, v_i) \quad = \quad [\Phi_{ii} + p(1 - \Phi_{ii})]\sigma_{uv},$$

where $\Phi_{ii} = \frac{1}{2}$ for i non-inbred and p is an additional parameter confined to the interval [0,1]. In Mendel's standard output, p is designated as XX_ADD or XX_QTL.

The observant reader will notice that the two X-linked components assign greater trait variance to males than to females. The source of this extra variance is X-inactivation,

which tends to smooth trait contributions to females. When you use the X-QTL component in the absence of the X-additive polygenic component, you run the risk of detecting spurious linkage to a major gene. If male variance exceeds female variance, then the evidence for an X-linked QTL can be driven entirely by this difference rather than excess correlations between relatives tied to a specific X-linked marker. Including an X-additive component safeguards against this possibility. A simpler and probably better safeguard is to standardize each trait separately by sex. Because of the limited map length of the X chromosome, the polygenic assumption of many different X-linked loci acting additively and independently to influence a trait is suspect. Thus, in X-linked QTL mapping, we recommend degendering the trait and including additive, X-QTL, and environmental variance components. Dominance and household components can be used as the need arises. If you get a strong linkage signal, then see if it persists after adding an X-additive component.

Keep in mind that a trait value for a person is discarded whenever one or more of its predictors are missing. Mendel will use a quantitative trait for a person even if he or she is missing other quantitative traits. Ascertainment correction proceeds by conditioning on proband trait values. Because Mendel assumes that the trait values within a pedigree follow a multivariate normal distribution, it is a good idea to subject quantitative traits to a power or log transform as instructed in Section 0.5.4.6. These transformations tend to eliminate skewness. If your traits display excess kurtosis, you can substitute the multivariate t distribution for the multivariate normal distribution in analysis. This works best for univariate traits since all subtraits of a multivariate trait are assumed to involve the same level of kurtosis.

19.3 Input Files

Option 19 uses standard pedigree and variable files. A map file is unnecessary unless you want to map a QTL. The control file Control19a.in

```
DEFINITION_FILE = Def19a.in
PEDIGREE_FILE = Ped19a.in
OUTPUT_FILE = Mendel19a.out
ANALYSIS_OPTION = Variance_Components
PROBAND_FACTOR = PROBAND
PROBAND = P
QUANTITATIVE_TRAIT = Left
QUANTITATIVE_TRAIT = Right
PREDICTOR = Sex :: Left
PREDICTOR = Sex :: Right
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
OUTLIERS = True
```

is fairly typical. It specifies a definition file because of the necessity for a factor defining proband status. Multiple probands per pedigree are permitted. Two variables, Left and Right, are designated as traits, and the mean of each trait is assumed to be a constant, the so called grand mean, offset by an increment depending on sex. It is possible to construct more complicated mean models using the syntax described in Analysis Option 14. The exception to this rule is that the major gene models discussed there are not allowed. If you want to use allele counts at a candidate locus as predictors, then run Analysis Option 20 directly or take advantage of the imputed counts it deposits in a new pedigree file.

The control file in this example also postulates an additive polygenic effect and a random environmental effect. The menu of possible covariance classes is listed in Table 19.1. If you invoke a household effect, then you must specify a corresponding household field by defining the keyword `GROUP_FACTOR` in the control file. In contrast to Analysis Option 2, the keyword `NUMBER_OF_MARKERS_INCLUDED` is irrelevant in QTL mapping. If you want to conduct a single-marker analysis, then restrict the map file to a single marker.

Finally, Mendel is instructed to flag outlier people and pedigrees. The additional commands

```
PEDIGREE_CUT_POINT = 0.1
PERSON_CUT_POINT = 0.02
```

in the control file change the defaults of 0.05 and 0.01 for the cutoffs determining what fraction of pedigrees and people are flagged under the null hypothesis implied by your model. If you want to switch to a more robust analysis involving the multivariate t distribution, insert the command

```
MULTIVARIATE_NORMAL = False
```

in the control file.

19.4 Examples

Total finger ridge count is a highly heritable trait for which an abundance of pedigree data exists. The pedigree file Pedigree19a.in contains Tables 1 and 3 of the classical reference [46]. Three extra sibships from Table 2 of the same source appear as ascertained nuclear families at the bottom of Pedigree19a.in. These families were selected on the basis of an exceptionally high total finger ridge count for the first listed sib.

Pedigree19a.in actually records ridge counts for each hand separately. These bivariate data are ideal for exploring the biometrical tools of Mendel. One reason for studying multivariate traits is to assess the degree to which the component traits are under common genetic and/or environmental control. To perform such an analysis, it is helpful to decompose the theoretical covariances between different traits of two related individuals in much

the same way that univariate traits are typically decomposed [60]. The control file Control19a.in discussed earlier instructs Mendel to implement a bivariate model with additive polygenes and random environment.

The following output on left and right-hand counts is near the bottom of Mendel19a.out:

SUMMARY FOR MEAN PARAMETERS

TRAIT	PARAMETER	PREDICTOR	ESTIMATE	STD ERR
Left	1	GRAND	63.6014	2.4108
Left	3	FEMALE	-4.3150	1.4847
Left	4	MALE	4.3150	1.4847
Right	2	GRAND	66.5788	2.4394
Right	5	FEMALE	-3.6234	1.5074
Right	6	MALE	3.6234	1.5074

SUMMARY FOR COVARIANCE PARAMETERS

TRAIT	TRAIT	VARIANCE	ESTIMATE	STD_ERR_1	STD_ERR_2
Left	Left	ADDITIVE	647.5283	66.1414	66.1845
Right	Left	ADDITIVE	653.8506	65.4853	65.5265
Right	Right	ADDITIVE	660.2347	68.9137	68.9606
Left	Left	ENVIRONMENTAL	32.1066	7.8986	7.9284
Right	Left	ENVIRONMENTAL	-6.8826	5.1432	5.1927
Right	Right	ENVIRONMENTAL	38.4807	10.2620	10.4761

TRAIT	TRAIT	TOTAL VARIANCE OR COVARIANCE
Left	Left	679.6349
Right	Left	646.9681
Right	Right	698.7154

These maximum likelihood estimates are quite revealing. First, it is clear that the two ridge counts are highly heritable traits. More surprising is the very high additive genetic correlation and very low environmental correlation between the traits. If these data are credible, then ridge counts on the left and right hands are basically determined by the same set of genes. Furthermore, because the right-left environmental covariance is less than one standard error away from zero, the environmental determinants for the two hands may act independently. Note that the total variances reported permit immediate calculation of heritabilities. Mendel always reports the wide-sense heritability of any univariate trait determined by additive, dominance, household, or environmental effects.

The outlier analysis in Mendel19a.out suggests that the model fits the data extremely well. Only one outlier person and one outlier pedigree are noted. The outlier statistics' p-values of 0.0051 and 0.048 are consistent with what one would expect for a study sample of this size. None of the empirical distribution function tests for the pedigree outlier statistics is significant at the 0.1 level. Further analysis of these data show that including dominance effects does not improve the fit of the model to the data. However, a likelihood ratio test shows that the sex effect is significant. Observe that the female and male predictors are constrained to sum to zero.

Our second example deals with QTL mapping for an anonymous bivariate quantitative trait. Here we apply the novel factor analytic decomposition discussed in Bauman et al. (2005) [8] to the QTL contribution to the trait. Classical factor analysis explains the covariation among the components of a random vector by approximating the vector by a linear transformation of a small number of uncorrelated factors. In variance component analysis and QTL mapping, factor analysis makes it possible to discover the coordinated control of multiple traits by common environment, common polygenes, or a single major gene. To instruct Mendel to limit the QTL contribution to a specific number of factors, say one, you must insert the line

```
COVARIANCE_FACTORS = 1 :: Qtl
```

in the control file. If you want to limit another contribution, say the additive polygenic effect to two factors, insert the line

```
COVARIANCE_FACTORS = 2 :: Additive
```

in the control file. The number of factors for any effect must not exceed the number of traits. The default number of factors equals this upper bound.

In addition to the advantage of parsimony, factor analytic decompositions have the desirable side effect of stabilizing maximum likelihood estimation for multivariate traits. We treat QTL factor analytic decompositions somewhat differently from factor analytic decompositions for other effects. When you specify the number of QTL factors, then Mendel considers all QTL decompositions with this or a smaller number of factors. Not only is such a tactic biologically revealing, but it also minimizes the chance that Mendel's maximum likelihood algorithm will be trapped at a local maximum on a bumpy likelihood surface.

The instructions

```
ANALYSIS_OPTION = Variance_Components
MAP_DISTANCE_UNITS = cM
GRID_INCREMENT = 0.5
QUANTITATIVE_TRAIT = Trait1
QUANTITATIVE_TRAIT = Trait2
```

```
PREDICTOR = SEX :: Trait1
PREDICTOR = AGE :: Trait1
PREDICTOR = BMI :: Trait1
PREDICTOR = SEX :: Trait2
PREDICTOR = AGE :: Trait2
PREDICTOR = BMI :: Trait2
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
COVARIANCE_CLASS = Qtl
```

from Control19b.in set up the analysis model. The mean of each trait is parameterized by a grand mean (always present) and regression coefficients on sex, age, and body mass index (BMI). Three variance effects — additive polygenes, random environment, and an additive QTL effect — contribute to the two traits. Here we use the command `GRID_INCREMENT = 0.5` to specify computation of a sequence of QTL location score curves at map points spaced 0.5 cM apart. (Recall that `GRID_INCREMENT` is in the same units as the distances in the Map file.) Examination of these curves shows where a putative QTL might reside. There is one curve for each permitted level of the QTL factor.

QTL mapping requires standard definition, map, and pedigree files. It also employs two files exported by SimWalk under its kinship option. In the current example, SimWalk's coefficient file is identified by the command

```
COEFFICIENT_FILE = Coefficient19b.in
```

in Control19b.in. Coefficient19b.in contains the name of a second file, IKKEY-19.TXT, generated by SimWalk. You must move both of these files to your working directory prior to execution of Mendel. You can change the name of the coefficient file, but we recommend that you retain the name of the key file.

On small pedigrees, Option 19 will substitute exact conditional kinship coefficients for the approximate ones exported by SimWalk. This task is accomplished by a hidden call to Option 10 for kinship matrix computation. Since these exact computations are quite time-consuming, for Option 19 the default value of the pedigree-complexity keyword `MAX_ADJUSTED_MEIOSES` is set to 12. You can override this low default by issuing a command such as

```
MAX_ADJUSTED_MEIOSES = 14
```

in the control file. You should avoid values above 20 for this keyword. If the coefficient file is absent, then Mendel will substitute theoretical kinship coefficients for conditional kinship coefficients on large pedigrees. Although this action blurs the distinction between additive polygenic variation and additive QTL variation, it does tend to preserve the overall additive

genetic variance. Finally, the current study sample contains ascertained pedigrees. Identification of probands is handled as in the previous example. See also the discussion of probands in Section 0.5.8 and in Option 14.

The fragment

POINT NUMBER	MARKER NAME	CHR NAME	NEUTER MAP (cM)	LOCATION SCORE (LOG-10)	AIC VALUE	NUMBER OF FACTORS
1	Marker01	AUTO	0.0000	1.6936	24.2006	1
2	Marker02	AUTO	0.1000	1.6893	24.2206	1
3	--	-	0.5000	1.7958	23.7300	1
4	--	-	1.0000	1.9892	22.8393	1
5	--	-	1.5000	2.2333	21.7151	1
6	--	-	2.0000	2.5074	20.4528	1
7	Marker03	AUTO	2.2800	2.6633	19.7351	1
8	Marker04	AUTO	2.3800	2.6672	19.7170	1
9	--	-	2.5000	2.6475	19.8078	1
10	--	-	3.0000	2.5391	20.3070	1
11	--	-	3.5000	2.3367	21.2389	1
12	Marker05	AUTO	3.8000	2.1927	21.9024	1
1	Marker01	AUTO	0.0000	1.6962	26.1887	2
2	Marker02	AUTO	0.1000	1.6898	26.2184	2
3	--	-	0.5000	1.7958	25.7300	2
4	--	-	1.0000	1.9892	24.8393	2
5	--	-	1.5000	2.2333	23.7151	2
6	--	-	2.0000	2.5074	22.4528	2
7	Marker03	AUTO	2.2800	2.6633	21.7352	2
8	Marker04	AUTO	2.3800	2.6672	21.7170	2
9	--	-	2.5000	2.6475	21.8078	2
10	--	-	3.0000	2.5391	22.3070	2
11	--	-	3.5000	2.3367	23.2390	2
12	Marker05	AUTO	3.8000	2.1927	23.9024	2

```
THE BEST AIC VALUE OF  19.717
OCCURS AT POINT    8 FOR  1 FACTOR(S).
```

of the summary file Summary19b.out provides location scores and the Akaike information criterion (AIC) for our bivariate trait. For the sake of brevity, we report here only results between the first and fifth markers. The location score curve plots the log base 10 of the ratio of the maximum likelihood of the data with and without the QTL effect included. For a single trait, a location score above 2.0 is suggestive and above 3.0 is impressive evidence

in favor of a QTL near the given map position. By definition, a location score never dips below 0.0. The AIC criterion for a given model is $-2L + 2n$, where L is the maximum standardized loglikelihood (base e) and n is the number of parameters minus the number of constraints under the model. All map distances are given in Morgans. It is obvious from the quoted AIC numbers that these data favor a one-factor model.

The standard output file Mendel19b.out contains a complete record of the maximum likelihood estimation process over all map points. Examination of this file shows that Mendel sweeps first from left to right and then from right to left along the genetic map at each factor level. The previous point visited supplies good starting values for the maximum likelihood search. In the left to right sweep, Mendel also uses starting values for the current map point from one factor level below. If you would like the pedigree deviances at the best AIC point reported at the bottom of the standard output file, then set DEVIANCES equal to true in the control file.

Our last example deals with X-linked QTL mapping. The trait of interest is head circumference in autistic children, who tend to have larger heads on average. The AGRE database furnishes the raw material for this exercise [36]. As demonstrated in the lines

```
ANALYSIS_OPTION = Variance_Components
QUANTITATIVE_TRAIT = HeadCirc
TRANSFORM = Degender :: HeadCirc
PREDICTOR = SEX :: HeadCirc
PREDICTOR = AGE :: HeadCirc
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
COVARIANCE_CLASS = X-Qtl
```

from Control19c.in, the model we adopt includes age and sex as trait predictors and additive polygenic, random environmental, and X-linked QTL variance components. Since trait values are available only on the children in these nuclear families, more complicated variance models are impractical. In this case, the female and male trait variances are nearly equal, so some of the qualms expressed earlier are not pertinent. However, we still follow the normal procedure for X-QTL mapping and standardize the trait by sex using the "Degender" transformation.

The output

```
                                                  LOCATION
 POINT    MARKER            CHR          NEUTER     SCORE
 NUMBER    NAME             NAME       MAP (cM)   (LOG-10)

    1     DXS9907           X             0.0000     0.0000
    2     DXS6807           X             5.3904     0.0000
```

```
    3    DXS9902          X          23.0402      0.0952
    4    DXS6810          X          64.5894      0.0000
    5    DXS7132          X          84.2994      0.0000
    6    DXS6800          X          94.1695      0.0000
    7    DXS6789          X         104.5601      0.5320
    8    DXS6799          X         108.4204      0.6749
    9    DXS6797          X         113.8899      0.7438
   10    DXS1047          X         144.2402      0.4641

 THE BEST LOCATION SCORE OF 0.74382     OCCURS AT POINT    9.
```

from Summary19c.out does not suggest linkage to any of the 10 X chromosome markers. Replacing the multivariate normal distribution by the multivariate t distributions in the analysis does not have much impact on this conclusion. Substituting X-linked polygenic background for autosomal polygenic background does increase the maximum location score, but only marginally.

19.5 Germane Keywords

```
ANALYSIS_OPTION = Variance_Components
COEFFICIENT_FILE
COVARIANCE_CLASS
COVARIANCE_FACTORS
DEVIANCES
GRID_INCREMENT
GROUP_FACTOR
MAX_ADJUSTED_MEIOSES
MULTIVARIATE_NORMAL
OUTLIERS
PEDIGREE_CUT_POINT
PERSON_CUT_POINT
PREDICTOR
PROBAND
PROBAND_FACTOR
QUANTITATIVE_TRAIT
TRANSFORM
```

Random Quote

I heard a Fly buzz — when I died.

Emily Dickinson in a poem with this as first line

The richer your friends, the more they will cost you.

Elizabeth Marbury

20 Analysis Option 20: QTL Association

20.1 Background

Quantitative traits are inherently more informative than disease-health dichotomies. One of the best ways of approaching association testing is through variance component models, treating genotypes at the marker locus as predictors modifying the mean of a quantitative trait [68]. This "measured genotype" approach controls for random environment and polygenic background while remaining in the frequentist domain of maximum likelihood estimation and likelihood ratio tests [13, 35, 47]. The current option implements this strategy for (a) multivariate traits, (b) both random and ascertained pedigrees, and (c) non-codominant markers.

Ascertainment is handled by conditioning on proband values. Non-codominant markers are dealt with by imputing marker allele counts to each person conditional on the marker phenotypes observed throughout his or her pedigree. These counts may be fractional. Mendel internally calculates the exact expected value for these counts at each marker, ignoring phenotypes at other markers and trait values. If you want better imputations, Mendel can use the expected values for these counts estimated by SimWalk. SimWalk's imputations are conditioned on all marker data from the pedigree and use an error model to account for possible genotype mistypings. Unfortunately, SimWalk cannot handle super-locus haplotype markers with unresolved phases, so you will have to revert to Mendel's internal method in this case. Once Mendel obtains allele counts, it then computes the maximum likelihood estimates of the regression coefficients for these predictors. It simultaneously estimates the variance and covariance components capturing the polygenic and environmental background of the traits. Mendel assesses association by conducting a likelihood ratio test to determine whether allelic regression coefficients differ significantly from 0.

A major objection to the measured genotype strategy is that the asymptotic distribution of the likelihood ratio statistic is sensitive to small sample sizes and departures from normality. This objection is irrelevant to permutation tests. These tests depend on exchangeable groups of people such as full siblings, spouse pairs, or a random sample from an ethnically uniform population. Each such group we call a permutation unit. Under the null hypothesis of no trait-genotype association, every permutation of trait values within a unit should be equally likely. Because permutation tests are computationally intensive, test statistics should be kept as simple as possible.

Model 2 of Option 20 relies on the minimum of the sum of squares

$$T(\beta,\mu) \;=\; \frac{1}{2}\sum_i\sum_j (x_{ij}-\mu_i-\sum_k a_{ijk}\beta_k)^2,$$

where i denotes a permutation unit, j a person within permutation unit i, x_{ij} his or her trait value, and a_{ijk} his or her imputed number of marker alleles of type k. The parameters of the model are the permutation unit effects μ_i and the allelic effects β_k. In order for all parameters to be identifiable, we assume $\sum_k \beta_k = 0$. Mendel evaluates the statistic $T_{\min} = \min_{\beta,\mu} T(\beta, \mu)$ rapidly enough to make permutation testing feasible.

20.2 Appropriate Problems and Data Sets

Model 1 of Option 20 tests for association between the alleles of a candidate gene and one or more correlated quantitative traits. The data may consist of a random sample of people, nuclear families, extended pedigrees, or some mixture of these three. The more complex the sampling units, the more complicated the variance component model can be for trait variation. Thus, if the data consists of a random sample of unrelated people, it would be foolish to estimate polygenic background. Only purely random variation from person to person can be estimated. Nuclear family data can sustain both random environment and polygenic background. In nuclear families, dominance effects and shared environment tend to be confounded, but even these two can be teased apart in extended pedigrees. Thus, it is important to match the complexity of the model to the complexity of the study sample.

Model 1 analyzes each marker common to the definition and map files separately. Imputation of allele counts is done independently of trait values. Outlier pedigrees and people are flagged if desired, and pedigree deviances are computed to help identify which pedigrees are driving a significant association. These additional features promote a more nuanced understanding of genetic heterogeneity. Determining what group of traits to analyze is the user's responsibility. Choose your traits wisely. Computation times increase with the cube of the number of traits. Transform all quantitative traits to normality if possible. Analysis Options 19 and 20 have the capacity to estimate parameters robustly by substituting the multivariate t distribution for the multivariate normal distribution.

We make no pretense of fully explaining the polygenic model for multivariate traits here. This is a subtle subject that takes a long time to master [59]. Roughly speaking, however, the polygenic model assumes that a large number of independent loci contribute small increments to a trait. Dominance effects capture departures from additivity among the alleles at each contributing locus. Random and household environments account for nongenetic increments to a trait independently mediated through individuals and households, respectively. On top of these random effects, are mean effects that directly influence a trait. For instance, sex, cigarettes smoked per day, and contraceptive use may all perturb a trait such as a lipid level. Mendel allows you to incorporate such mean effects in addition to the mean allelic effects of a candidate locus. Keep in mind, however, that a trait value for a person is discarded if one or more of the relevant predictors are missing. Mendel will use

a quantitative trait of a person even if he or she is missing other quantitative traits.

Model 2 operates somewhat differently. Only a single trait can be handled at a time, and ascertainment is ignored. Assigning individuals to permutation units should be done with care. To ensure exchangeability, you may also want to regress trait values on predictors such as sex and age and then analyze the resulting residuals. There are some identifiability issues in estimating parameters. For instance, if a different homozygous genotype is assigned to each separate permutation unit, then the β_k and μ_i parameters will be confounded. Taking permutation units too small reduces the power of the permutation test. For example, a permutation unit consisting of a single person contributes nothing to the test statistic. Because of the possible loss of power, we recommend model 1 as the primary vehicle of inference in most circumstances. Model 2 can be used for confirmation or when departures from normality cannot be corrected by data transformation. For example with the trait coronary calcification, no transformation can achieve normality since a large fraction of people show zero coronary calcification.

20.3 Input Files

Option 20 employs standard definition, map, and pedigree files. It outputs bottom-line information to a summary file. As with Option 19, all quantitative traits must be named. The commands

```
QUANTITATIVE_TRAIT = Gc_conc
PREDICTOR = Sex :: Gc_conc
PREDICTOR = Age :: Gc_conc
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Dominance
COVARIANCE_CLASS = Environmental
!COVARIANCE_CLASS = Household
```

from Control20a.in set up a model 1 problem involving the trait Gc concentration. Average Gc concentration is determined by a grand mean (always present) and regression coefficients on sex and age in addition to regression coefficients on allele counts. Note that it is possible to construct more complicated mean models using the syntax described in Analysis Option 14. The sole exception to this rule is that the major gene models discussed there are not allowed.

In this model, trait variation depends on polygenic additive and dominance effects and random environment. Table 19.1 lists the variance components Mendel recognizes. To analyze a bivariate trait, we must define the keyword `QUANTITATIVE_TRAIT` twice in the control file, as seen in Option 19. We could add household environment by deleting the exclamation point preceding the command `COVARIANCE_CLASS = Household` in the above

partial control file and by defining the keyword GROUP_FACTOR. The value of GROUP_FACTOR tells Mendel what field holds household information. Two people with the same label in this field belong to the same household. Households cannot extend beyond pedigree boundaries. If you want to override this limitation, then you can list multiple pedigrees as a single pedigree in the pedigree file. Note that this can adversely affect computation times. Household contributions are random effects rather then mean effects. When you model group contributions as mean effects, you should use the categories of a factor to label the groups, and name the factor as a predictor in the control file.

Redefining the keywords DEVIANCES and OUTLIERS to true allows deviances to be output and outlier pedigrees and people to be flagged. Deviances are explained in Section 0.9.4. If the keyword MULTIVARIATE_NORMAL is set to false, then the multivariate t is substituted for the multivariate normal in maximum likelihood estimation. Probands are designated through a proband field. This field and the symbol for a proband are defined in the control file as explained in Option 14.

To request that Mendel employ the expected allele counts estimated by SimWalk under its mistyping option, you must invoke the keyword ALLELE_COUNT_FILE to designate the file exported by SimWalk. For example, the command

```
ALLELE_COUNT_FILE = AEF-20.in
```

in the control file instructs Mendel to read the allele counts from the file AEF-20.in. That file names a second file generated by SimWalk, called AEKEY-20.TXT in this case, containing a key to some of the values in the allele count file. You must move both of these files to your working directory prior to execution of Mendel. You can change the name of the allele count file, but we recommend that you retain the name of the key file.

If a marker is associated with a trait, then you may wish to use the imputed allele counts for that marker in further statistical analysis. To capture the imputed counts for loci common to the definition and map files, name new definition and new pedigree files in the control file, using the keywords NEW_DEFINITION_FILE and NEW_PEDIGREE_FILE. The imputed counts will then be appended person by person in the new pedigree file after all other quantitative variables. The order of the appended variables follows the order of the loci in the map file, and for a particular locus the order of the alleles in the definition file. If you want to avoid creating a massive new pedigree file and are interested in the imputed allele counts for just a single locus, then list just that locus in the map file.

Control files for model 2 problems tend to be simpler. The last three commands

```
MODEL = 2
GROUP_FACTOR = Group
QUANTITATIVE_TRAIT = Gc_conc
```

in Control20b.in name the model and trait and identify the factor defining the various permutation units. Only those individuals assigned a non-blank value at this factor participate in testing. Although the keyword GROUP_FACTOR functions differently in models 1 and 2, no conflict arises because permutation units are ignored in one case and variance components in the other. If you use pedigree repeat numbers in model 2, then bear in mind the consequences this has for the size of some permutation units.

The accuracy of the p-value approximation depends on the number s of data permutations undertaken. The value of s is determined by the keyword SAMPLES. A command such as SAMPLES = 100000 in the control file overrides the default value of 10,000 for s. If none of the results from the data permutations are more extreme than the value from the observed data, then the p-value is reported as less than $1/s$. Each reported p-value $\hat{p}$ has an attached range of plus or minus twice $\sqrt{\hat{p}(1-\hat{p})/s}$, the approximate standard error of $\hat{p}$. If $\hat{p}$ is reported as 1.0 or less than $1/s$, then the range is omitted.

20.4 Examples

Our model 1 example considers data on plasma concentration of human group specific component. The Gc locus determines qualitative variation in this transport protein for vitamin D. A question of some interest is whether the genotypes at the Gc locus also determine quantitative differences in plasma concentrations. Data bearing on this question appear in an article by Daiger et al. [24]. The study sample consist of 31 monozygous twin pairs, 13 dizygous twin pairs, and 45 unrelated controls. Gc concentrations and Gc genotypes are available on all individuals. The two Gc alleles, 1 and 2, are codominant.

The output

```
                    SUMMARY FOR MEAN PARAMETERS

TRAIT      PARAMETER      PREDICTOR         ESTIMATE         STD ERR

Gc_conc        1          GRAND              29.5171          0.7052
Gc_conc        2          FEMALE              0.4305          0.3632
Gc_conc        3          MALE               -0.4305          0.3632
Gc_conc        4          Age                -0.0309          0.0394
Gc_conc        5          1                   1.3322          0.2494
Gc_conc        6          2                  -1.3322          0.2494

                  SUMMARY FOR VARIANCE COMPONENTS

TRAIT      PARAMETER      VARIANCE          ESTIMATE         STD ERR

Gc_conc        7          ADDITIVE            3.5989         17.2604
```

```
 Gc_conc       8          DOMINANCE          6.3836        17.0598
 Gc_conc       9          ENVIRONMENTAL      2.4770         0.6391

 TRAIT     TOTAL VARIANCE

 Gc_conc      12.4594
```

near the bottom of Mendel20a.out shows a total of six mean parameters and three variance parameters. Note that the female and male regression coefficients are constrained to sum to zero. Likewise, the allelic effects are constrained to sum to zero. The total variance for the trait is reported to facilitate the computation of heritability.

The subsequent output for the Gc locus,

```
THE LIKELIHOOD RATIO TEST STATISTIC IS  0.2485E+02 AT LOCUS Gc.
THIS CHI-SQUARE STATISTIC HAS   1 DEGREE OF FREEDOM.
THIS HAS APPROXIMATE P-VALUE 0.619E-06.

VARIABLES IN NEW PEDIGREE FILE CORRESPONDING TO IMPUTED ALLELE COUNTS

 LOCUS     STARTING   ENDING
 NAME      VARIABLE  VARIABLE
 Gc            3         4
```

shows the likelihood ratio test statistic and its p-value. Finally, if a new pedigree file is requested, Mendel lists the numbers of the quantitative variables giving the expected allele counts.

Setting the keyword STANDARD_ERRORS equal to true in the control file provides us with the asymptotic standard errors of the parameter estimates. If we set the keyword OUTLIERS equal to true in the control file, then the standard output file shows several significant empiric distribution statistics. These statistics monitor the number of outlier pedigrees and suggest a departure from normality in the raw data. If we reanalyze the data under the t distribution, then all outliers either disappear or moderate.

The output in Summary20a.out

```
LOCUS     TRAIT      MOST NEGATIVE EFFECT    MOST POSITIVE EFFECT    P-VALUE
NAME      NAME       ALLELE      ESTIMATE    ALLELE      ESTIMATE

Gc        Gc_conc    2            -1.3322    1             1.3322   0.619E-06
```

tells us that allele 2 tends to lower plasma Gc concentrations, allele 1 tends to raise plasma Gc concentrations, and these effects are highly significant. If we reanalyze the data with a multivariate t model, then the estimates and p-values change very little.

Our second example takes the same data and computes a p-value for the permutation statistic T_{min}. Each pair of twins constitutes a separate permutation unit; the unrelateds are lumped in a single permutation unit. Somewhat at our peril, we have neglected sex and age effects. The estimates of the allelic effects shown in Summary20b.out are now -1.23 and 1.23 versus the earlier estimates of -1.33 and 1.33. The p-value 0.0039 ± 0.0013 of T_{min} is much less impressive than the p-value of the likelihood ratio statistic. Part of the reason for this change is that the identical twins contribute almost no information. Two identical twins share the same genotype, and the permutation unit effect μ_i assigned to them explains all systematic variation.

20.5 Germane Keywords

```
ANALYSIS_OPTION = QTL_Association
ALLELE_COUNT_FILE
COVARIANCE_CLASS
DEVIANCES
GROUP_FACTOR
MODEL
MULTIVARIATE_NORMAL
NEW_DEFINITION_FILE
NEW_PEDIGREE_FILE
OUTLIERS
PREDICTOR
QUANTITATIVE_TRAIT
SAMPLES
STANDARD_ERRORS
```

Random Quotes

> Truth often suffers more by the heat of its defenders than from the arguments of its opposers.
>
> *William Penn* in *Some Fruits of Solitude*

> When people are least sure, they are most dogmatic.
>
> *John Kenneth Galbraith*

> I can picture in my mind a world without war, a world without hate. And I can picture us attacking that world, because they'd never expect it. *Jack Handey*

> Before you criticize someone, you should walk a mile in their shoes. That way . . . you're a mile away — and you have their shoes. *Jack Handey*

21 Analysis Option 21: Trim Pedigrees

21.1 Background

The pedigrees encountered in genetic epidemiology vary enormously in size and complexity. Since computational times rarely scale linearly in pedigree size, this variety has adverse consequences for statistical analysis. Option 21 deals with the problems of trimming irrelevant members from a pedigree. These individuals are typically dead or otherwise unavailable for study. Left in the pedigree, they slow likelihood evaluation. Of course, each pedigree will have some core people who must be retained. To maintain the proper relationships among the core people, we often must retain some of the relatives connecting them. Option 21 finds this skeleton of core people and connecting relatives and deposits it in a new pedigree file for further analysis [67]. When a moderate sized pedigree is depicted graphically, it is usually easy to spot the skeleton by visual inspection. Humans, after all, are good at this kind of pattern recognition. However, it is tedious to trim pedigrees manually, and it is better to rely on good software. Computers never tire or complain.

Option 21 has the secondary goal of splitting a disconnected pedigree into connected subpedigrees. This tactic also accelerates likelihood evaluation under the Lander-Green-Kruglyak algorithm. Pedigree trimming is always accompanied by pedigree splitting.

21.2 Appropriate Problems and Data Sets

Any pedigree analysis problem suffering from long computation times will benefit from trimming pedigrees and splitting them into their connected components. Option 21 can handle arbitrarily large pedigrees. Be careful to choose a sensible trimming criterion. Option 21 can also operate on the binary SNP data files discussed in section 0.6.

21.3 Input Files

Option 21 employs standard definition, map, and pedigree files. It can also operate on binary SNP data. In either case, Option 21 outputs the trimmed pedigrees to a new pedigree file. For binary SNP data, the trimmed SNP genotypes are output to the new SNP data and phase files that must be named in the control file. The typical control file using standard data files

```
DEFINITION_FILE = Def21a.in
MAP_FILE = Map21a.in
PEDIGREE_FILE = Ped21a.in

OUTPUT_FILE = Mendel21a.out
```

```
SUMMARY_FILE = Summary21a.out
NEW_PEDIGREE_FILE = Ped21a.out

ANALYSIS_OPTION = Trim_pedigrees
AFFECTED_LOCUS_OR_FACTOR = Retain
AFFECTED = in
```

for Option 21 names the new pedigree file and an affected locus or factor for flagging the core people in each pedigree. An analogous example control file for pedigree trimming with binary SNP data would be

```
DEFINITION_FILE = Def21b.in
PEDIGREE_FILE = Ped21b.in
SNP_DEFINITION_FILE = SNP_def21b.in
SNP_DATA_FILE = SNP_data21b.bin
SNP_SUBSET_FILE = SNP_subset21b.in

OUTPUT_FILE = Mendel21b.out
SUMMARY_FILE = Summary21b.out
NEW_DEFINITION_FILE = Def21b.out
NEW_PEDIGREE_FILE = Ped21b.out
NEW_SNP_DEFINITION_FILE = SNP_def21b.out
NEW_SNP_DATA_FILE = SNP_data21b_out.bin
NEW_SNP_PHASE_FILE = SNP_phase21b_out.bin

ANALYSIS_OPTION = Trim_pedigrees
AFFECTED_LOCUS_OR_FACTOR = Retain
AFFECTED = in
```

Note the use of a SNP subset file, as described in section 0.6.3. The resulting new SNP definition and data files will be restricted to the subset of SNPs designated in this file.

There are ten models for Option 21. In the above examples, the default model 1 is operative, and a core person must have the phenotype “in” appearing in the field “Retain.” Model 2 identifies a core person as anyone whose has a non-blank phenotype in the designated field. Model 3 includes all people but splits the pedigree into separate components. For example, if the permitted values for “Retain” are “in”, “possible”, and blank, then model 1 considers only those people with value “in” as core, model 2 considers anyone with either value “in” or “possible” as core, and model 3 considers everyone as core. The core people in model 4 include everyone who is affected, a proband, has a recorded value for a designated quantitative trait, or has a non-blank phenotype or non-trivial penetrance for at least one model locus. The two keywords `AFFECTED_LOCUS_OR_FACTOR` and `AFFECTED` need only be defined for models 1 and 2. Mendel’s pedigree preprocessing step invokes model 4

Table 21.1: Pedigree Trimming Models

Model Number	Core Field Required	Indicator of Core Person	Trimming or Splitting
1	Yes	Unique Value at Core Field	Trimming*
2	Yes	Non-blank Value at Core Field	Trimming*
3	No	Everyone	Splitting
4	No	See Text	Trimming*
5	No	See Text	Splitting
6	No	Founder	Both
7	No	Parent	Both
8	No	Child	Both
9	No	Non-blank Value at Quantitative Trait	Both
10	No	Combine Everyone into One Pedigree	Neither

*If trimming causes disjoint components within a pedigree, the pedigree will also be split.

unless you set the keyword PRETRIM_PEDIGREES equal to false. See Section 0.5.6.4 for a description of the default trimming criteria. To enable very quick analyses, model 5 splits a pedigree into its constituent nuclear families. Models 6 through 9 extract respectively, all founders, all parents, all children, or all individuals with a non-blank value at the first quantitative trait, and then deposits them into the new pedigree file as unrelateds. Finally, model 10 combines everyone into one overall pedigree. Any parent-child relationships in the original pedigrees remain intact in the single, expanded pedigree. Table 21.1 summarizes the various models.

21.4 Example

Our Option 21 example considers a 20 member pedigree with two inbreeding loops, one pair of identical twins, and three core individuals. The control file, Control21a.in, for this example appears above. The new pedigree file, Ped21a.out,

```
A:Test  ,1      ,       ,       ,M      ,       ,1\2    ,       ,
A:Test  ,5      ,1      ,2      ,M      ,       ,1\1    ,1\1    ,
A:Test  ,7      ,1      ,2      ,M      ,       ,1\1    ,1\1    ,
A:Test  ,2      ,       ,       ,F      ,       ,1\1    ,       ,
A:Test  ,13     ,5      ,6      ,F      ,       ,1\1    ,1\1    ,in
A:Test  ,6      ,       ,       ,F      ,       ,1\1    ,1\1    ,
A:Test  ,14     ,7      ,8      ,M      ,       ,1\1    ,1\1    ,in
A:Test  ,8      ,       ,       ,F      ,       ,1\1    ,1\2    ,
```

```
B:Test  ,16     ,       ,       ,M      ,T      ,       ,       ,
B:Test  ,17     ,       ,       ,M      ,T      ,       ,       ,in
```

shows a skeleton of 10 individuals scattered over two subpedigrees, A:Test and B:Test, corresponding to the original single pedigree Test. These two subpedigrees imply the correct relationships among the core people. Note that both twins are retained even though only one of them is designated as a core person in Ped21a.in. Also note that Mendel lists pedigree members in Ped21a.out in a different order than in Ped21a.in. This is a harmless change since person and parent names are respected.

21.5 Germane Keywords

```
ANALYSIS_OPTION = Trim_Pedigrees
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
MODEL
NEW_DEFINITION_FILE
NEW_PEDIGREE_FILE
NEW_SNP_DATA_FILE
NEW_SNP_DEFINITION_FILE
NEW_SNP_PHASE_FILE
SAMPLE_SUBSET_FILE
SNP_DATA_FILE
SNP_DEFINITION_FILE
SNP_PHASE_FILE
SNP_SUBSET_FILE
```

Random Quotes

> In the end we will conserve only what we love. We love only what we understand. We will understand only what we are taught.
>
> *Baba Dioum*

> Expectations tend to increase in direct proportion to the amount of money being spent, and if you are spending a fortune you expect perfection. Details that we would consider trivial assume enormous significance (to the rich); the breakfast egg is inedible because it is marginally underboiled, the silk shirt is unwearable because of a barely visible wrinkle, the chauffeur is insupportable because he's been eating garlic again, the doorman is either insufficiently attentive or overfamiliar — the list of maddening blots on the landscape of life just goes on and on. How can you have a nice day if some fool hasn't warmed your socks or ironed your newspaper properly?
>
> *Peter Mayle* in *Acquired Tastes*

22 Analysis Option 22: Association Given Linkage

22.1 Background

Although the forces of linkage and association are difficult to disentangle, doing so can yield useful insights. Analysis Option 22 implements a model that captures the two simultaneously in pedigree data [16, 118]. In particular, it estimates both recombination and linkage disequilibrium parameters and conducts a likelihood ratio test for association given linkage. To avoid some of the complexities of multipoint analysis, the option proceeds marker by marker. Recombination is parameterized by the recombination fraction θ separating the trait locus and a marker, and linkage disequilibrium is parameterized by the conditional frequencies of the disease allele at the trait locus given each marker allele.

More specifically, let d and n denote the disease and normal alleles at the disease locus, number the marker alleles from 1 through m, and fix the marginal population frequency p_i of marker allele i at its observed frequency. Option 22 estimates θ and the conditional frequency $q(d|i) = 1 - q(n|i)$ of the disease-marker haplotype d–i given marker allele i for each i. During the estimation process, the linear constraint $\sum_{i=1}^{m} p_i q(d|i) = q$ is imposed, where q is the population frequency of the disease allele. You should estimate q beforehand from disease prevalence data or from ascertained pedigrees via segregation analysis and insert the estimated q in the definition file. The joint null hypothesis of no linkage and no association amounts to $\theta = \frac{1}{2}$ and $q(d|i) = q$ for all i. The null hypothesis of no association given linkage is just $q(d|i) = q$ for all i with no restriction on θ beyond $\theta \in [0, \frac{1}{2}]$. Mendel tests the latter hypothesis by comparing the maximum LOD score assuming linkage equilibrium with the maximum LOD score taking linkage disequilibrium into account.

22.2 Appropriate Problems and Data Sets

Option 22 uses the same kind of definition, map, pedigree, and penetrance files as Option 2. Option 22 is one option where it is imperative to input good marker allele frequencies. We urge users to estimate these via model 1 or 3 of Option 6. If your data set is large, then you can estimate allele frequencies from it. If you estimate allele frequencies from another data set, say a large random sample, then be certain that the ethnicity of this data set matches the ethnicity of your own data set, which should be of a single ethnicity to prevent population stratification. If you use Option 18 to combine neighboring loci into a super-locus for single point analysis, then remember to estimate the frequencies of the super-locus haplotypes using Option 6. Under no circumstances, should you rely on linkage equilibrium among the markers unless you test this assumption beforehand. Finally, to avoid failures in the large sample assumptions behind the likelihood ratio test of asso-

ciation given linkage, we urge lumping of infrequent marker alleles, say with five or fewer representatives in your sample, prior to analysis.

If you suspect reduced penetrance or phenocopies at the trait locus, then you should also include a penetrance file. The parameters in the penetrance file should be consistent with the population prevalence r of the disease. Thus, if someone with trait genotype i/j has probability $f_{i/j}$ of displaying the disease, then the equation

$$r = q^2 f_{d/d} + 2q(1-q) f_{d/n} + (1-q)^2 f_{n/n}$$

should hold connecting the disease allele frequency q to r and the $f_{i/j}$. These considerations particularly apply to traits constructed by dichotomizing a quantitative variable since all three genotypes are apt to produce individuals with values above the trait threshold. In view of the loss of information, it probably is better not to dichotomize at all. Option 14 can help you estimate penetrance parameters from your data.

22.3 Input Files

In Option 22, the trait locus should have exactly two alleles, the second of which is considered the disease allele. If disease genotypes have reduced penetrance or phenocopies, then you should spell this out using either PENETRANCE keywords or a penetrance file. The control file

```
DEFINITION_FILE = Def22.in
MAP_FILE = Map22.in
PEDIGREE_FILE = Ped22.in
OUTPUT_FILE = Mendel22.out
SUMMARY_FILE = Summary22.out
ANALYSIS_OPTION = ASSOCIATION_GIVEN_LINKAGE
COMPLEXITY_THRESHOLD = 0.120000E+09
!DEVIANCES = True
AFFECTED_LOCUS_OR_FACTOR = TRAIT
AFFECTED = 2
PENETRANCE = 0.05 :: d|d
PENETRANCE = 0.99 :: d\D
PENETRANCE = 0.99 :: D\D
```

for the Option 22 example is unremarkable except for the command increasing the complexity threshold and the penetrance construction. Because all computations are done by the Elston-Stewart algorithm, increasing the maximum adjusted meioses has no impact on Mendel's ability to process large pedigrees in Option 22. Ordinarily, the default complexity threshold is adequate. Construction of a simple penetrance function using the PENETRANCE keyword in the control file is discussed in Section 0.5.7.2. The default value SEARCH of

the keyword TRAVEL cannot be overridden. Finally, Option 22 makes no provision for a fraction of unlinked pedigrees. It is possible to print pedigree deviances to assess which pedigrees favor the alternative of association given linkage. Simply insert the command DEVIANCES = True in your control file. Deviances are explained in Section 0.9.4.

22.4 Example

Our Option 22 example deals with a dominant trait with reduced penetrance. The summary file Summary22.out

```
MAXIMUM LOD SCORES AND TEST OF ASSOCIATION GIVEN LINKAGE

              MAX LOD FOR    MAX LOD FOR     P-VALUE FOR
 MARKER         LINKAGE      LINKAGE AND     ASSOCIATION      DEGREES OF
  NAME           ALONE       ASSOCIATION   GIVEN LINKAGE      FREEDOM

SNP1             1.3618         1.9702        0.094176           1
SNP2             1.6887         2.7063        0.030409           1
SNP3             1.5395         2.3857        0.048373           1
S1+S2            1.7761         2.9707        0.063879           2
S2+S3            0.9798         1.7566        0.058569           1
S1+S3            1.6203         2.6652        0.090175           2
S1+S2+S3         0.8599         2.3915        0.029399           2
```

reports the results for three SNPs and all of their two and three-way combinations derived from Option 18. Marker allele and haplotype frequencies were estimated via model 3 of Option 6, and then rare marker haplotypes combined via Option 16. The super-locus formed from the first two SNPs gives the greatest LOD score for linkage and association jointly. However, the combination of all three SNPs yields the most impressive p-value for association given linkage. In each case, the degrees-of-freedom is one less than the number of lumped marker alleles.

In concluding our discussion of this option, it is worth reminding readers of the pitfalls in inference caused by incorrect specification of allele or haplotype frequencies. Other sorts of model miss-specification such as penetrance mistakes are also risky. Note that estimated recombination fractions should be low because high values are inconsistent with linkage disequilibrium. If recombination fractions or conditional haplotype frequencies fall on a boundary, then the nominal degrees-of-freedom stated in the summary file may be wrong. Thus, it is a good idea to check parameter estimates in the standard output file before leaping to conclusions.

22.5 Germane Keywords

```
ANALYSIS_OPTION = Association_given_Linkage
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
COMPLEXITY_THRESHOLD
DEVIANCES
PENETRANCE
```

Random Quotes

Their bodies were so close together that there was no room for real affection.

Stanislaw Lec

What a country calls its vital economic interests are not the things which enable its citizens to live, but the things which enable it to make war. Gasoline is much more likely than wheat to be a cause of international conflict.

Simone Weil in *The Need for Roots*

Men often oppose a thing merely because they have no agency in planning it or because it may have been planned by those whom they dislike.

Alexander Hamilton

We sleep in separate rooms; we have dinner apart. We take separate vacations. We're doing everything we can to keep our marriage together.

Rodney Dangerfield

Sometimes I think the surest sign that intelligent life exists elsewhere in the universe is that none of it has tried to contact us.

Bill Waterson

The best thing is to look natural, but it takes makeup to look natural.

Calvin Klein

. . . Every word is at home,
Taking its place to support the others,
The word neither diffident nor ostentatious,
An easy commerce of the old and the new,
The common word exact without vulgarity,
The formal word precise and not pedantic,
The complete consort dancing together.

T.S. Elliot in *Little Gidding*

23 Analysis Option 23: SNP Imputation

23.1 Background

It is possible to impute genotypes and haplotypes in two very different ways. If pedigree data are available, then the genotypes and phases of individuals are constrained by the genotypes of their relatives. Sometimes such constraints fully determined missing genotypes or phases; more often, the constraints render some configurations more probable and other configurations less probable. In the absence of pedigree data, one can use linkage disequilibrium (LD) patterns to make similar inferences. With the advent of dense SNP genotyping, the LD method has become increasingly attractive.

This option of Mendel performs both methods of imputation on pedigrees genotyped on a dense set of SNP markers. Of course, the singleton pedigrees present in a sample undergo only LD inference. Genotype imputation is a crucial prelude to SNP association analysis as implemented in Analysis Option 24. Haplotyping reveals some of the fine-scale structure of the human genome. In particular, conserved haplotypes are signatures of disease predisposing mutations.

In LD inference Mendel estimates haplotype frequencies along a sliding genomic window and assigns most probable SNP genotypes and haplotypes based on these estimates. In most small genomic regions, only a handful of haplotypes prevail in any population. Mendel accordingly adds a penalty term to the loglikelihood of the sample in haplotype frequency estimation [6]. The penalty enforces parsimony and automatically discards potential haplotypes with low explanatory power. Estimation is carried out by a very fast MM algorithm closely related to the classical EM algorithm for haplotype frequency estimation [76].

In LD inference Mendel imputes missing genotypes by one of two methods. The highest posterior probability method, Mendel's default, assigns each genotype at the current SNP a posterior probability and selects the best genotype on this basis. In haplotyping, the best ordered genotype is chosen. The best haplotype pair method selects the most probable haplotype pair consistent with surrounding genotypes and imputes the genotype displayed by the haplotype pair at the current SNP. The highest posterior probability method tends to be slightly more accurate.

As with Mendel's other SNP based analyses, SNPs and individuals with low genotyping rates are ignored during SNP imputation. Instructions for controlling this filtering appears in Section 0.6.1.1. To force Mendel to impute genotypes at all SNPs and individuals, set the keywords `MIN_SUCCESS_RATE_PER_INDIVIDUAL` and `MIN_SUCCESS_RATE_PER_SNP` to 0.0. In this case, if a SNP is completely untyped in a sample, then all genotypes at the SNP are assigned the value 0/0.

23.2 Appropriate Problems and Data Sets

Analysis Option 23 requires the use of binary SNP data files, as described in Section 0.6. The data should present either pedigrees or unrelated individuals genotyped at a large number of autosomal or X-linked SNPs. Each data set must contain either all autosomal or all X-linked SNPs. Mixing them in one data set is not allowed in this option. The SNPs must display substantial LD. If this is not the case, then the most common genotype in the population will be imputed for all missing genotypes. Furthermore, haplotype frequency estimation and imputation will be less accurate. Analysis Option 23 relies crucially on the SNP order. Thus, all SNPs should be listed in the SNP definition file in their genomic order. SNPs without position information are omitted from this analysis. If there are additional SNPs with ambiguous ordering, omit these using the SNP Subset file described in Section 0.6.3.

Model 1, the default model, performs genotype imputation. The imputed genotypes are unordered, unless phases can be deduced from genotypes of available relatives. Model 2 performs haplotype imputation and outputs ordered genotypes. Suppose we label the two alleles at a SNP as 0 and 1. In the bit compressed SNP file input by this option, unordered genotypes are coded in two bits with $00 \to 0/0$, $01 \to 0/1$, $10 \to 1/1$, and $11 \to$ missing. Here the symbol / represents the unordered allele separator. Mendel converts the missing data to imputed data. In haplotyping, the ordered genotypes delivered in the output file are coded as $00 \to 0|0$, $01 \to 0|1$, $10 \to 1|0$, and $11 \to 1|1$. Here the symbol | represents the ordered allele separator, with maternal alleles to the left and paternal alleles to the right.

The output of Analysis Option 23 is stuffed in a bit compressed file largely unreadable by the user. If requested, the haplotyping model will deposit in the summary file the discovered haplotypes and their population frequencies around a central SNP. This central SNP is named using the keyword `ZOOM_SNP`. Appended to this output are the imputed haplotypes for the sample individuals in the window surrounding the designated SNP.

23.3 Input Files

This option requires a pedigree file and SNP definition and data files. If your SNP data is partially phased, then you will also need a SNP phase file. Standard definition and map files are unnecessary.

The penalty in haplotype frequency estimation depends on two tuning constants. If p is the parameter of haplotype frequencies and $L(p)$ is the loglikelihood of the data, then Mendel maximizes the criterion $L(p) - \lambda \sum_i f(p_i)$, where

$$f(q) = \begin{cases} q & q \leq \delta \\ \delta & q \geq \delta \end{cases}.$$

The commands

```
TUNING_CONSTANT = 0.005 :: DELTA
TUNING_CONSTANT = 10000 :: LAMBDA
```

in the control file specify new values for the tuning constants δ and λ. Mendel's default values are $\delta = .01$ and $\lambda = 1000$. Do not change these without good reason.

There are two more keywords pertinent to this option. The sliding window employed in haplotype frequency estimation extends a certain number of FLANKING_SNPS on either side of the central SNP. Mendel's default value of 7 flanking SNPs entails a window of length $7 + 1 + 7 = 15$ surrounding the central SNP. For a dense marker map, you might consider increasing the number of flanking SNPs. As this number increases, computing times tend to increase and imputation and phase errors at the central SNP tend to decrease. The keyword IMPUTATION_METHOD takes one of the two values Highest_Posterior_Probability (or simply HPP) and Best_Haplotype_Pair (or simply BHP). The former value is the default.

23.4 Example

The sample data for this option consist of 129 parent-offspring trios genotyped at 103 SNPs on chromosome 5q31 by Daly et al [27]. The control file for Mendel's sample problem

```
!
!  Input Files
!
PEDIGREE_FILE = Ped23a.in
SNP_DEFINITION_FILE = SNP_def23a.in
SNP_DATA_FILE = SNP_data23a.bin
!
!  Output Files
!
OUTPUT_FILE = Output23a.out
SUMMARY_FILE = Summary23a.out
NEW_DEFINITION_FILE = Def23a.out
NEW_PEDIGREE_FILE = Ped23a.out
NEW_SNP_DEFINITION_FILE = SNP_def23a.out
NEW_SNP_DATA_FILE = SNP_data23a_out.bin
NEW_SNP_PHASE_FILE = SNP_phase23a_out.bin
!
!  Analysis Parameters
!
ANALYSIS_OPTION = SNP_IMPUTATION
MODEL = 2
FLANKING_SNPS = 7                    ! default 7
ZOOM_SNP = IGR3029a_2                ! output haplotypes centered here
```

```
MIN_SUCCESS_RATE_PER_INDIVIDUAL = 0.0 ! default 0.98
MIN_SUCCESS_RATE_PER_SNP = 0.0        ! default 0.98
```

names several new output data files, including a phase file. Model 2, the haplotyping sub-option, is chosen with 7 flanking SNPs.

The top of Summary23a.out first displays the window centered at the SNP named using the keyword ZOOM_SNP, in this case IGR3029a_2. The list

```
SNP NUMBER     SNP NAME
     58        IGR3018a_2
     59        IGR3019a_2
     60        IGR3020a_1
     61        IGR3022a_1
     62        IGR3023a_1
     63        IGR3023a_3
     64        IGR3029a_1
     65        IGR3029a_2  <-
     66        IGR3030a_1
     67        IGR3039a_1
     68        IGR3044a_1
     69        IGR3045a_1
     70        IGR3051a_1
     71        IGR3053a_1
     72        IGR3061a_1
```

is followed by the truncated snippet

```
ESTIMATED HAPLOTYPE FREQUENCIES

HAPLOTYPE    FREQUENCY         HAPLOTYPE
  NUMBER
       1    0.65071039       110011110000101
       2    0.26237983       111100001100100
       3    0.05177407       000111101111011
       4    0.00403025       111000001100100
       5    0.00357183       111100001100110
```

from the middle of Summary23a.out showing the top 5 haplotypes by frequency out of the 33 identified haplotypes. The original allele symbols are replaced here by 0's and 1's. Mendel compares the two SNP names and converts them to 0 and 1 consistent with their lexicographic order in the standard ASCII collating sequence. Finally, the fragment

```
BEST HAPLOTYPES AROUND SNP IGR3029a_2
```

```
PEDIGREE     PERSON            HAPLOTYPES
  NAME        NAME

PED054       430             110011110000101
PED054       430             111100001100100

PED054       412             110011110000101
PED054       412             111100001100100

PED054       431             110011110000101
PED054       431             110011110000101
```

from Summary23a.out displays the best haplotype pairs for the first parent-offspring trio. The father is individual 430, the mother 431, and the daughter 412. Here the SNP alleles are listed from left to right along each of the two haplotypes. Maternal chromosomes appear on the top, and paternal on the bottom. As they should be, the daughter's genotypes are consistent with those of her parents. See Summary23a.out for the full output. The haplotypes over all SNPs and individuals are deposited in the new data output files.

23.5 Germane Keywords

```
ANALYSIS_OPTION = SNP_Imputation
FLANKING_SNPS
IMPUTATION_METHOD
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
MODEL
NEW_DEFINITION_FILE
NEW_PEDIGREE_FILE
NEW_SNP_DATA_FILE
NEW_SNP_DEFINITION_FILE
SAMPLE_SUBSET_FILE
SNP_DATA_FILE
SNP_DEFINITION_FILE
SNP_PHASE_FILE
SNP_SUBSET_FILE
TUNING_CONSTANT
ZOOM_SNP
```

Random Quote

Young men have a passion for regarding their elders as senile.
Henry Adams in *The Education of Henry Adams*

24 Analysis Option 24: GWAS (SNP Association)

24.1 Background

Association testing is one of the dominant themes of Mendel. Here we take up the subject from the perspective of large-scale SNP genotyping. We narrow our focus even more by concentrating on just two models. Model 1, the default model, involves ordinary linear regression for random samples of a population with a quantitative trait. Model 2 involves logistic linear regression in case-control samples. Even these simple settings are a little sinister because the potential predictors vastly outnumber the number of observations. Most investigators ignore the multivariate nature of the predictors and opt for simple linear regression one predictor at time. In addition to delivering these simpler, marginal results, Mendel can perform penalized regression [115, 116], a form of continuous model selection on all predictors simultaneously. The penalty discourages predictors with low explanatory power from entering the model. Our implementation also allows the user to force individual predictors and groups of predictors to be retained in the model. These might be known or candidate predictors of the trait. Finally, each SNP predictor can be assigned a relative weight for more nuanced model selection.

Rare SNPs that affect traits are inherently hard to find using association testing since so few individuals have the minor allele. A potential solution is to group SNPs into sets. These sets can be based on proximity to genes or participation in common molecular pathways. You make group assignments in the SNP definition file as instructed in Section 0.6.2. If a SNP is not assigned to a group, then the SNP is considered a singleton group. Mendel runs penalized regression analyses that act both on individual predictors and groups of predictors [120, 121]. You can even specify the proportion of the total weight given to individual predictors versus group predictors. Thus, one can ignore group information (the default) or focus entirely upon it.

To more formally present penalized regression methods, let y_i be the response for case i of n cases, x_{ij} the jth of p predictors for case i, β_j the regression coefficient corresponding to x_{ij}, and μ the intercept. For notational convenience also let $\boldsymbol{\theta} = (\mu, \beta_1, \ldots, \beta_p)^t$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t$. In ordinary linear regression, the objective function to be minimized is

$$g(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - \mu - \boldsymbol{x}_i^t \boldsymbol{\beta}\right)^2. \tag{4}$$

In its more nuanced analyses, Mendel also requires a weight s_j for the jth predictor and a weight r_k for the kth of g groups. If $\boldsymbol{\beta}_k$ denotes the subvector of regression coefficients for predictors assigned to group k, then penalized regression is implemented by minimizing

the modified objective function

$$f(\boldsymbol{\theta}) \;=\; g(\boldsymbol{\theta}) + \lambda_P \sum_{j=1}^{p} \frac{1}{s_j} N_P(\beta_j) + \lambda_G \sum_{k=1}^{g} \frac{1}{r_k} N_G(\boldsymbol{\beta}_k), \tag{5}$$

where N_P and N_G are the predictor and group penalty functions; the tuning constants λ_P and λ_G control the strength of the penalties. We restrict the penalty functions to be non-negative and the weights s_j and r_k to be positive. The penalty terms not only shrink each β_j toward the origin, they actually force most β_j to vanish. Several penalty functions have been investigated [5]. For the predictor penalty function, Mendel allows the user to choose between the classic lasso penalty [20] and the minimax concave penalty (MCP) [119]. The MCP is the default. (To force the use of the lasso penalty use the command `LASSO_PENALTY = True` in the control file.)

The lasso penalty, $N_P(\beta_j) = |\beta_j|$, makes moderate computational demands because it permits the use of cyclic coordinate descent algorithms [115]. In practice, the lasso may shrink the regression estimates too much. Severe shrinkage encourages false positives to enter a model to compensate. One remedy to this over shrinkage is the MCP. The MCP

$$N_P(\beta_j) \;=\; \int_0^{|\beta_j|} \left(1 - \frac{s}{\gamma \lambda_P}\right)_+ ds$$

includes an additional tuning constant γ. Moreover, the MCP can be majorized by a v-shaped function similar to the lasso's absolute value function [65]. Thus, with minor differences the coordinate descent algorithm carries over to the MCP, and there is almost no extra computational burden. Under MCP, model selection is achieved without severe shrinkage, and inference in GWAS improves [44].

For the group penalty function, Mendel always uses $N_G(\boldsymbol{\beta}_k) = \|\boldsymbol{\beta}_k\|_2$ where $\|z\|_2 = \sqrt{\sum_j z_j^2}$ is the Euclidean (ℓ_2) norm. This introduces little additional computational burden because it allows continued use of coordinate descent algorithms [65]. Note that the intercept μ is ignored in all penalty terms.

In logistic regression we maximize the penalized loglikelihood

$$\begin{aligned} f(\boldsymbol{\theta}) \;=\; & \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi\left(\mu + \boldsymbol{x}_i^t \boldsymbol{\beta}\right)\right] + (1 - y_i) \ln\left[1 - \pi\left(\mu + \boldsymbol{x}_i^t \boldsymbol{\beta}\right)\right] \right\} \\ & - \lambda_P \sum_{j=1}^{p} \frac{1}{s_j} N_P(\beta_j) - \lambda_G \sum_{k=1}^{g} \frac{1}{r_k} N_G(\boldsymbol{\beta}_k), \end{aligned} \tag{6}$$

where $\pi(z) = e^z/(1 + e^z)$.

As mentioned, by default Mendel ignores SNP groupings, and weights all predictors uniformly. However, if desired, the individual predictor weights s_j can be set for each SNP

in the SNP definition file as described in Section 0.6.2. We note that the larger the weight s_j, the more likely predictor j will remain in the model. If a weight is not assigned to a SNP, the weight is set to 1.0. However, you can change this default so that each unassigned SNP weight $s_j = 1/\sqrt{4q(1-q)}$, where q is the minor allele frequency at SNP j. All non-SNP predictors currently have individual weights 1.0. Also, all group weights, r_k, are currently fixed at 1.0. Keep in mind that you may force specified predictors or groups of predictors to be retained in all standard models. Of course all predictors that are members of a retained group are also set to be retained, even if Mendel is ignoring group weights.

In Option 24, Mendel tunes the constants λ_P and λ_G to select a predetermined number of predictors. Besides SNPs, the predictors can be environmental indices such as height, weight, or smoking behavior. All quantitative predictors should be standardized. This puts them on a common footing. Once the final model is selected, the intercept and regression coefficients β_j for the selected predictors are re-estimated with the predictor and group tuning constants reset to 0; all other predictors are omitted in the process. The summary file reports the re-estimated parameters.

The order with which predictors enter the final model is highly correlated with how significant the predictors are in a marginal analysis. Recall that a p-value can be attached to any regression coefficient β_j by conducting an F test or a likelihood ratio test of the null hypothesis $\beta_j = 0$. In simple linear regression, the alternative model has two parameters μ and β_j, while the null model has the single parameter μ. Mendel reports these marginal p-values. In its multivariate analyses Mendel also reports leave-one-out indices. These indices are the p-values derived from the final multivariate model by leaving one predictor out at a time. Thus, the null model in this setting retains all of the selected predictors except the current one. Because of the complicated selection procedure, the leave-one-out indices are not legitimate p-values. We include them in the output because they give insight into how correlated the final predictors are. A larger index value is a sign that other predictors can partially take up the slack caused by omitting the current predictor. SNPs are often correlated because of linkage disequilibrium.

In addition to ranking single predictors, Mendel can also find the most significant interactions among the predictors. Mendel takes either of two approaches to interaction analysis — pairwise or higher-order. In testing pairwise interactions, Mendel looks at all two-way interactions either within a base set of predictors or between each member of the base set and all other predictors. The base set can be either a set of named predictors or the top predictors found in Mendel's marginal or penalized regression analysis. As always, the predictors can be either SNPs or environmental covariates.

Mendel can also identify interactions via penalized regression. Here, two-way, three-way and even higher-order interactions are fair game. This application of penalized regression is distinct from its application in identifying good marginal predictors. In penalized

interaction analysis, Mendel examines only interactions within a pool of top marginal predictors. Interaction predictors are constructed by taking products of marginal predictors. Mendel is in danger of being overwhelmed by the sheer number of possible interactions. You can avoid these computational bottlenecks by setting reasonable limits on (a) the number of marginal predictors in the pool of top marginal predictors, (b) the number of interactions to be singled out, and (c) the maximum level of interactions allowed. Once the desired number of interactions is identified, Mendel reports the same kind of marginal p-values, leave-one-out indices, and parameter estimates for the interaction predictors that it did for the penalized regression analysis of the marginal predictors. It is helpful to keep in mind that the marginal predictors are viewed as potential one-way interactions and may turn up among the interaction predictors selected via penalized regression. We recommend setting the desired number of predictors in the penalized interaction round higher than the pool size.

24.2 Appropriate Problems and Data Sets

Option 24 requires the use of binary SNP data files as described in Section 0.6. The data should present unrelated individuals genotyped at a large number of autosomal or X-linked SNPs. This option performs up to three analyses: a marginal analysis on each predictor, a penalized regression analysis on the single predictors, and an interaction analysis. There are two models: logistic regression for case-control data, and linear regression for a random sample from a population with a quantitative trait. Any number or type of co-variates can be included in the marginal and penalized regression analyses, but these are not include in the interaction analyses. Any specified co-variates are include in the null and alternate models.

For ordinary, linear regression, any quantitative variable can be used as the response variable. If you have case-control data, then Mendel normally expects a quantitative response variable that has value 1 for cases and 0 for controls. Section 0.5.4.6 explains how to create this indicator variable easily from other factors or variables, using transformation commands within a control file such as `Indicator`, `Above_Threshold`, or `Below_Threshold`. To more easily accommodate data sets that use 2 for cases and 1 for controls, Mendel allows the command `CASE2_CONTROL1 = True`. The default value for this keyword is False, unless using Plink-format input files.

For case-control data, that is, the logistic regression procedure, Mendel carries out a score test to quickly obtain a good estimate of the p-value for each predictor. Only if the p-value is below a threshold does Mendel bother to carry out a full likelihood calculation and estimate effect size. This threshold can be modified from its default of 10^{-5} using the keyword `ESTIMATION_THRESHOLD`.

As mentioned earlier, standardize all quantitative predictors except for the case-control

Table 24.1: The genotype encodings for the possible dominance models. The additive model is also known as the codominant model and is the default. In the genotype column, "1" and "2" represents the first and second alleles for each SNP in the input data files. An effect size estimate reflects the change in trait values due to each positive unit change in the encodings. For example, when using the default additive model, the effect size estimates the trait difference moving from an individual with a 1/2 genotype to a 2/2 individual.

Genotype	Additive	Dominant	Recessive
1/1	–1	–1	–1
1/2	0	–1	+1
2/2	+1	+1	+1

trait indicator. Individuals missing the trait value are ignored during analysis, as are SNPs and individuals with low genotyping rates. This latter filtering step is an important tool for removing likely false positives. Instructions for filtering on genotyping success rates appear in Section 0.6.1.1.

SNP genotypes can be analyzed using any of three models: additive (which is also known as codominant and is the default), first allele dominant, or first allele recessive. The SNP model is chosen by setting the keyword SNP_DOMINANCE_MODEL to either Additive (Codominant), Dominant, or Recessive (case insensitive). The genotype encoding schemes for each of these three models is shown in Table 24.1. In this table the "1" and "2" alleles represent the first and second alleles at the SNP. The actual allele names can be listed in the SNP definition input file. For the top SNPs, the results of the SNP Association analysis will include an estimate for the effect size, which is the change in trait value for each positive unit change seen in Table 24.1. For example, when using the default additive (codominant) model, the effect size estimates the trait difference moving from an individual with a 1/2 genotype to a 2/2 individual. In terms of equation (4), the genotype encodings for each SNP are the x_i and the regression coefficients β are the estimated effect sizes.

In addition, each SNP can be analyzed by all three dominance models, with the most significant model reported in the marginal analysis and then carried forward into any subsequent penalized regression or interaction analyses. To maximize over SNP dominance models, set the keyword SNP_DOMINANCE_MODEL to Maximize. Of course this will almost triple the computation time for the marginal analysis. Other analysis times will be unaffected. Unfortunately, the usual Q-Q plot, used to detect systematic bias and inflated p-values, is no longer feasible when you maximize over the three dominance model since the null distribution becomes nonstandard.

We recommend the use of imputation to fill in any missing SNP genotypes before using

Option 24. Of course, Option 23 is one good choice to perform the imputation. If you do not use imputation and missing SNP genotypes remain in the data set, then for a marginal analysis with no co-variates Mendel will only use observed genotypes. For a marginal analysis with co-variates or any penalized regression or interaction analyses, Mendel will replace the missing SNP genotypes with "average" genotypes, that is, genotypes chosen at random based on the allele frequencies in your data set. These replacements are neutral but sub-optimal corrections.

For a qualitative predictor such as sex, Mendel usually ties together the various categories by introducing a sum-to-zero constraint for the regression coefficients. This does not mesh well with penalized regression, so in this analysis Mendel reparameterizes and eliminates the regression coefficient corresponding to the most common category for a factor. Missing quantitative predictors are set to 0, and missing factors are set to their most common category.

24.3 Input Files

The sample control file snippet

```
! Input Files
!
Definition_file = Def24a.in
Pedigree_file = Ped24a.in
SNP_definition_file = SNP_def24a.in
SNP_data_file = SNP_data24a.bin
!
! Output Files
!
Output_file = Mendel24a.out
Summary_file = Summary24a.out
Plot_file = Plot24a.out
!
! Analysis Parameters
!
Analysis_option = GWAS
Model = 1
Quantitative_Trait = SimTrait
```

illustrates the basic setup needed to run Option 24. Most of the input and output files are standard. The standard definition file specifies the quantitative trait of interest, SimTrait in this example, and any non-SNP predictors. As always, the order of the loci in the standard definition file must be genetic-loci, factors, and then quantitative variables. The pedigree

file lists all people in the sample, their non-SNP predictor values, and their values at the quantitative trait. The SNP definition and data files convey to Mendel SNP names and genotypes. Because the default value for the keyword ECHO is No, and it is not reset in this example, Mendel's standard output file will contain little beyond error messages and a reference to the summary file. A plot file is also generated with marginal results for all environmental and genetic predictors. The contents of the plot file are described below. The preceding control file instructs Mendel to carry out ordinary linear regression, `MODEL = 1`, on the trait SimTrait.

To implement logistic regression, `MODEL = 2`, you must have an indicator variable that takes on the value 1 for cases and 0 for controls. (To directly accommodate data sets that use 2 for cases and 1 for controls, Mendel allows the command `CASE2_CONTROL1 = True`. The default value for this keyword is False, unless using Plink-format input files.) Section 0.5.4.6 discusses three transformations that create one–zero indicator variables. The `Above_Threshold` and `Below_Threshold` transformations convert a quantitative variable based on a threshold value. The `Indicator` transformation converts a locus or factor based on the presence or absence of a user-defined affected label. Alternatively, the indicator variable may be predefined in your data set.

In addition to the automatic and mandatory intercept and SNP predictors, you may also include any number and type of environmental predictors and non-SNP genetic predictors using the keyword `PREDICTOR`. For example, the commands

```
Predictor = Sex :: SimTrait
Predictor = Medicated :: SimTrait
Predictor = BMI :: SimTrait
```

in a control file indicate that sex, the factor Medicated, and the variable BMI should be analyzed as possible predictors for the trait SimTrait. Of course SNPs do not need to be listed as predictors as they are all automatically included. We do not recommend regressing on environmental predictors prior to genetic analysis because this precludes the possibility of finding interactions between environmental predictors and SNPs. As much as possible, Mendel treats all predictors equally.

For marginal and penalized regression analyses, any of the predictors can be specified to be retained in the null model and all alternate models using the `RETAINED_PREDICTOR` keyword. For a co-variate to be retained in all models, it must be included as a predictor in the first place (recall that SNPs are automatically included in the list of predictors). Consider the commands

```
Retained_Predictor = Sex
Retained_Predictor = Medicated
```

in a control file. This will result in a null model for the trait that has at least three terms: an intercept, sex, and the Medicated factor. The alternate models will include the null model terms plus each additional predictor under analysis.

In general, Option 24 involves three stages of analysis. Any of the three stages may be omitted, but whenever they are present, they are always run in the order: marginal analysis first, penalized regression analysis second, and interaction analysis third. (The interaction analysis can take one of two forms, either pairwise or penalized regression-based.) Which of the three stages is actually run is controlled by setting the keywords `MARGINAL_ANALYSIS`, `PENALIZED_REGRESSION`, `PAIRWISE_ANALYSIS`, and `PENALIZED_INTERACTION`. For example, the commands

```
Marginal_Analysis = True
Penalized_Regression = False
Pairwise_Analysis = False
Penalized_Interaction = False
```

instruct Mendel to run the simpler, marginal analysis, output the results, and then stop. This is the default setting. To set the number of top marginal predictors printed in the summary file, use a command such as

```
Desired_Predictors = 100 :: Marginal
```

The 100 top marginal predictors are then listed in a table in the summary file in ascending order according to their marginal p-values. Again, 100 is the default.

The summary file begins with a count of the number of tests run and the resulting Bonferroni significance threshold. In the table of the most significant predictors, the intercept term, with its grand mean estimate, is always listed first. For each of the remaining top predictors, the summary file records its position, its marginal p-value, an estimate of its effect size, and the standard error of the estimate. The summary file also reports the name of the allele regressed against and whether this was the major or minor allele. (If the allele names were not listed in the SNP definition file, then the names default to 1 and 2.) In the default additive model, the effect size is the estimate for the change in trait value for each additional copy of the regression allele; see Table 24.1 for more information. Finally, the summary file reports the minor allele frequency, Hardy-Weinberg p-value, genotyping success rate, selected dominance model, and group assignment of each top SNP. As usual, with case-control data, Mendel calculates which allele was the minor allele, the minor allele frequency, and the Hardy-Weinberg p-value using only the controls.

Following the summary of the marginal results for the top SNPs, Mendel presents a brief false discovery rate analysis. This analysis is designed to assist in assessing the significance of the SNPs. If one is willing to accept, for example, a 1% chance of a false

positive, then one should pursue all SNPs with marginal p-value at or below the corresponding p-value threshold.

Option 24 creates a plot file designed to ease graphing of the results of marginal analysis. To name the plot file, invoke the command `PLOT_FILE` in the control file. The plot file contains output similar to the summary file, but for all predictors, both SNP and non-SNP, not just the top few. The columns in the plot file are described in Table 24.2. A Manhattan plot of the results is simply the Chromosome and Base-Pair columns graphed against the $-\log_{10}$(P-Value column), or graphed directly against the QQ_Observ column, which already has the logarithm values listed. It may be instructive to also look at the Q-Q plot of the observed $-\log_{10}$(p-values) against the expected quantiles based on the null distribution. A Q-Q plot can show if you have inflated or realistic p-values. To ease creation of a Q-Q plot, the necessary observed and expected values are listed in adjacent columns in the plot file. The plot file also contains the chromosome name and base-pair position of a SNP whenever these items are supplied in the input files. In the plot file the non-SNP predictors are listed first; after these the SNPs are listed in the same order as they appear in the SNP definition file.

Control of penalized regression analysis, the second stage of analysis, is very similar to the control of marginal analysis. The commands

```
Penalized_Regression = True
Desired_Predictors = 10 :: Penalized
```

instructs Mendel to run the marginal and penalized regression analyses, output the results, and then stop. Mandating penalized regression analysis forces marginal analysis to be run first. The latter of the above two commands sets the number of penalized regression predictors output to the summary file to 10. This is the default. The number of predictors selected in the penalized regression model is automatically increased to at least the number of user-specified retained predictors. With group penalties, Mendel permits some leeway in the number of predictors allowed in the penalized regression model to account for groups of predictors entering and leaving the model en masse. Again, the predictors are listed in ascending order according to their marginal p-values. For each penalized regression predictor listed, the summary file contains its marginal p-value, leave-one-out index as described above, and regression estimate. The summary file also reports the Hardy-Weinberg p-value and minor allele frequency of each selected SNP.

In Option 24 there are two basic types of interaction analysis, the third stage of analysis. In each Mendel run with an interaction analysis, you must chose either a pairwise or penalized regression interaction analysis. To keep the number of computations within reasonable bounds in pairwise analysis, you can select the base set defining the predictors to be analyzed. The choice of a pairwise analysis is made using a command such as

Table 24.2: Description of the Data Columns in the SNP Association Plot File

Column Heading	Description
Predictor_Name	Name of predictor
Chromosome	Chromosome name from SNP definition file*
Base-Pair	Base-pair position from SNP definition file*
P-Value	P-value that the predictor is associated with the trait
Effect_Size	Estimate of effect size (regression coefficient)
Std_Error	Standard error of effect size estimate
Regress	Name of regression allele (if unnamed in input: "1" or "2") *
MA_Freq	Minor allele frequency*†
HW_P-value	P-value for Hardy-Weinberg equilibrium*†
QQ_Observ	For Q-Q plots: observed $-\log_{10}$(p-value)
QQ_Expect	For Q-Q plots: expected quantile value under null model
Weight	Weight of predictor for penalized regression analysis
Typing_Rate	Genotyping success rate*
Model	Dominance model selected*
Group	Name of group that predictor is a member of

*no entry for non-SNP predictors

†in case-control data sets, only controls are used to calculate these values

```
Pairwise_Analysis = Base-All
```

The keyword PAIRWISE_ANALYSIS should be assigned one of the values: Base-All, Base-Base, or None (the default). For this keyword, False and No are considered synonyms for None; True and Yes synonyms for Base-All. If PAIRWISE_ANALYSIS is set to Base-All, then Mendel considers all two-way interactions where one predictor is a member of the base set, and the other can be any other predictor. If PAIRWISE_ANALYSIS is set to Base-Base, then Mendel considers all two-way interactions where both predictors are members of the base set. Obviously, the number of predictors in the base set for the Base-All option should be smaller than the number for the Base-Base option. For example, one can either find the best two-way interactions where at least one of the predictors is among the top 100 marginal predictors (Base-All), or find the best two-way interactions where both predictors are among the top 10,000 marginal predictors (Base-Base). If there are 1 million total predictors, then both these analyses will require 100 million comparisons. All possible two-way interactions on this data set would require 1 trillion comparisons, not a practical number for normal runs even with modern computer systems.

The base set of predictors can be constructed in one of three ways: the top predictors from the marginal analysis, the top predictors from the penalized regression analysis, or the predictors named by the user in the control file. First, a command such as

```
Base_Set = 10 :: Marginal
```

makes the base set the top 10 predictors from the marginal analysis. This is the default value for the keyword BASE_SET. Obviously, this choice requires the marginal analysis be performed first. Alternatively, a command such as

```
Base_Set = 15 :: Penalized
```

makes the base set the top 15 predictors from the penalized regression analysis. Obviously, this requires the penalized regression analysis to be performed first. Finally, commands such as

```
Base_Predictor = rs2256412
Base_Predictor = rs1935681
Base_Predictor = rs289332
```

make the base set consist of the three named SNPs. Any number of SNPs may be named in this way. This type of base set does not require either the marginal or penalized regression analysis to be performed first. Although not required, data initialization proceeds more quickly if the base predictor SNPs are listed in the same order as they appear in the SNP definition file, particularly if that file is large.

The output in the summary file changes slightly depending on whether a Base-All or Base-Base analysis is selected. For a Base-All analysis, no matter which of the three types of base sets is used, for each predictor in the base set, Mendel ranks all two-way interactions in ascending order by p-value. You inform Mendel how many of these interaction models to output to the summary file by inserting a command such as

```
Desired_Predictors = 20 :: Interactions
```

in the control file. This is the default value. Under this command, the initial output to the summary file is the top 20 two-way interactions that include the first predictor in the base set. Next, the top 20 two-way interactions that include the second predictor in the base set will be output. Then, the top 20 that include the third predictor from the base set, and so on. For a Base-Base pairwise analysis, the above command results in the output of the top 20 two-way interactions where both predictors are in the base set. Again the results are listed in ascending order by p-value.

Mendel can also perform interaction analysis via penalized regression. This type of analysis is selected by commands such as

```
Penalized_Interaction = True
Interaction_Levels = 3
Desired_Predictors = 10 :: Penalized
Desired_Predictors = 20 :: Interactions
```

that compel Mendel to consider all one-way (marginal), two-way, and three-way interactions among the top 10 predictors from the earlier penalized regression analysis on the individual predictors. Penalized regression is used to determine the top 20 most potent interactions for output to the summary file. For this interaction analysis via penalized regression, we recommend setting the desired number of predictors in the interaction stage higher than the desired number in the earlier penalized regression analysis on the individual predictors. Please note that the number of possible interactions escalates rapidly. If you want to investigate all d-way and less complex interactions among the s selected predictors from the marginal penalized regression stage, then Mendel must consider $\sum_{j=0}^{d} \binom{s}{j}$ predictors in the interaction stage. If we take $d = s$, then this sum of binomial coefficients equals 2^s, a large number even for s as small as 10. Hence, restrain your ambition in exploring interactions. Two-way interactions are the natural place to start.

24.4 Examples

The sample problem employs simulated data with two contributing SNPs and an interaction between them. The first example control file, Control24a.in, is discussed in the previous section and shown in its entirety here

```
!  Input Files
!
Definition_file = Def24a.in
Pedigree_file = Ped24a.in
SNP_definition_file = SNP_def24a.in
SNP_data_file = SNP_data24a.bin
!
!  Output Files
!
Output_file = Mendel24a.out
Summary_file = Summary24a.out
Plot_file = Plot24a.out
!
!  Analysis Parameters
!
Analysis_option = GWAS
Model = 1
Quantitative_Trait = SimTrait
```

```
Marginal_Analysis = True
Desired_Predictors = 15 :: Marginal
Penalized_Regression = False
Pairwise_Analysis = False
Penalized_Interaction = False
```

The start of the Summary file with the first ten columns for the first ten SNP entries

```
              ADDITIVE (CODOMINANT) MODEL FOR ALL SNPS

           ORDINARY REGRESSION: BEST MARGINAL PREDICTORS

  THIS ANALYSIS IS FOR THE QUANTITATIVE TRAIT NAMED: SimTrait

  SINCE THERE WERE NO USER-SPECIFIED RETAINED PREDICTORS,
  THE NULL MODEL CONTAINS ONLY THE GRAND MEAN TERM,
  WHICH IS: 0.14058

  BELOW ARE THE TOP RESULTS FROM ANALYZING ALL THE ALTERNATIVE MODELS, EACH
  OF WHICH INCLUDED THE NULL MODEL PREDICTORS PLUS ONE ADDITIONAL PREDICTOR.

  IN THIS RUN, P-VALUES WERE CALCULATED FOR 10000 MARGINAL PREDICTORS.
  THUS, THE BONFERRONI SIGNIFICANCE THRESHOLD IS 0.50000E-05,
  WHICH ON THE -LOG10 SCALE IS 5.30103.

  IN THIS RUN, THE ANALYSIS USED 2200 INDIVIDUALS.
```

PREDICTOR NAME	CHR NAME	POSITION IN BP	MARGINAL P-VALUE	MARGINAL -LOG10(P-VAL)	EFFECT ESTIMATE	EFFECT STANDARD ERROR	REGRESS ALLELE	MINOR ALLELE FREQ.	HARDY-WEINBERG P-VALUE
rs2256412	12	41913108	0.15706E-33	33.80393	0.38573	0.03093	2=Major	0.47409	0.18438
rs1935681	6	112651205	0.10517E-10	10.97811	0.25004	0.03658	2=Major	0.25227	0.36657
rs6974676	7	41007729	0.32918E-04	4.48256	0.19062	0.04581	2=Major	0.14364	0.65088
rs11895528	2	34277978	0.75169E-04	4.12396	-0.12766	0.03218	2=Major	0.41023	0.05506
rs3796226	3	71737106	0.92959E-04	4.03171	-0.17797	0.04545	2=Major	0.14977	0.69397
rs7130524	11	69660481	0.00016	3.79990	0.13675	0.03614	2=Major	0.28500	0.36297
rs9473372	6	48436924	0.00017	3.78080	0.12056	0.03195	2=Major	0.49432	0.28911
rs506233	11	75012534	0.00021	3.67741	-0.12185	0.03282	2=Major	0.42955	0.60646
rs2048741	9	120637261	0.00028	3.54952	-0.11700	0.03217	2=Major	0.42114	0.09983
rs4708557	6	169339600	0.00029	3.53319	0.11978	0.03302	2=Major	0.38909	0.72392

of Summary24a.out single out in order the two real predictors and eight non-significant predictors. Consider, for example, the top SNP, rs2256412, since this is analyzed under the additive model, the effect estimate, 0.38573, indicates the change in the trait value expected for each additional second allele, which is the major allele in this data set. In the next section of Summary24a.out we find the results

```
FALSE DISCOVERY RATE THRESHOLDS FOR MARGINAL PREDICTORS:
```

FDR	P-VALUE THRESHOLD	NUMBER OF PASSING PREDICTORS
0.010000	0.10517E-10	2
0.050000	0.10517E-10	2
0.100000	0.10517E-10	2
0.200000	0.92959E-04	5
0.300000	0.00021	8
0.400000	0.00046	14
0.500000	0.00074	15
0.600000	0.00092	19
0.700000	0.00173	29
0.800000	0.00273	39
0.900000	0.00552	62

of the Simes false discovery rate (FDR) procedure [10]. In accord with our expectations, these tallies suggest that only the top two predictors are significant.

The second example relies on the same data set but employs the commands

```
Marginal_Analysis = False
Penalized_Regression = False
Pairwise_Analysis = Base-All
Desired_Predictors = 4 :: Interactions
Base_Predictor = rs2256412
Base_Predictor = rs1935681
Base_Predictor = rs289332
Base_Predictor = rs918119
Base_Predictor = rs9613221
```

to request Mendel to perform a two-way interaction analysis using the named SNPs. The top half of the output file Summary24b.out

```
           ORDINARY REGRESSION: TWO-WAY BASE INTERACTION ANALYSIS

           ALL PAIRWISE INTERACTIONS OF A BASE SET AGAINST ALL PREDICTORS
           THE BASE SET IS 5 USER-SPECIFIED PREDICTORS
```

BASE PREDICTOR NAME	MOST SIGNIFICANT INTERACTING PREDICTORS	LIKELIHOOD RATIO P-VALUE	INTERACTION ESTIMATE	INTERCEPT ESTIMATE	FIRST SLOPE ESTIMATE	SECOND SLOPE ESTIMATE
rs2256412	rs1935681	0.30904E-10	0.32618	0.00326	0.22000	0.23081
rs2256412	rs10975605	0.63122E-04	0.25036	0.04674	0.20324	0.09865
rs2256412	rs2049513	0.00027	0.23984	0.20784	0.20313	-0.11157
rs2256412	rs7220838	0.00031	0.35318	0.12639	0.06828	-0.00821

```
rs1935681     rs2256412      0.30904E-10      0.32618     0.00326     0.23081     0.22000
rs1935681     rs7951348          0.00037     -0.19087     0.01821     0.25028     0.04819
rs1935681     rs4662054          0.00043      0.23562     0.06432     0.09999    -0.07591
rs1935681     rs6750362          0.00115     -0.33196    -0.18308     0.53354     0.23265

rs289332      rs6913778      0.13702E-04      0.21635     0.15248    -0.02261    -0.11152
rs289332      rs3823662          0.00020     -0.18931     0.14070     0.03552     0.00548
rs289332      rs9317530          0.00023     -0.19167     0.12253     0.05161     0.07231
rs289332      rs2943584          0.00030     -0.26407     0.08700     0.18451     0.07700
```

shows results of several interaction models, including the rs2256412 $\times$ rs1935681 interaction assumed in simulating the data. The p-value used to rank these interactions comes from the likelihood ratio test comparing the model with both predictors to the model with both predictors plus their interaction.

The quantitative variable pertinent to these examples was generated by an additive model based on the SNPs rs2256412 and rs1935681, plus their interaction, with the regression coefficients of 0.2, 0.2, and 0.3 respectively. To generate the trait value of each person in the sample, we added a standard normal random deviate to the computed mean effect. In the marginal analysis, only the two true SNPs were judged significant. Not surprisingly, given the limitation of the marginal model compared to the generating model, the estimates of 0.38573 and 0.25004 for the SNP effect sizes are higher than their generating values. In contrast, the interaction analysis estimates of 0.22000 and 0.23081 for the marginal effects and 0.32618 for the interaction effect are gratifyingly close to their generating values. Of course, this is expected since the generating model and analysis model are in close agreement.

The same data set also serves to test the logistic regression model for case-control samples. In the file Control24c.in we deem all individuals with trait values at or above the threshold 0.75 to be cases; the remaining individuals are controls. To convert to this case-control indicator variable, we employ a threshold transformation as discussed in Section 0.5.4.6. The relevant input section of Control24c.in

```
Transform = Above_Threshold :: SimTrait
Indicator_Threshold = 0.75
```

illustrates the appropriate commands. The analysis section of Control24c.in

```
Analysis_option = GWAS
Model = 2
Quantitative_Trait = SimTrait
Marginal_Analysis = True
Penalized_Regression = True
```

```
Penalized_Interaction = True
Interaction_Levels = 2
Desired_Predictors = 100 :: Marginal
Desired_Predictors =  10 :: Penalized
Desired_Predictors =  20 :: Interactions
Uniform_Weights = True
Predictor_Penalty_Proportion = 0.9
Retained_Predictor = rs918119
Retained_Group = 4651
```

requests logistic regression (`MODEL = 2`) and the performance of marginal, standard penalized regression, and penalized regression-based interaction analyses. The value assigned to `UNIFORM_WEIGHTS` indicates whether all SNPs with unassigned weight should be given a uniform weight of 1.0 (the default) or weight $1/\sqrt{4q(1-q)}$, where q is the minor allele frequency at the SNP. The keyword `PREDICTOR_PENALTY_PROPORTION` is used to control whether the group information is used during the standard penalized regression analysis. In terms of the tuning constants in equations (5) and (6), `PREDICTOR_PENALTY_PROPORTION` $= \lambda_P/(\lambda_P + \lambda_G)$. In our example, by setting `PREDICTOR_PENALTY_PROPORTION` to a positive value less than 1.0, Mendel is instructed to weight the group information as well as the data on the individual SNPs. The value assigned to `PREDICTOR_PENALTY_PROPORTION` must be in the interval $[0, 1]$. The value 1.0, which is the default, implies no group weighting; the value 0.0 implies only group weighting. Finally, the last two commands listed above mandate that the candidate SNP rs918119 and the candidate group 4651 always be retained in the standard penalized regression analysis. Users may retain as many SNPs and groups as desired by repeating these commands in their control file.

As mentioned, Mendel usually internally tunes the constants λ_P and λ_G, conditioned on the `PREDICTOR_PENALTY_PROPORTION` value, to select the number of predictors that the user specified via a command such as `Desired_Predictors = 10 :: Penalized` in the control file. Mendel does not vary the value of the third tuning constant γ that is used in the default MCP-based regression. The default value for γ is 3. However, if desired, one can set specific values for all these tuning constants. Once set by the user, their values are not altered. For example, the commands

```
Predictor_Penalty_Proportion = 0.8
Tuning_Constant = 100.0 :: Lambda
```

set Predictor_Penalty_Proportion (PPP) $= \lambda_P/(\lambda_P + \lambda_G) = 0.8$ and $\lambda = \lambda_P + \lambda_G = 100.0$. Thus

$$\begin{aligned} \lambda_P &= \lambda \times \text{PPP} = 100.0 \times 0.8 = 80 \\ \lambda_G &= \lambda - \lambda_P = \lambda \times (1 - \text{PPP}) = 20. \end{aligned}$$

One can also change γ using a command such as `TUNING_CONSTANT = 1.0 :: Gamma` in the control file. The summary file will list the final values of the tuning constants. The drawback of setting specific λ tuning constant values, rather than letting Mendel set them, is that the number of predictors placed in the model may not be close to the desired count. The advantage is that by trying several iterations one can discover the order that the predictors enter the model.

The reader can check the results of the example data set 24c. As usual, in all sections of the summary file, Summary24c.out, the predictors, or sets of predictors, are listed in ascending order by p-value. In this small simulated data set, the standard penalized regression analysis flags most of the same top predictors flagged by the marginal analysis, including the two true predictors. Similarly, the penalized regression interaction analysis again correctly flags the two true predictors and their interaction.

24.5 Germane Keywords

```
ANALYSIS_OPTION = GWAS
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
BASE_PREDICTOR
BASE_SET
CASE2_CONTROL1
DESIRED_PREDICTORS
ESTIMATION_THRESHOLD
INDICATOR_THRESHOLD
INTERACTION_LEVELS
LASSO_PENALTY
MARGINAL_ANALYSIS
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
MODEL
PAIRWISE_ANALYSIS
PENALIZED_INTERACTION
PENALIZED_REGRESSION
PLOT_FILE
PREDICTOR
PREDICTOR_PENALTY_PROPORTION
QUANTITATIVE_TRAIT
RETAINED_PREDICTOR
RETAINED_GROUP
SNP_DOMINANCE_MODEL
TRANSFORM
TUNING_CONSTANT
```

```
UNIFORM_WEIGHTS
```

Random Quotes

Do all the good you can,
By all the means you can,
In all the ways you can,
In all the places you can,
At all the times you can,
To all the people you can
As long as ever you can.

John Wesley

Nina: I love you.
Jamie: I love you.
Nina: I really love you.
Jamie: I really truly love you.
Nina: I really truly madly love you.
Jamie: I really truly madly deeply love you.
Nina: I really truly madly deeply passionately love you.
Jamie: I really truly madly deeply passionately remarkably love you.
Nina: I really truly madly deeply passionately remarkably deliciously love you.
Jamie: I really truly madly passionately remarkably deliciously juicily love you.
(Jamie skips "deeply" and Nina wins the game.)

Anthony Minghella from the film *Truly, Madly, Deeply*

If I had to live my life again, I'd make the same mistakes, only sooner.

Tallulah Bankhead

Several excuses are always less convincing than one.

Aldous Huxley in *Point Counter Point*

Let us cross over the river, and rest under the shade of the trees.

dying words of *Stonewall Jackson*

. . . I cannot find much common ground with someone who believed that the principal source of human woe over the last twenty centuries has been a tragic shortage of selfishness. *David Bentley Hart* in *The Trouble with Ayn Rand*

People demand freedom of speech as a compensation for the freedom of thought that they never use.

Soren Kierkegaard

25 Analysis Option 25: File Conversion

25.1 Background

With the advent of low-cost, high-density, genome-wide SNP genotyping, data files containing billions of genotypes have become common. These large data sets can be compressed from multi-gigabyte text files to binary files of a few hundred megabytes as discussed in Section 0.6. The binary files have the advantage of much easier transportation and quicker loading into analysis programs, but the disadvantage of not being easily edited or even viewed by humans.

Analysis Option 25 uses the input data files, which may be either standard text-based Mendel data files or binary SNP files, to create new input data files that the user specifies to be either text-based or binary. The user can easily specify the set of loci that should be included in the new files. Conversion from text-based to binary input files, or the reverse, is an obvious use for this option. Currently only Analysis Options 10, 15, 21, 23, 24, 25, 27, 28, and 29 can use the binary SNP data sets as input. So, to perform further analysis on a region suggested, for example, by SNP association testing, one needs to convert data on the SNPs in that region from the binary files to standard Mendel input files.

In the past Mendel data files were commonly in column-specific text formats, which are defined in Section 0.7. This older, more restrictive format is often easier to read but harder to edit than the more modern comma-separated format. Analysis Option 25 can also convert data files from column-specific format to files with comma-separated values.

25.2 Appropriate Problems and Data Sets

Binary SNP files are beneficial when very large numbers of SNPs are genotyped and few if any non-SNP loci. A genome-wide association study (GWAS) is the obvious example. However, after a small, genome region of interest has been identified, it is often useful to extract the binary data from that region into ordinary, text-based data files for visualization, ease of editing, and further analysis. This is a common use for Option 25. In this model 1 paradigm, the input files are the standard definition and pedigree files without SNPs and the binary, SNP definition and data files. If the SNP genotypes are of mixed phase, then a SNP phase file must also be included. Also pertinent, if present, is a SNP subset file as defined in Section 0.6.3. This file is used to specify the subset of the SNPs that will be transferred to the standard-format files. If a subset file is not named in the control file, then all SNPs are transferred. Beware that transferring all SNP data will produce multi-gigabyte text files for large SNP data sets. New comma-separated definition, map, and pedigree files are created containing all information on the transferred SNPs and their genotypes, including all phase information.

A secondary use of model 1 is to move data from legacy column-specific definition, map, and pedigree files to more modern list-directed files. (The old-style locus file can also be used in place of the input definition file.) The default is to only move the model loci to the new files. If you wish to convert all loci, regardless of the model loci status, then set the keyword COMPLETE_CONVERSION to true. In that case, the resulting comma-separated definition, map, and pedigree files will contain exactly the same information as the old files but in an easier to edit package. (If the input map file used interval formatting, that is, used distances or recombination fractions rather than positions, then the new map file will use centi-Morgan distances. Thus, a change in the Control file may be necessary to reflect the map distance units when using the new data files.)

For the creation of binary SNP files, which is the model 2 paradigm, Option 25 requires standard definition and pedigree files input in either comma-separated or column-specific format. In addition to storing pedigrees and phenotypes at any non-SNP loci, these two files determine factors and quantitative traits. As always the order of the loci in the standard definition and pedigree files must be genetic-loci, factors, and then quantitative variables. There should not be a penetrance file. Distances are transferred to the SNP definition file when physical distances are listed. New standard definition and pedigree files are created with all SNPs removed. Mendel interprets any locus with two codominant alleles as a SNP. The SNPs and their genotypes are transferred to new SNP definition and binary data files, which should be named in the control file. The genotypes maintain their phase status: unordered genotypes remain unordered, ordered genotypes remain ordered, and missing data remain missing. Thus, if there is a mixture of ordered and unordered SNP genotypes, a new SNP phase file is produced as well.

A secondary use of model 2 is to move a subset of the data from large binary files to new, region-specific binary files. For this purpose, the keyword SNP_SUBSET_FILE (described in Section 0.6.3) and the commands to control filtering on genotyping success rates (described in Section 0.6.1.1) are particularly relevant.

25.3 Input Files

The commands for file conversion are few and simple. Placing MODEL = 1 in the control file informs Mendel to convert the input files to standard, comma-separated files. MODEL = 2 implies conversion to binary SNP data files. The only other requirement is to name the new files. The keywords NEW_DEFINITION_FILE, NEW_MAP_FILE, and NEW_PEDIGREE_FILE should be used to name the corresponding new files in model 1. In model 2, the keywords NEW_SNP_DATA_FILE, NEW_SNP_DEFINITION_FILE, and NEW_SNP_PHASE_FILE play a similar role.

25.4 Examples

Our first example converts a 30 individual and 10,000 SNP data set from comma-separated files to binary SNP data files. The Control25a.in file

```
!
!  Input files
!
DEFINITION_FILE = Def25a.in
MAP_FILE = Map25a.in
MAP_DISTANCE_UNITS = BP
PEDIGREE_FILE = Ped25a.in
!
!  Output files
!
OUTPUT_FILE = Mendel25a.out
SUMMARY_FILE = Summary25a.out
NEW_DEFINITION_FILE = Def25a.out
NEW_MAP_FILE = Map25a.out
NEW_PEDIGREE_FILE = Ped25a.out
NEW_SNP_DEFINITION_FILE = SNP_def25a.out
NEW_SNP_DATA_FILE = SNP_data25a_out.bin
NEW_SNP_PHASE_FILE = SNP_phase25a_out.bin
!
!  Analysis parameters
!
ANALYSIS_OPTION = File_Conversion
MODEL = 2
```

demonstrates the simple commands used in Option 25. The key command is `MODEL = 2`, dictating conversion to binary SNP files. The input pedigree file containing the SNP genotypes is over 1.2 MB with very little wasted space. The new SNP data file is only 80 KB. Obviously this new binary file will be much easier to move about and faster to read into Mendel, although much harder to edit. Since there are mixed ordered and unordered genotypes in the input file, a new SNP phase file is also created. This new file is always close to half the size of the corresponding SNP data file, here another 40 KB. In this example, since there were no non-SNP loci, factors, or variables, the resulting standard definition and map files are empty.

Our second example converts a small subset of the binary SNP data files output by the first example back to comma-separated files. The subset of SNPs we want to convert are listed in the SNP subset file named in Control25b.in

```
!
```

```
!  Input files
!
PEDIGREE_FILE = Ped25b.in
SNP_DEFINITION_FILE = SNP_def25b.in
SNP_DATA_FILE = SNP_data25b.bin
SNP_PHASE_FILE = SNP_phase25b.bin
SNP_SUBSET_FILE = SNP_subset25b.in
!
!  Output files
!
OUTPUT_FILE = Mendel25b.out
SUMMARY_FILE = Summary25b.out
NEW_DEFINITION_FILE = Def25b.out
NEW_MAP_FILE = Map25b.out
NEW_PEDIGREE_FILE = Ped25b.out
!
!  Analysis parameters
!
ANALYSIS_OPTION = File_Conversion
MODEL = 1
MIN_SUCCESS_RATE_PER_SNP = 0.925
MIN_SUCCESS_RATE_PER_INDIVIDUAL = 0.98
```

Here the command `MODEL = 1` dictates conversion to comma-separated, text-based files. Comparing the output pedigree file with the input pedigree from our first example above shows that all information at the converted SNPs is identical, including all phase information, except for the six individuals filtered out due to low genotyping success. At these six individuals no SNP genotypes are listed in the new pedigree file. As mentioned in Section 0.6.1.1, to turn off all filtering, and thus convert all data, set both filtering rate keywords `MIN_SUCCESS_RATE_PER_SNP` and `MIN_SUCCESS_RATE_PER_INDIVIDUAL` to zero.

25.5 Germane Keywords

```
ANALYSIS_OPTION = File_Conversion
COMPLETE_CONVERSION
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
MODEL
NEW_DEFINITION_FILE
NEW_MAP_FILE
NEW_PEDIGREE_FILE
NEW_PENETRANCE_FILE
NEW_SNP_DATA_FILE
```

```
NEW_SNP_DEFINITION_FILE
NEW_SNP_PHASE_FILE
SAMPLE_SUBSET_FILE
SNP_DATA_FILE
SNP_DEFINITION_FILE
SNP_DOMINANCE_MODEL
SNP_PHASE_FILE
SNP_SUBSET_FILE
```

Random Quotes

Yes, I'm sorry. I've learned my lesson and I won't do it again. I've been a bad boy. . . . I'm trying to be nice up here. That may be something you guys are not acquainted with, but you have to understand my position. I'm a very good boy.

Mike Tyson, heavyweight boxer

There is nothing so terrible as the pursuit of art by those who have no talent.

W. Somerset Maugham in *Of Human Bondage*

He has all the characteristics of a dog except loyalty.

Sam Houston

Oh my son's my son till he gets him a wife,
But my daughter's my daughter all her life.

Dinah Maria Mulock Craik in *Young and Old*

I stopped believing in Santa Claus when my mother took me to see him in a department store, and he asked for my autograph.

Shirley Temple

She runs the gamut of emotions from A to B.

Dorothy Parker

I do not like broccoli, and I haven't liked it since I was a little kid, and my mother made me eat it, and I'm president of the United States, and I'm not going to eat any more broccoli.

George Bush, the elder

William McKinley, a kindly soul in a spineless body . . .

Samuel Elliot Morrison

[*Of two boring congressmen*:] They never open their mouths without subtracting from the sum of human knowledge.

Thomas Reed, Speaker of the US House of Representatives in the 1890's

26 Analysis Option 26: Maternal-Fetal Genotype (MFG) Incompatibility Test

26.1 Background

Maternal-fetal genotype (MFG) incompatibility represents an adverse interaction between the genes of a mother and her fetus at a particular locus. Such interactions increase disease risk to the child [98]. The Maternal-Fetal Genotype Incompatibility Test [19, 48, 54, 53, 96] is an affecteds-only, likelihood ratio test (LRT) relying on the joint estimation of offspring allelic effects, maternal allelic effects, and interactions between maternal and offspring genotypes (MFG incompatibility). One can view the MFG test more generally as testing the joint effects of maternal and offspring genotypes on a dichotomous trait. The MFG option of Mendel implements the Extended-MFG (EMFG) Test [19] appropriate for both small and large pedigrees. This test models observed genotypes conditional on certain pedigree members being affected. Let G and D denote the genotypes and phenotypes of a pedigree. The conditional likelihood $L(G|D)$ computed in MFG testing differs from the typical pedigree likelihood [86] because offspring penetrances depend on both maternal and offspring genotypes. Mendel implements conditioning in basically the same way it implements ascertainment correction [59]. The most important difference from earlier implementations of the MFG test is Mendel's parameterization through genotypes rather than mating types. The corresponding assumption of random mating at the locus of interest makes it possible to handle arbitrary pedigrees. When genotypes are missing for some individuals, or genotype phases are unknown, $L(G|D)$ is computed by summing over all possible ordered genotypes consistent with the observed genotypes G in the family. Unaffected offspring are assigned a disease penetrance of 1 and do not contribute directly to the MFG test except for limiting the possible genotypes of parents with missing genotypes.

The MFG test does not require that controls or their family members be collected. As a generalization of the Case Parent Triad approach [109], it presents a flexible alternative to the TDT and gamete competition models. For example, the current option can test for either offspring or maternal allelic effects by imposing appropriate constraints on the parameters. The biggest limitations of the current MFG implementation are that covariates cannot be included and that only biallelic loci and qualitative traits can be analyzed.

For the convenience of the user, Mendel automatically maximizes the loglikelihood under the null hypothesis, which postulates that all mother-offspring genotype combinations confer the same risk of disease. Genotype frequencies are the only parameters estimated under the null hypothesis. Mendel calculates the likelihood ratio test statistic for your alternative model versus the null model and reports the relevant asymptotic p-value. If you want to compare your model to a different null model, then you must run Mendel twice and use

the reported loglikelihoods to calculate the corresponding likelihood ratio test and p-value.

26.2 Appropriate Problems and Data Sets

Pedigrees of any shape or size are admissible. However, they should all be from the same general ethnicity, with similar allele frequencies at the loci of interest. Only biallelic loci will be analyzed. To test a hypothesis using multi-allelic loci, prepare the data using Option 16 to reduce the number of alleles to two. Mendel's affecteds-only analysis relies on treating unaffected individuals and founders as phenotype unknown. Accordingly, the MFG test is inappropriate when the prevalence of the trait or disease exceeds 10%. Since only affecteds are phenotypically identified, the baseline risk to the population cannot be estimated. Pedigrees with affected founders can only be analyzed by supplying those founders with parents that have no phenotype or genotype data in the input files. Pedigrees with a single, affected individual also follows this restriction.

Mendel's parameterization allows for inclusion of arbitrary family structures by assuming random mating; test results are sensitive to severe violations of this supposition [19]. Whenever possible, we recommend testing for departures from random mating in a sample of randomly ascertained parent pairs from the same population as the study sample. When a severe violation of random mating is observed, use the nuclear-family version of the MFG test that does not assume random mating [53, 96]. Alternatively, consider the case-mother, control-mother approach [18].

26.3 Input Files

Option 26 of Mendel estimates the relative risks on a log-scale and the genotype frequencies on their original scale. Table 26.1 describes the three models available to users. In the table, allele "2" represents the risk allele used under models 1 and 2. In your dataset, alleles do not have to be labeled 1 and 2. Some of the parameter names appearing in Table 26.1 for a given model need explaining. The label 0 indicates that the corresponding parameter is fixed at 0. The double name `MFG_M` and `MFG_F` indicates that there are separate male and female offspring penetrance parameters for the listed maternal-fetal genotype combination.

The best known example of MFG incompatibility is RHD incompatibility [103]. Under the default analysis (`MODEL = 1`), Mendel estimates sex-specific RHD-type incompatibility effects. In our discussion of RHD incompatibility, allele 2 corresponds to the antigen coding allele (often coded as D) and allele 1 corresponds to the null allele (often coded as d). RHD incompatibility occurs when the immune system of a mother with genotype `1/1` recognizes the protein product from the fetal `1/2` genotype as foreign and mounts an immune response that can be detrimental to her offspring. The `1/1` – `1/2` genotype combination has an

Table 26.1: MFG Parameters for Various Models of Relative Disease Risk

Mother	Offspring	Model 1: RHD	Model 2: NIMA	Model 3: Generalized Risk
2/2	2/2	0	HOMOZ	U_22
2/2	1/2	0	HETER	U_21
1/2	2/2	0	HOMOZ	U_12
1/2	1/2	0	HETER	U_11
1/2	1/1	0	MFG_M & MFG_F	U_10
1/1	1/2	MFG_M & MFG_F	HETER	U_01
1/1	1/1	0	0	U_00

All risk parameters are on a log-scale. Under models 1 and 2, allele "2" represents the risk allele; the allele labels in the input files are arbitrary.

increased risk over the baseline risk. All other maternal-offspring genotype combinations are at the baseline risk. By default, sex-specific effects are included in model 1. However, the user can make the analysis gender neutral by simply equating the relevant male and female risk parameters (MFG_M and MFG_F).

Another known example of MFG incompatibility involves the effect of non-inherited maternal antigens (NIMA) on rheumatoid arthritis (RA). At HLA-DRB1, the interaction between offspring alleles and NIMA appears to increase the risk of RA in offspring [41, 48, 84, 106]. Under MODEL = 2, Mendel estimates a sex-specific NIMA incompatibility effect, as seen in Table 26.1. NIMA effects occur to 1/1 offspring whose mother has genotype 1/2. As in model 1, the default sex-specific effects can easily be made gender neutral. Also, offspring allelic effects are included in model 2's default parameterization to better match the case of HLA-DRB1 and RA. However, the user can easily remove these offspring effects by simply fixing the relevant parameters (HOMOZ and HETER) at zero.

For MODEL = 3, Mendel implements a general, two-allele model for gender-neutral, maternal-offspring genotype marginal effects and interactions. There are seven possible maternal-offspring genotype combinations, as seen in Table 26.1. Here U_ij denote the log relative risk when the mother's genotype contains i risk alleles and the child's genotype contains j risk alleles. Since these seven parameters are log relative risks, it is necessary to constrain at least one of them to zero. If you do not constrain one parameter to zero, Mendel will terminate your analysis with an appropriate error message.

Model 3 is useful when you want to screen a number of loci and have no prior hypothesis of the mechanism of maternal-fetal genotype incompatibility. Under this general model, the default likelihood ratio test has six degrees-of-freedom and may be underpowered relative to a correctly specified, more detailed model. To avoid over parameterization, we do

not allow for sex-specific effects in model 3. Model 3 allows you to fit various submodels by judiciously constraining the parameters U_00, ..., U_22. The advantage of model 3 over models 1 and 2 is generality; the disadvantages are a loss of power over more parsimonious models and potential difficulties in biological interpretation.

To run the MFG test in Mendel, the user must provide control, pedigree, definition, and map files. The definition, map, and pedigree files are standard, text-based Mendel input files. At each locus that has alleles listed in the definition file, the second allele listed is considered the risk allele for models 1 and 2. If there is no prior hypothesis to the contrary, we recommend designating the least frequent allele as the risk allele. If, for some locus, allele names are not specified in the definition file, then for that locus the second allele lexicographically is considered the risk allele. For each locus, Mendel's output for models 1 and 2 specifies the risk allele used in the analysis.

The control file deserves special comment. In addition to the usual commands to set the input and output files, the analysis option to MFG, and the model number, the control file must also specify the affection status factor and the affection designator. For example, the commands

```
AFFECTED = 1
AFFECTED_LOCUS_OR_FACTOR = HEALTH
```

indicate that all individuals with a "1" phenotype at the factor named HEALTH are affected, everyone else is of unknown affection status. For the current option, affection status must be indicated at a factor, not a locus.

For hypothesis testing or when prior knowledge allows it, the number of parameters can be reduced by choosing an appropriate model or introducing constraints. To constrain the parameters, use the keyword PARAMETER_EQUATION as described in Section 0.10.2. The names of the parameters for each model can be found in Table 26.1. For example, under models 1 and 2, if you want to estimate the MFG parameters under a gender-neutral model, then add the constraint

```
PARAMETER_EQUATION = MFG_M - MFG_F :: 0
```

to the control file. This constraint (MFG_M − MFG_F = 0) forces the male and female relative risk parameters to be equal throughout the analysis. Alternatively, to estimate an MFG effect in female offspring assuming there is no MFG effect in male offspring, add the command

```
PARAMETER_EQUATION = MFG_M :: 0
```

to the control file.

In model 2, the offspring effects can be set to their null values of zero while still allowing the estimation of sex-specific NIMA effects. This is accomplished by including the two constraints

```
PARAMETER_EQUATION = HETER :: 0
PARAMETER_EQUATION = HOMOZ :: 0
```

in the control file. To impose the constraint that the homozygote risk is the square of the heterozygote risk for offspring, recall that the risk parameters are on the log-scale. Thus, the appropriate constraint is

```
PARAMETER_EQUATION = 2*HETER - HOMOZ :: 0
```

For model 3, the user must specify the maternal-fetal genotype combination whose risk parameter is fixed at zero. For example, the command

```
PARAMETER_EQUATION = U_00 :: 0
```

makes all other genotype risk assessments relative to the `1/1` – `1/1` maternal-offspring genotype combination. Enforcing additional constraints on parameters allows one to explore many different scenarios under model 3. For example, to fit a model in which there are maternal and offspring allelic effects (the main effects) but no maternal-offspring genotype interactions, place in the control file the constraints

```
PARAMETER_EQUATION = U_00 :: 0
PARAMETER_EQUATION = U_11 - U_10 - U_01 :: 0
PARAMETER_EQUATION = U_22 + U_11 - U_21 - U_12 :: 0
```

26.4 Examples

All five examples exploit the same simulated data at the two unlinked markers and one trait factor listed in Def26a.in. In the first example, we set `MODEL = 1` in Control26a.in to test for sex-specific MFG effects. By using model 1's default parameterization, we are testing if a gene's effects act in a similar fashion to RHD incompatibility. Summary26a.out contains the results

```
 RESULTS FOR LOCUS NAMED: SNP1
 USING RISK ALLELE NAMED: 2
 *** WARNING *** THE RISK ALLELE IS THE MOST FREQUENT ALLELE.

 NULL MODEL: LOG(LIKELIHOOD) =    -423.9787
 USER MODEL: LOG(LIKELIHOOD) =    -417.7314
```

```
GENOTYPE FREQUENCY ESTIMATES:
     1/1         1/2         2/2
   0.14272     0.43911     0.41817

MODEL PARAMETER ESTIMATES (LOG-SCALE):
    MFG_M       MFG_F
    0.7034      0.1830

THE LIKELIHOOD RATIO TEST COMPARING THE USER MODEL TO THE NULL MODEL
FOR LOCUS SNP1 HAS P-VALUE:  0.00193569


RESULTS FOR LOCUS NAMED: SNP2
USING RISK ALLELE NAMED: 2
*** WARNING *** THE RISK ALLELE IS THE MOST FREQUENT ALLELE.

NULL MODEL: LOG(LIKELIHOOD) =    -334.9988
USER MODEL: LOG(LIKELIHOOD) =    -334.9695

GENOTYPE FREQUENCY ESTIMATES:
     1/1         1/2         2/2
   0.05998     0.32214     0.61788

MODEL PARAMETER ESTIMATES (LOG-SCALE):
    MFG_M       MFG_F
    0.0597     -0.0892

THE LIKELIHOOD RATIO TEST COMPARING THE USER MODEL TO THE NULL MODEL
FOR LOCUS SNP2 HAS P-VALUE:  0.97114614
```

For both SNPs, Mendel warns us that the chosen risk allele is actually the most frequent. Considering the p-values, we see that there is evidence at SNP1, but not SNP2, of increased risk to offspring with the 1/2 genotype if their mothers have the 1/1 genotype. We note there is a much larger effect in male offspring for SNP1 than the female offspring. To test whether the effect for SNP1 is limited to male offspring we rerun model 1 with the additional constraint

```
PARAMETER_EQUATION = MFG_F :: 0
```

listed in the control file Control26b.in. Since Mendel always analyzes only the loci in both the definition and map files, we can restrict our analysis to SNP1 by only listing SNP1 in Map26b.in. Summary26b.out contains the results

```
NULL MODEL: LOG(LIKELIHOOD) =    -423.9787
USER MODEL: LOG(LIKELIHOOD) =    -417.9410

GENOTYPE FREQUENCY ESTIMATES:
     1/1        1/2        2/2
   0.14802    0.43447    0.41751

MODEL PARAMETER ESTIMATES (LOG-SCALE):
    MFG_M      MFG_F
    0.7099     0.0000

THE LIKELIHOOD RATIO TEST COMPARING THE USER MODEL TO THE NULL MODEL
FOR LOCUS SNP1 HAS P-VALUE:  0.00051090
```

Clearly, there is a substantial male effect even when the female effect is removed. Moreover, by treating this male only analysis as the null hypothesis for the previous unrestricted analysis on SNP1, we can test whether there is an effect on females by forming the appropriate likelihood ratio test. The LRT statistic roughly follows the χ^2 distribution. In this case, the LRT statistic is $2 \times (-417.7314 - -417.9410) = 0.419$. With one degree-of-freedom, the corresponding p-value is 0.517. Thus, we fail to find support for an effect in females.

In our third example, we set `MODEL = 2` in Control26c.in to test whether there are offspring allelic or sex-specific MFG effects. In using model 2's default parameterization, we are testing if a gene's effects act in a similar fashion to HLA-DRB1's effects on RA. The first half of Summary26c.out shows the test was not significant for SNP1. However, the second half

```
RESULTS FOR LOCUS NAMED: SNP2
USING RISK ALLELE NAMED: 2
*** WARNING *** THE RISK ALLELE IS THE MOST FREQUENT ALLELE.

NULL MODEL: LOG(LIKELIHOOD) =    -334.9988
USER MODEL: LOG(LIKELIHOOD) =    -322.3683

GENOTYPE FREQUENCY ESTIMATES:
     1/1        1/2        2/2
   0.17809    0.40921    0.41271

MODEL PARAMETER ESTIMATES (LOG-SCALE):
    MFG_M      MFG_F      HETER      HOMOZ
   -0.1610    -0.0315     0.9977     1.7820

THE LIKELIHOOD RATIO TEST COMPARING THE USER MODEL TO THE NULL MODEL
FOR LOCUS SNP2 HAS P-VALUE:  0.00004458
```

indicates there is a significant effect at SNP2 under this model. Since the parameter estimates for MFG_M and MFG_F are close to zero, this result might be explained by the offspring genotypes alone. To test if the offspring genotype effects alone are significant, we remove both sex-specific MFG effects by placing the commands

```
PARAMETER_EQUATION = MFG_M :: 0
PARAMETER_EQUATION = MFG_F :: 0
```

in Control26d.in. We restrict our analysis to SNP2 by only listing SNP2 in Map26d.in. Summary26d.out contains the results

```
 NULL MODEL: LOG(LIKELIHOOD) =    -334.9988
 USER MODEL: LOG(LIKELIHOOD) =    -322.3842

 GENOTYPE FREQUENCY ESTIMATES:
      1/1         1/2         2/2
    0.18016     0.40822     0.41162

 MODEL PARAMETER ESTIMATES (LOG-SCALE):
     MFG_M       MFG_F       HETER       HOMOZ
     0.0000      0.0000      1.0625      1.8510

 THE LIKELIHOOD RATIO TEST COMPARING THE USER MODEL TO THE NULL MODEL
 FOR LOCUS SNP2 HAS P-VALUE:  0.00000332
```

Thus, there is strong support for offspring allelic effects. To test whether there is an effect of NIMA, we compare the unrestricted model from example 26c to the model without MFG effects. Here the LRT statistic is $2 \times (-322.3683 - -322.3842) = 0.0318$. With two degrees-of-freedom, the p-value is 0.9842. Thus, there is no support for the NIMA effect.

Given MODEL = 3 in Control26e.in, we can estimate separate relative risks for each mother-offspring genotype combination. The results in Summary26e.out

```
 RESULTS FOR LOCUS NAMED: SNP1

 NULL MODEL: LOG(LIKELIHOOD) =    -423.9787
 USER MODEL: LOG(LIKELIHOOD) =    -417.7630

 GENOTYPE FREQUENCY ESTIMATES:
      1/1         1/2         2/2
    0.12485     0.43170     0.44344

 MODEL PARAMETER ESTIMATES (LOG-SCALE):
    U_00        U_01        U_10        U_11        U_12        U_21        U_22
```

```
    0.0000     0.3210    -0.0824    -0.1687    -0.3732    -0.3544    -0.2365

 THE LIKELIHOOD RATIO TEST COMPARING THE USER MODEL TO THE NULL MODEL
 FOR LOCUS SNP1 HAS P-VALUE:  0.05300676


 RESULTS FOR LOCUS NAMED: SNP2

 NULL MODEL: LOG(LIKELIHOOD) =    -334.9988
 USER MODEL: LOG(LIKELIHOOD) =    -322.1755

 GENOTYPE FREQUENCY ESTIMATES:
     1/1        1/2        2/2
   0.19021    0.41980    0.38999

 MODEL PARAMETER ESTIMATES (LOG-SCALE):
    U_00       U_01       U_10       U_11       U_12       U_21       U_22
    0.0000     1.0153    -0.1078     1.0867     1.8027     1.0577     1.9171

 THE LIKELIHOOD RATIO TEST COMPARING THE USER MODEL TO THE NULL MODEL
 FOR LOCUS SNP2 HAS P-VALUE:  0.00025905
```

show marginal significance for SNP1 and more significance for SNP2. Examples 26a and 26b above showed more significance for SNP1, as did examples 26c and 26d for SNP2. The better performance of the more parsimonious models illustrates our contention that the more general model is good for screening SNPs but often suffers from a relative loss of power.

26.5 Germane Keywords

```
ANALYSIS_OPTION = MFG
AFFECTED
AFFECTED_LOCUS_OR_FACTOR
MODEL
PARAMETER_EQUATION
```

Random Quotes

Lord Illingworth: All women become like their mothers. That is their tragedy.
Mrs. Allonby: No man does. That is his.

Oscar Wilde in *A Woman of No Importance*

You have delighted us long enough.

Jane Austen in *Pride and Prejudice*

27 Analysis Option 27: Inbred Strains Analysis

27.1 Background

Gene mapping in model organisms can be easier than gene mapping in humans. With inbred mice, generation times are short, and environmental effects can be rigidly controlled. Any genes mapped can be quickly located in humans by synteny. All members of an inbred strain are genetically identical and uniformly homozygous. Diversity is regained by crossing inbred strains, and the more strains involved in a cross, the greater the chance of mapping a relevant gene. For this reason, geneticists are contemplating more ambitious crosses with more contributing strains. Unfortunately, these complex crosses are harder to analyze statistically. This option is designed for QTL mapping in dense SNP genome scans. Pedigree crosses may be arbitrarily complex, and multivariate traits are fair game.

In humans the dominant mapping strategies are linkage and association. The former is more robust; the latter has better resolution. It is not altogether obvious how to make the transition from linkage analysis to association mapping with inbred strains. Linkage mapping operates by tracking recombination events. Association mapping exploits linkage disequilibrium. The great depth of some mouse pedigrees makes linkage mapping very accurate. The rearrangements wrought by recombination enable mapping a trait to the smallest region of overlap defined by conserved strain blocks. However, linkage mapping is plagued by impossibly long computation times in complex pedigrees.

In association mapping, QTL fixed effects are tied to the current marker. The marker is viewed as a candidate gene whose genotypes or alleles directly influence trait means. In effect, association mapping relies on a single SNP to distinguish local strain origins. Here we pursue a strategy that substitutes imputed strain origins for observed genotypes at a single SNP. As the QTL location slides along the genome of an animal, the imputed maternal and paternal origins of the animal change in response to observed genotypes at nearby SNPs. Imputed strain origins serve as predictors of the mean level of a quantitative trait in a statistical model that combines fixed and random effects. Polygenic background and random measurement errors are modeled as random effects. Non-SNP predictors such as age, sex, and diet should be incorporated as mean effects. This option's underlying model mimics what linkage mapping is seeking to accomplish in tracking recombination events and strain blocks.

The two articles [9, 123] provide a full explanation of how Mendel captures polygenic background and imputes local strain origins (maternal strain | paternal strain). These are complicated stories motivated by the need for model realism and computational speed. Our calculation of polygenic correlations between mouse relatives solves a problem dating back to R.A. Fisher. To summarize the theory relevant to modeling polygenic background, suppose there are s strains and t traits. If X_{ik} denotes the polygenic contribution

to trait k of animal i, then the means and covariances of these random variables can be expressed in terms of certain combinatorial entities called strain fractions and strain coefficients, denoted by the symbols f_{ia} and $C_{ij}(a,b)$, respectively. Here a and b index the strains assigned to the two animals i and j. The relevant means and covariances are

$$\begin{aligned} \mathrm{E}(X_{ik}) &= 2\sum_{a=1}^{s} f_{ia}\mu_{ak} \\ \mathrm{Cov}(X_{ik}, X_{jl}) &= 4\,\mathrm{tr}(C_{ij}\Omega_{kl}) = 4\sum_{a=1}^{s}\sum_{b=1}^{s} C_{ij}(a,b)\Omega_{ij}(a,b), \end{aligned}$$

where μ_a is a $t \times 1$ vector of mean effects for strain a and $\Omega_{kl}(a,b)$ is an $s \times s$ matrix of covariance effects for the trait pair k and l. The $st \times st$ matrix Ω constructed from the blocks Ω_{kl} is symmetric and positive semidefinite. Mendel estimates the vectors μ_a and the blocks Ω_{kl} as part of its statistical analysis of the measured traits. Prior to conducting maximum likelihood estimation, Mendel calculates the global strain fractions f_{ia} and strain coefficients $C_{ij}(a,b)$ based on the available pedigree data, ignoring the trait data.

In the absence of SNP data, Mendel can estimate polygenic background and non-genetic mean and variance components such as cage effects and random environment. This exercise in classical biometrical genetics provides insight into non-genetic predictors, the heritability of trait residuals, the structure of multivariate traits, and the impact of shared environment. In QTL mapping, local strain origins are viewed as predictors of trait values on the mean level. Dense SNP genotyping allows one to impute local strain origins with high accuracy. Mendel employs a dynamic programming algorithm with computational complexity linear in both time and computer memory. Dense SNP genotyping also make it easy to approximate the global strain fractions f_{ia} and strain coefficients $C_{ij}(a,b)$ empirically without relying on pedigree structure.

27.2 Appropriate Problems and Data Sets

Analysis Option 27 can operate on either text-based or binary-type data files. The data should include quantitative traits, non-genetic predictors, and a dense SNP map. Pedigree relationships supply prior information and improve imputation accuracy and mapping power, so we encourage the use of carefully kept pedigree records. When full pedigree information is lacking in QTL mapping, remember to group animals by pedigree with the parents of each animal omitted. There is no limit on the number of founding strains. For each founding strain we recommend including at least one representative with good genotyping information. Conflicts among the genotypes of founders of the same strain are resolved by considering the genotypes of descendants. Missing founder genotypes are filled in similarly. Obviously, complete and accurate data are always preferable to incomplete or

inaccurate data. Imputation is occasionally prone to error, particularly at the boundaries of recombination blocks. In simulated problems with a small fraction of missing data, Mendel imputation accuracy usually exceeds 99% [123]. Strains sometimes share chromosome blocks identical by descent. To deal with this complicating feature, Mendel attempts to find the blocks and adjusts its QTL test statistic accordingly.

Trait values are assumed to be normally distributed. If your trait or traits depart from this ideal, consider transforming them as instructed in Section 0.5.4.6. With multivariate traits, some animals will be measured on only a subset of the traits. Mendel will salvage as much information as possible from incomplete trait data. Missing environmental predictors have more severe consequences. Animals with missing predictors are simply dropped from analysis. Fortunately, missing SNP genotypes do not fall into this category because of Mendel's capacity for strain imputation. Model 1 (the default) for the current option carries out association testing. The quicker model 2 is intended for estimation of mean and variance components in the setting of biometrical genetics. Model 3 facilitates visualization of imputed strain origins in a narrow slice of the genome.

27.3 Input Files

This option can use standard pedigree, definition, and map files. For large SNP genome scans, it is also possible to exploit compressed SNP data in binary files. Quantitative traits and non-genetic predictors are stored in the pedigree file and named in the standard definition file. Analysis Option 27 depends crucially on SNP order. Thus, all SNPs should be listed in their genomic order. A few features of this analysis are specific to binary format input files. For example, SNPs without position information in the SNP definition file are omitted from analysis. If there are additional SNPs with ambiguous ordering that should be excluded, omit these using the SNP Subset file described in Section 0.6.3. As usual the individuals and SNPs in the binary input files can be filtered by their genotyping success rates as explained in Section 0.6.1.1. For the current analysis option, the default value for `MIN_SUCCESS_RATE_PER_INDIVIDUAL` is set to 0.0. Thus, no animals are filtered out unless this value is altered.

In QTL mapping Mendel tests for association at sampled points along the genome of a species. It is pointless and time consuming to test each and every SNP because the test statistics for close SNPs are highly correlated. The key word `SNP_SAMPLING_INCREMENT` determines what fraction of the SNPs are actually tested. For instance, if this key word is reset to 1000 from its default value of 100, then every thousandth SNP location along the genome is tested. The test statistic asymptotically follows a χ^2 distribution with $s-1$ degrees-of-freedom, where s is the number of strains. An exception to this rule occurs when one or more strains are identical by descent in a small window surrounding the current location. In this situation s is reduced to reflect the effective number of different

strains in the window. The width of the comparison window is determined by the key word FLANKING_SNPS. For example, resetting this key word to 25 from its default value of 7 creates a comparison window of length $25 + 1 + 25 = 51$.

Most items in the following sample control file should be obvious.

```
!
!  Input Files
!
DEFINITION_FILE = Def27a.in
PEDIGREE_FILE = Ped27a.in
SNP_DEFINITION_FILE = SNP_def27a.in
SNP_DATA_FILE = SNP_data27a.bin
POPULATION_FACTOR = STRAINS              ! strain factor
POPULATIONS = 4                          ! number of strains
!
!  Output Files
!
OUTPUT_FILE = Mendel27b.out
SUMMARY_FILE = Summary27b.out
!
!  Analysis Parameters
!
ANALYSIS_OPTION = Inbred_Strains
MODEL = 1                                ! default model => QTL analysis
FLANKING_SNPS = 20                       ! defines strain identity window
SNP_SAMPLING_INCREMENT = 200             ! defines stride between QTL locations
!
!  Parameters defining the model
!
QUANTITATIVE_TRAIT = trait
PREDICTOR = STRAIN_BACKGROUND :: trait   ! strain polygenic mean effect
COVARIANCE_CLASS = Inbred-Strain         ! strain polygenic variance effect
COVARIANCE_CLASS = Environmental         ! random environment
```

The analysis commands in the last section of this control file name the quantitative trait and specify a minimal model. The model always includes a grand mean, even when, as above, it is not explicitly stated. The listed commands also specify mean effects determined by the different contributing strains and random effects summarizing polygenic background and random environment. Recall that under the polygenic model, a trait is viewed as the sum of a large number of small increments determined by independently acting loci. Mendel automatically adds the QTL mean effect. Other mean effects such as age, sex, and diet can be added as suggested in Option 19. Secondary random effects such as cage effects can also be added, again as described in Option 19.

Adding the command

```
ZOOM_SNP =  rs36775006
```

to the control file allows one to zoom in on the promising SNP rs36775006 and test its QTL location and the locations of all flanking SNPs within the comparison window. (If the data set includes SNP binary files, then the ZOOM_SNP must be one of the SNPs listed in the binary files, not from the standard data input files.) Although using a ZOOM_SNP means locations outside this comparison window are not tested, we still recommend retaining a reasonable SNP_SAMPLING_INCREMENT such as the default value 100. Mendel uses these uniformly scattered locations along the genome to construct the global strain coefficients and strain fractions crucial to recovery of polygenic background when pedigree relationships are unavailable. A small value of SNP_SAMPLING_INCREMENT entails far more computation than necessary in this regard.

27.4 Examples

Our four examples all use the same data simulated to mimic a complex recombinant inbred cross. The data contain three independently simulated pedigrees, each with 154 mice in 15 generations. Each pedigree originates in four mice, one from each of four inbred strains. Values for the quantitative trait, named "trait" in our data set, are available only for the last five generations and the founding generation. Genotypes are available for all mice on 1000 SNPs on each of 19 chromosomes. The locus definitions, pedigree structures, and trait values are provided in the text-based files Def27a.in, Ped27a.in, and SNP_def27a.in. The genotypes are stored in the binary file SNP_data27a.bin.

Our first example uses the command MODEL = 2 in Control27a.in to instruct Mendel to ignore all genotype information and estimate the polygenic background and random environment effects. The results in Summary27a.out

```
                        INBRED STRAINS OPTION

                    SUMMARY FOR MEAN PARAMETERS

 TRAIT      PARAMETER    PREDICTOR          ESTIMATE        STD ERR

 trait          1        GRAND               6.8342         0.0562
 trait          2        129S1              -4.5000         0.0487
 trait          3        A                   1.6667         0.0487
 trait          4        CAST               -2.8333         0.0487
 trait          5        PWK                 5.6667         0.0487
```

record estimates for the grand mean, mean effects for each strain, and an overall random environmental effect. These results are in reasonable agreement with the following values used to generate the data:

Predictor	Value
GRAND	6.75
129S1	-4.34
A	1.47
CAST	-2.89
PWK	5.76

Our second example uses the available genotypes to perform QTL mapping on the simulated inbred cross. Note that the control file Control27b.in is the same as Control27a.in, except for resetting MODEL to its default value of 1 and adding the commands

```
FLANKING_SNPS = 20
SNP_SAMPLING_INCREMENT = 200
```

to control the size of the computation window and the increment between sampled SNPs. For each of the 19 chromosomes, Summary27b.out shows mapping results for five SNPs spaced roughly evenly across the chromosome. The results for chromosome 6

LOCUS NAME	CHROM NAME	MAP POSITION	TRAIT NAME	MOST NEGATIVE EFFECT STRAIN	MOST NEGATIVE EFFECT ESTIMATE	MOST POSITIVE EFFECT STRAIN	MOST POSITIVE EFFECT ESTIMATE	P-VALUE
rs32488890	6	5535423	trait	PWK	-5.9171	A	2.1607	0.668308
rs38640145	6	37053726	trait	CAST	-5.2627	129S1	2.6777	0.557564
rs36775006	6	62054089	trait	PWK	-6.6872	129S1	3.0865	0.101E-06
rs31037202	6	93368729	trait	A	-0.8446	129S1	0.3443	0.313437
rs31713686	6	120429819	trait	CAST	-0.5489	PWK	0.9075	0.229702

shows a particularly promising SNP named rs36775006, whose 10^{-7} p-value easily surpasses the Bonferroni correction level of roughly 10^{-4} for this data set. The next smallest p-value of 0.006 occurs on chromosome 16. At each SNP, the strains and effect sizes for the most positive and most negative strain effects are listed.

In our third example we zoom in on the promising SNP rs36775006 by adding the command

```
ZOOM_SNP = rs36775006
```

to Control27c.in. The summary file Summary27c.out now lists results only for the central SNP rs36775006 and the 20 flanking SNPs on either side. The number of flanking SNPs is controlled by the keyword FLANKING_SNPS. The first half of the results in the output file Summary27c.out

LOCUS NAME	CHROM NAME	MAP POSITION	TRAIT NAME	MOST NEGATIVE EFFECT STRAIN	MOST NEGATIVE EFFECT ESTIMATE	MOST POSITIVE EFFECT STRAIN	MOST POSITIVE EFFECT ESTIMATE	P-VALUE
rs38513387	6	60676228	trait	PWK	-6.5734	129S1	3.9001	0.526E-09
rs36501608	6	60762414	trait	PWK	-6.5734	129S1	3.9001	0.526E-09
rs30433366	6	61078990	trait	PWK	-6.3218	A	3.0073	0.441E-09
rs30436512	6	61111885	trait	PWK	-6.3218	A	3.0073	0.441E-09
rs30436787	6	61203897	trait	PWK	-6.3218	A	3.0073	0.441E-09
rs30440937	6	61209472	trait	PWK	-6.3218	A	3.0073	0.441E-09
rs30443496	6	61215854	trait	PWK	-6.3490	A	3.0661	0.830E-10
rs30445988	6	61222084	trait	PWK	-6.3490	A	3.0661	0.830E-10
rs30437719	6	61225867	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs30441149	6	61273446	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs30442246	6	61297152	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs30447261	6	61378836	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs37752222	6	61712277	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs36537041	6	61715057	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs36269246	6	61753207	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs38724212	6	61794807	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs37394050	6	61795055	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs37114416	6	61795734	trait	PWK	-6.2359	129S1	3.0513	0.787E-10
rs37258809	6	62048149	trait	PWK	-6.6873	129S1	3.0865	0.101E-06
rs37120866	6	62053235	trait	PWK	-6.6873	129S1	3.0865	0.101E-06
rs36775006	6	62054089	trait	PWK	-6.6873	129S1	3.0865	0.101E-06

show even more significant p-values at SNPs near the ZOOM_SNP, which appears last in this half of the summary output. All SNPs in the second half of the output are less significant. These results clearly pinpoint the region of the QTL, which occurs at 61222084 bp (rs30445988) in the model generating the data.

Our fourth example illustrates Mendel's capacity for local strain imputation. The analysis section

```
ANALYSIS_OPTION = Inbred_Strains
MODEL = 3
```

of Control27d.in omits all mention of traits and predictors. At the top of the summary file, Summary27d.out, is a Key with labels for the founder strains

```
 KEY FOR INBRED STRAIN LABELS USED IN IMPUTATION RESULTS
  1   ,  129S1
  2   ,  A
  3   ,  CAST
  4   ,  PWK
```

The imputation results appear next. In particular, for animal 2005 of pedigree number 1 we find the partial results

```
LOCUS          , CHROMOSOME,       BASE, FIRST HAPLOTYPE , SECOND HAPLOTYPE
NAME           , NAME      ,   POSITION, STRAIN SOURCES  , STRAIN SOURCES
rs36687350     , 1         ,    6851210, 1               , 2
rs32251261     , 1         ,   31173089, 1               , 2
rs33433050     , 1         ,   43600025, 1 2             , 1 2
rs32841571     , 1         ,   65687369, 1 2             , 1 2
rs30197853     , 1         ,   80928168, 1               , 2
rs31369358     , 1         ,  103050712, 1               , 2
```

At the SNPs rs33433050 and rs32841571, the first and second strains, 129S1 and A, are identical within a window surrounding each SNP. The length of the window is determined by the keyword FLANKING_SNPS. Thus, Mendel lists both strains as possible sources for each of the two alleles at these SNPs.

27.5 Germane Keywords

```
ANALYSIS_OPTION = Inbred_Strains
COVARIANCE_CLASS
FLANKING_SNPS
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
MODEL
POPULATION_FACTOR
POPULATIONS
PREDICTOR
QUANTITATIVE_TRAIT
SAMPLE_SUBSET_FILE
SNP_DATA_FILE
SNP_DEFINITION_FILE
SNP_PHASE_FILE
SNP_SAMPLING_INCREMENT
SNP_SUBSET_FILE
ZOOM_SNP
```

Random Quotes

> This structure has novel features which are of considerable biological interest . . . It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.
>
> *James Watson and Francis Crick*

> History will be kind to me, for I intend to write it.
>
> *Winston Churchill*

28 Analysis Option 28: Trait Simulation

28.1 Background

Trait simulation is vital in power calculations, development of new statistical methods, and construction of classroom exercises in statistical genetics. The current option aims to make trait simulation as painless as possible. Although complete generality is hardly consistent with this objective, some of the most common simulation tasks can be automated. Analysis Option 28 takes existing genotypes or the output of gene dropping (Option 17) and simulates trait values. The trait simulation can be either univariate via a generalized linear model (`MODEL` = 1) or multivariate via a variance component model (`MODEL` = 2). The vocabulary and syntax for specifying simulation models is accordingly coordinated with Options 14 and Option 19. Readers may want to consult these options for a fuller discussion of some of the fine points met here.

The biggest limitations of Option 28 are the restriction to a single major locus and the generalized linear model assumption that trait correlations are driven solely by this locus. Fortunately, variance component models permit greater flexibility in modeling polygenic background and environmental effects. Most variance component models are built on Gaussian distributions, but Mendel allows one to replace these by multivariate t-distributions in the hope that users will investigate robust statistics less prone to distortion by outliers. Option 28 also permits simulation of pure variance component models with no major locus effects. If a major locus is included, it is not limited to two alleles or to mean effects tied to simple additive, dominant, or recessive models. More complicated scenarios such as imprinting can be simulated.

28.2 Appropriate Problems and Data Sets

Option 28 is appropriate for both pedigree and random sample data. No provision is made for ascertainment via probands. Generally, simulation is fast enough that one can cull simulated pedigrees not meeting the ascertainment criteria. It is the user's responsibility to weed out the inadmissible pedigrees deposited in the new pedigree file. Option 28 ignores pedigree copy numbers for the simple reason that whenever a major locus is involved, every simulation of trait values should be preceded by resimulation of the major locus genotypes. Otherwise, the true level of randomness is reduced. In compensation, Option 17 does respect copy number commands in gene dropping.

In the output, simulated trait values may be left blank for a number of reasons. First, for a particular individual and trait, if relevant predictors, including a defined major locus, are missing values, then no value will be simulated for the trait at that individual. Second, similar to the simulation of genotypes in Option 17, one can define an overall missing

data pattern using the keyword MISSING_DATA_PATTERN. Setting MISSING_DATA_PATTERN to "None" (case insensitive) mandates no global pattern of missing trait values in the output. Setting the value to "Existing_Data" (case insensitive and the default value) causes the simulated trait values in each output pedigree to mimic the missing data pattern at that trait in the corresponding input pedigree. Finally, setting MISSING_DATA_PATTERN to the name of a locus or factor in the Definition file causes the missing data pattern at that locus or factor to be replicated at the simulated trait. For example, to output simulated trait values only for non-founders, create a factor that has non-blank values only at non-founders and assign that factor's name to this keyword. The third reason for missing trait values is that one can superimpose on the overall missing data pattern a rate of randomly missing values, set by the keyword MISSING_AT_RANDOM. This keyword has default value zero, indicating no randomly missing trait values. Assigning the value 0.02 indicates that of all the values that pass muster based on existence of all relevant predictors and the overall missing data pattern, a random 2% will still be left blank. More examples of possible uses of MISSING_DATA_PATTERN and MISSING_AT_RANDOM are described in Section 17.

The keyword SEED discussed in Section 0.11 allows you to simulate different data during each Mendel run. If the keyword MULTIVARIATE_NORMAL is set to false, then the multivariate t-distribution is substituted for the multivariate normal distribution in simulation. In this situation you must also insert the command

```
PARAMETER_INITIAL_VALUE = df :: T-PARAM
```

in the control file, where df is the desired degrees-of-freedom of the t-distribution. Non-integer values of df are allowed.

28.3 Input Files

Option 28 can use either text-based definition and pedigree files or high-density data sets that use binary files. Any map file is unused and ignored. To indicate a major locus set the keyword MAJOR_LOCUS to the name of the locus, which may be listed in either the definition file or the SNP definition file. The alleles of the major locus should be codominant. More than two alleles are allowed if the locus is in the text-based files. If you wish to accommodate imprinting models, the genotypes at the major locus must be coded as ordered. (The output from gene dropping can be either ordered or unordered genotypes; see Option 17 for the relevant instructions.) Finally, variance component models for inbred strains require strain indicators for pedigree founders. Remember to define the keywords POPULATIONS and POPULATION_FACTOR in the control file to designate the number of strains and the field in the pedigree file for reading strain membership.

The simulated trait values will overwrite the values of a quantitative variable that must already be listed in the definition and pedigree files. Any pre-existing values at this variable

are ignored. A new pedigree file will be created that includes the simulated values at this variable. The order of the individuals in the new pedigree file may differ from the input file if there are families listed. Therefore, if binary data files were input, new binary data files are also created since they must be exactly coordinated with the order of the individuals in their corresponding pedigree file.

The simulation file, named using the keyword SIMULATION_FILE, is peculiar to this option. It is comma delimited and functions to initialize mean and variance parameters prior to simulation. Spacing within the file is arbitrary. Consider the simulation file named Simulation28a.in

```
3 Mean Effects
Htcm, GRAND,  175
Htcm, FEMALE, -10
Htcm, MALE,    10
4 Genotype Effects
Htcm, 1, 1,     0.0
Htcm, 1, 2,    10.0
Htcm, 2, 1,    10.0
Htcm, 2, 2,    10.0
```

pertinent to our first example. The top line gives the number of items specifying mean effects unrelated to the major gene. Each of the next three lines of the file specifies a trait, a mean effect for the trait, and the regression coefficient corresponding to the effect. Note that the mean effects for the categories of a predictor should sum to zero, e.g., the female and male categories above for the sex predictor. Mendel presets all mean and variance effects to 0. Line 5 prepares Mendel to read 4 genotypic effects for the major locus named in the control file. Lines 6 through 9 each list a trait, an ordered major locus genotype, and the mean effect of that genotype. The symbols 1 and 2 conveying an ordered genotype refer to the first and second alleles at the major locus. The left allele in an ordered genotype is the maternal allele, and the right allele is the paternal allele. In this example the major locus acts in a dominant fashion, with allele 2 raising height by 10 cm on average. If the input data for an individual at the major locus is an unordered, heterozygous genotype, then the assigned genotype effect will be the average of the two corresponding ordered genotype effects.

In variance components models, one must also initialize the variance components. In our second example, the third part of the simulation file reads

```
6 Covariance Effects
Additive,      Htcm, Htcm, 100
Additive,      Htcm, Wtkg,  80
Additive,      Wtkg, Wtkg, 100
```

```
Environmental, Htcm, Htcm, 100
Environmental, Htcm, Wtkg,  80
Environmental, Wtkg, Wtkg, 100
```

Here the top line supplies the number of items conveying variance components. Each variance component requires the entire covariance matrix for an isolated individual. Covariance between relatives are constructed from these building blocks. In this example, there are two traits, height in cm and weight in kg. Because of symmetry only three covariances must be initialized. Each pertinent line of the simulation file lists the variance component, two traits, and the covariance between the traits contributed by the component. Mendel does not check whether your data specify a legitimate covariance matrix, so be careful. In the special case of an Inbred-Strain variance component, the covariance effects line uses a slightly different format described in the next section. Also be sure to list all effects in the simulation file in the following order: overall mean effects, mean effects of the genotypes at a major locus, and covariance effects. It is a good idea to examine the sample simulation files in preparing your own simulation files.

28.4 Examples

The control file, Control28a.in,

```
DEFINITION_FILE = Def28a.in
PEDIGREE_FILE = Ped28a.in
SIMULATION_FILE = Simulation28a.in
OUTPUT_FILE = Mendel28a.out
SUMMARY_FILE = Summary28a.out
NEW_PEDIGREE_FILE = Ped28a.out
!
! Analysis Parameters
!
ANALYSIS_OPTION = Simulate_traits
MODEL = 1
MAJOR_LOCUS = Marker1
PENETRANCE_MODEL = Normal :: Distribution
PENETRANCE_MODEL = 20 :: Scale
QUANTITATIVE_TRAIT = Htcm
PREDICTOR = SEX :: Htcm
MISSING_DATA_PATTERN = None
MISSING_AT_RANDOM = 0.02
```

for our first example is fairly typical. In addition to naming the standard input and output files, this control file names a simulation file and a new pedigree file. The keyword

MAJOR_LOCUS names the locus whose genotypes shift the trait mean values. The major locus genotype effect values are listed in the second section of the simulation file. Note that no major locus is required. The analysis information in this control file together with the parameter values defined in the previous simulation file specify that the trait variable, height in centimeters, is to be univariate normal with standard deviation 20 and mean 175+10 = 185 in males and mean 175-10 = 165 in females. This trait is also affected by the genotype values at the major gene locus.

The keywords for defining mean effects are discussed in more detail in Option 14. In particular, Table 14.1 summarizes the supported distributional families and their required scale parameters. Table 14.2 lists supported link and inverse link functions. The references [28, 78] cover the basic theory and simple applications of generalized linear models. You can also consult our association testing guide [68] specifically tailored to the needs of genetic epidemiology.

The grand mean (intercept) and male and female offsets are initialized in the simulation file, Simulation28a.in, and echoed in the summary file, Summary28a.out. The scale (standard deviation) and major locus effects are also echoed in the summary file, which reads in part

```
                        SIMULATE_TRAITS

 GENOTYPE EFFECTS ARE BASED ON MAJOR LOCUS: Marker1

                    SIMULATION MEAN EFFECTS:

                                          REGRESSION
  TRAIT               PREDICTOR          COEFFICIENT

 Htcm                 1/1                  0.00000
 Htcm                 1/2                 10.00000
 Htcm                 2/2                 10.00000

 Htcm                 GRAND              175.00000
 Htcm                 FEMALE             -10.00000
 Htcm                 MALE                10.00000
 Htcm                 SCALE               20.00000
```

Following this brief recapitulation of parameter initial values, the summary file reports the sample statistics for the simulated trait variable and predictor variables untouched by simulation. These statistics should help you determine whether simulated trait values match your intended model. Comparison of the input and output pedigree files shows that the variable Htcm has changed, but all other data are preserved. Note that Htcm is not simulated for Jenna Bush because she lacks a major locus genotype. The first and second

alleles of the major locus are called 213 and 217 in the definition file, Definition28a.in, and the pedigree file, Ped28a.in.

Our second example is an elaboration of the first example. The choices

```
ANALYSIS_OPTION = Simulate_traits
MODEL = 2
MAJOR_LOCUS = Marker1
QUANTITATIVE_TRAIT = Htcm
PREDICTOR = SEX :: Htcm
QUANTITATIVE_TRAIT = Wtkg
PREDICTOR = SEX :: Wtkg
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
MISSING_DATA_PATTERN = Existing_Data
MISSING_AT_RANDOM = 0
```

in Control28b.in show that we have elected model 2 with a major locus effect and Additive and Environmental covariance components. We have also included a second trait Wtkg, weight in kilograms, to illustrate simulation of a bivariate trait. The variance component part of the simulation file, Simulation28b.in, was previously discussed. Examination of this file and the output in the summary file, Summary28b.in, make it clear that the major locus impacts only height.

Our third example deals with inbred strains. The bottom portion

```
ANALYSIS_OPTION = Simulate_traits
MODEL = 2
QUANTITATIVE_TRAIT = Trait1
PREDICTOR = SEX :: Trait1                ! sex effect
PREDICTOR = Strain_background :: Trait1  ! strain polygenic mean effect
QUANTITATIVE_TRAIT = Trait2
PREDICTOR = SEX :: Trait2                ! sex effect
PREDICTOR = Strain_background :: Trait2  ! strain polygenic mean effect
COVARIANCE_CLASS = Inbred-Strain         ! strain polygenic variance effect
COVARIANCE_CLASS = Environmental         ! random environment
POPULATION_FACTOR = Group
POPULATIONS = 2
MISSING_DATA_PATTERN = None
MISSING_AT_RANDOM = 0
```

of Control28c.in defines the keywords `POPULATIONS` and `POPULATION_FACTOR` determining strain. Only pedigree founders should be assigned a strain. For each trait the explicit model here includes a sex effect and a polygenic mean effect; an intercept (grand mean)

is automatically added to each model, if not explicitly stated. No major locus is named because the model is purely polygenic.

The simulation file may include a mean effect for each strain. For example, the top two sections of Simulation28c.in

```
5 Mean Effects
Trait1, GRAND,  0.7
Trait1, Male,   1.0
Trait1, Female,-1.0
Trait1, Black,  0.3
Trait1, Red,   -0.3
0 Genotype Effects
```

include the polygenic mean effect for strains Black and Red on Trait1. As noted above, there is an implicit sum to zero constraint for the strain mean effects for each trait because these effects constitute the categories for the predictor "strain-background". Note this is also always true for the categories, female and male, when "sex" is included as a predictor. Since this example includes no major locus, no genotype effects are listed in the simulation file.

Initialization is handled somewhat differently for the Inbred-Strain variance component than for the Additive and Environmental components we have seen up until now. Each of the lines in a simulation file specifying a covariance for the Inbred-Strain component must indicate which strain should be combined with each trait. In essence, the rows and columns of the relevant covariance matrix are indexed by trait-strain pairs. Therefore, in a simulation file, each line providing a covariance effect for the Inbred-Strain component will take the form:

Inbred-strain, first-trait, first-strain, second-trait, second-strain, covariance value

For instance, the two lines

```
Inbred-strain, Trait1, Red,   Trait1, Red, 1.0
Inbred-strain, Trait1, Black, Trait2, Red, 0.0
```

in Simulation28c.in specify a variance of 1 for the combination (Trait1,Red) and a covariance of 0 for the pair of combinations (Trait1,Black) and (Trait2,Red). (See the background discussion in Option 27 for further clarification of this point.) Mendel will fill in unspecified entries of the covariance matrix by symmetry whenever possible. Given the complexity of variance component initialization, you should check the echoed values in the summary file against the input values in the simulation file to determine if Mendel has used the values you intended.

Our last two example data set use binary input files. The Control28d.in file

```
!
! Input Files
!
DEFINITION_FILE = Def28d.in
PEDIGREE_FILE = Ped28d.in
SNP_DATA_FILE = SNP_data28d.bin
SNP_DEFINITION_FILE = SNP_def28d.in
SIMULATION_FILE = Simulation28d.in
!
! Output Files
!
NEW_PEDIGREE_FILE = Ped28d.out
NEW_DEFINITION_FILE = Def28d.out
NEW_SNP_DATA_FILE = SNP_data28d_out.bin
NEW_SNP_DEFINITION_FILE = SNP_def28d.out
OUTPUT_FILE = Mendel28d.out
SUMMARY_FILE = Summary28d.out
!
! Analysis Parameters
!
ANALYSIS_OPTION = Simulate_traits
MODEL = 2
SEED = 1
MAJOR_LOCUS = rs10412915
QUANTITATIVE_TRAIT = simTrait
PREDICTOR = sex :: simTrait
COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
MISSING_DATA_PATTERN = None
MISSING_AT_RANDOM = 0
```

illustrates the commands used to name new binary files. Even though this analysis does not change any genotypes, a new SNP data file is required because the order of the individuals in the new pedigree file may be different than in the input files. Inspection of this control file, and the corresponding simulation file Simulation28d.in, shows that a univariate trait is being simulated. Example data set 28e is similar to 28d (in fact they use the same SNPs and genotypes) except that a bivariate trait is simulated in 28e. We will use the output of data sets 28d and 28e as example input data sets when we discuss pedigree-based GWAS, Option 29.

28.5 Germane Keywords

```
ANALYSIS_OPTION = Simulate_traits
```

```
COVARIANCE_CLASS
MAJOR_LOCUS
MISSING_DATA_PATTERN
MISSING_AT_RANDOM
MODEL
MULTIVARIATE_NORMAL
NEW_PEDIGREE_FILE
PARAMETER_INITIAL_VALUE
PENETRANCE_MODEL
POPULATION_FACTOR
POPULATIONS
PREDICTOR
QUANTITATIVE_TRAIT
SEED
SIMULATION_FILE
```

Random Quotes

It is the education which gives a man a clear conscious view of his own opinions and judgments, a truth in developing them, an eloquence in expressing them, and a force in urging them. It teaches him to see things as they are, to get right to the point, to disentangle a skein of thought, to detect what is sophistical, and to discard what is irrelevant.

John Henry Newman in *The Idea of a University*

If you want people to think well of you, do not speak well of yourself.

Blaise Pascal

Mrs. Long is a selfish, hypocritical woman, and I have no opinion of her.

Jane Austen in *Pride and Prejudice*

Well, between the 16th century and the present day, [the comma] became a kind of scary grammatical sheepdog. As we shall see, the comma has so many jobs as a “separator” . . . that it tears about on the hillside of language, endlessly organizing words into sensible groups and making them stay put: sorting and dividing; circling and herding; and of course darting off with a peremptory “woof” to round up any wayward subordinate clause that makes a futile bolt for semantic freedom. Commas, if you don’t whistle at them to calm down, are unstoppably enthusiastic at this job.

Lynne Truss in *Eats, Shoots & Leaves*

You’re either part of the solution or part of the problem.

Eldridge Cleaver

29 Analysis Option 29: Pedigree GWAS

29.1 Background

Most programs for genome-wide association studies (GWAS), including Mendel's Analysis Option 24, are designed to exploit only unrelated individuals. However, data sets thought to consist of only unrelated individuals may include cryptic relationships that lead to false positives if not correctly treated. In addition, family designs possess compelling advantages for genetic analysis. They are better equipped to detect rare variants, control for population stratification, and facilitate the study of parent-of-origin effects. Pedigrees selected for extreme trait values often segregate a single gene with strong effect. Furthermore, pedigrees are often available as a legacy from the era of linkage analysis. For many years, the classical variance component model has been a powerful tool for mapping quantitative trait loci (QTL) in pedigrees [59]. Polygenic effects are effectively captured as a variance component through the kinship coefficient matrix. If a SNP is at or near a major locus influencing a QTL, the two alleles are considered to shift the trait mean. Fitting a model with all of these fixed effects simultaneously included is nearly impossible even in the absence of pedigree data. On pedigree data, the traditional solution, Option 20, carries out a likelihood ratio test for each marker separately. Unfortunately, likelihood ratio testing is frustratingly slow for large-scale studies.

Ped-GWAS Analysis Option 29 implements an extraordinarily fast score test for association mapping with dense genome-wide SNPs. Samples may include both related and unrelated individuals. Estimation of the kinship coefficients of the individuals in the data set can either be based on explicitly defined pedigrees or estimated from dense markers via covert calls to Option 10. Score tests require no additional iteration under the alternative model. All that is needed is evaluation of a quadratic form combining the score vector and the expected information matrix at the maximum likelihood estimates under the null model [122]. Option 29 (a) works for random sample data, pedigree data, or a mix of both, (b) allows for covariate adjustment, including correction for population stratification, (c) accommodates both univariate and multivariate quantitative traits with missing data, and (d) allows SNPs to be analyzed under additive (codominant), dominant, or recessive models.

29.2 Appropriate Problems and Data Sets

Option 29 tests for association between individual SNPs and one or more quantitative traits. The data may consist of a random sample of people, nuclear families, extended pedigrees, or some mixture of all three. Option 29 can use either text-based SNP data in pedigree files or the binary SNP data files as described in Section 0.6. Binary format obviously is preferred for GWAS data due to the smaller storage and memory requirements.

Missing genotypes are replaced by random genotypes generated conditional on the allele frequencies within the data set. This strategy is neutral but not optimal. It is a good idea to impute missing SNP genotypes using Option 23 before running pedigree GWAS. If several quantitative traits are named in the control file, then by default each is analyzed separately as a univariate trait. To analyze them all simultaneously as a multivariate trait, set the keyword `MULTIVARIATE_ANALYSIS` to true, as discussed in section 29.3.

Besides depending on the score test, the current option relies strongly on estimation of global kinship coefficients between all pairs of individuals in the pedigree file. Option 29 has three options for accomplishing this task. The keyword `KINSHIP_SOURCE` chooses one of the strategies. (The values for this keyword are case insensitive.) The first and fastest strategy, chosen by setting `KINSHIP_SOURCE = pedigree_structure`, uses only the explicit pedigree structures provided in the pedigree file. Clearly this is only a good strategy for constructing kinship coefficients if one is confident that the stated pedigrees are correct and complete, including that pedigree founders are unrelated. This procedure does not capitalize on SNP genotypes. The second fastest, and default, strategy, chosen by setting `KINSHIP_SOURCE = SNPs_within_pedigrees`, uses SNP genotypes to estimate global kinship coefficients for all pairs of individuals within each pedigree listed in the input files. This strategy is fast because most pedigrees are of moderate size and any global kinship Φ_{ij} is set to zero when i and j belong to different pedigrees. The third and final strategy, chosen by setting `KINSHIP_SOURCE = SNPs_using_everyone`, estimates global kinship coefficients from the available SNP data for all pairs of individuals regardless of whether they belong to the same input pedigree. All stated genealogies are ignored at the expense of longer computation times. Run time in kinship estimation is proportional to mn^2, where m is the number of SNPs employed and n is the number of individuals in the largest pedigree. The third strategy effectively lumps all individuals into one large pedigree. In practice, the third strategy still enjoys reasonable run times and should be applied whenever there is doubt about the reliability or completeness of reported pedigrees.

The remaining two options for global kinship estimation pertain to the details of how SNP genotypes are converted to kinship coefficients. (Thus, if the first strategy listed above is employed, namely, only using pedigree structure for kinship estimation, then the following two options have no effect.) First, there is the question of how many SNPs to use. In practice, with common large data sets, we have found using 20% of the SNPs gives fine estimates for the global kinship coefficients. Thus, the default setting for the keyword `SNP_SAMPLING_INCREMENT` is 5, i.e., sample every fifth SNP. Alternatively, setting this keyword to 100 implies only 1% of the SNPs are used, which results in somewhat less accurate estimates but shorter compute times. A value of 1 implies 100% of the SNPs are used, which results in somewhat more accurate estimates but longer run times. The value applied to this keyword can be any positive integer, however, we find 5 (20% of SNPs,

which is the default) to be a reasonable compromise between accuracy and compute times. As an example, using the commands

```
KINSHIP_SOURCE = SNPs_using_everyone
SNP_SAMPLING_INCREMENT = 1
```

in a control file will cause Mendel to use all SNP genotype data to estimate the global kinship coefficients of all pairs of individuals. Compared to Mendel's default settings, this will result in more accurate results although with longer run times. If the pedigree structures listed in the input files are of decent quality, then the increase in accuracy will be small, since members of different pedigrees will probably be at most distantly related.

Of course, for smaller data sets one needs to use a greater fraction of the SNPs and this is enforced using the keyword `MIN_SAMPLED_SNPS`, which has default value 5000. If the SNP sampling increment would result in fewer SNPs being sampled than the value chosen for `MIN_SAMPLED_SNPS`, then the SNP sampling increment is decreased just enough to bring the number of SNPs sampled above the minimum value.

The last option for kinship estimation is a choice in the actual formula that converts from the selected SNP genotypes to a global kinship coefficient. The case insensitive value of the keyword `KINSHIP_METHOD` determines which of two formulae is chosen. The command `KINSHIP_METHOD = GRM`, which is the default, tells Mendel to use the Genetic Relationship Matrix method. Under GRM, the estimate of the global kinship coefficient of individuals i and j is

$$\hat{\Phi}_{ij} = \frac{1}{2S}\sum_{k=1}^{S}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$$

where k ranges over the selected S SNPs, p_k is the minor allele frequency of SNP k, and x_{ik} is the number of minor alleles in individual i's genotype at SNP k. Alternatively, the command `KINSHIP_METHOD = MoM` tells Mendel to use the Method of Moments [26]. Under MoM, the estimate of the global kinship coefficient of individuals i and j is

$$\hat{\Phi}_{ij} = \frac{e_{ij}-\sum_{k=1}^{S}\left[p_k^2+(1-p_k)^2\right]}{S-\sum_{k=1}^{S}\left[p_k^2+(1-p_k)^2\right]},$$

where

$$e_{ij} = \frac{1}{4}\sum_{k=1}^{S}\left[x_{ik}x_{jk}+(2-x_{ik})(2-x_{jk})\right]$$

is the observed fraction of alleles identical by state (IBS) between i and j. Both methods are fairly quick to calculate and provide good estimates, given reasonably dense genome-wide data. When using the GRM method, very rare SNPs (those with minor allele counts

less than 3) are not used since they become over weighted. In general, one can think of the GRM method centering and scaling each genotype, while the MoM method uses the raw genotypes and then centers and scales the final result.

If an individual is missing a trait value, then he or she is not used when analyzing that trait alone. However, if one is analyzing a multivariate trait, then as long as an individual has values for at least one of the included traits, he or she is included in the analysis, as the proper adjustment are made automatically within the score test statistic for any partial data.

There is an analogous condition for each SNP. If the count of minor alleles of a SNP in the entire data set is below a threshold, then that SNP is not analyzed. For example, if the count is zero, that is, the SNP is mono-allelic in the data set, then there is no point in analyzing that SNP. In fact, the default threshold is set to 3, so SNPs with minor allele counts of 2 or less are not analyzed. (This is roughly analogous to the rule of thumb to not include cells with counts below 3 in ANOVAs.) To change this threshold, use a command such as `MIN_MINOR_ALLELE_COUNT = 1`. Setting this keyword to 1 allows analysis of any SNP that is not mono-allelic in the data set. A list of all SNPs not analyzed due to low minor allele counts is included in the standard, detailed output file, whenever `ECHO` is not off. When analyzing multivariate traits, a SNP may be below the threshold for some of the included traits but not for others, since each trait may have a different set of individuals with missing values. In this case, the SNP is included in the analysis, as again proper adjustments are made automatically within the score test statistic for any partial data.

SNP genotypes can be analyzed using any of three models: additive (which is also known as codominant and is the default), first allele dominant, or first allele recessive. The SNP model is chosen by setting the keyword `SNP_DOMINANCE_MODEL` to either Additive (or Codominant), Dominant, or Recessive (case insensitive). The genotype encoding schemes for each of these three models is shown in Table 24.1. In this table the "1" and "2" alleles represent the first and second alleles at the SNP. The actual allele names can be listed in the SNP definition input file. For the top SNPs, the summary results of the analysis will include an estimate for the effect size, which is the change in trait value for each positive unit change seen in Table 24.1. For example, when using the default additive model, the effect size estimates the trait difference moving from an individual with a 1/2 genotype to a 2/2 individual. For multivariate traits, there will be an effect size estimate for each included trait.

Option 29 by default uses multivariate normal distributions during analysis. If you want to switch to a more robust analysis involving the multivariate t distribution, insert the command `MULTIVARIATE_NORMAL = False` in the control file.

The command `OUTLIERS = True` in the control file instructs Mendel to flag outlier people and pedigrees. The default value for this keyword is `False`. The additional commands

```
PEDIGREE_CUT_POINT = 0.1
PERSON_CUT_POINT = 0.02
```

change the defaults of 0.05 and 0.01 for the cutoffs determining what fraction of pedigrees and people are flagged under the null hypothesis. The results of the outlier analysis are listed in the standard, more detailed output file, not in the summary file.

29.3 Input Files

The Control29a.in file,

```
!
!  Input Files
!
DEFINITION_FILE = Def29a.in
PEDIGREE_FILE = Ped29a.in
SNP_DATA_FILE = SNP_data29a.bin
SNP_DEFINITION_FILE = SNP_def29a.in
!
!  Output Files
!
OUTPUT_FILE = Mendel29a.out
SUMMARY_FILE = Summary29a.out
PLOT_FILE = Plot29a.out
!
!  Analysis Parameters
!
ANALYSIS_OPTION = ped-GWAS
QUANTITATIVE_TRAIT = simTrait
PREDICTOR = SEX :: simTrait
COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
DESIRED_PREDICTORS = 10 :: LRT

SNP_SAMPLING_INCREMENT = 5
KINSHIP_SOURCE = SNPs_within_pedigrees

OUTLIERS = True
```

illustrates the basic commands needed to run Option 29. It instructs Mendel to perform pedigree-based GWAS to test for association between the quantitative trait `simTrait` and the SNPs in the data files. The command `PREDICTOR = SEX :: simTrait` indicates that sex should be a covariate in the null and all alternate models. The `COVARIANCE_CLASS`

commands create a model in which trait variation depends on polygenic additive effects and random environment. At least these two covariance classes should usually be included in all user-specified models, i.e., in all ped-GWAS control files. An exception to this rule occurs when kinship estimation is determined pedigree by pedigree, but all pedigrees are singletons. In this case there is no additive information in the data set and the additive covariance class should be left out; Mendel checks for this automatically. Table 19.1 lists the complete set of variance components Mendel recognizes.

The command `DESIRED_PREDICTORS = 10 :: LRT` tells Mendel that after it performs the fast score test on all SNPs, it should rank them, and then perform a likelihood ratio test (LRT) on the top 10 SNPs. The LRT results are output to the summary file. The LRT is more computationally intensive than the score test but it provides effect size estimates and slightly more accurate p-values. Ten is the default value for the desired number of top predictors to undergo a likelihood ratio test.

In contrast to the summary file, the plot output file will contain results for all SNPs, listed in the same order as the input files. For the SNPs with LRT results, effect size estimates and their standard errors are listed; for all other SNPs, those columns are left blank. (For multivariate traits, effect size estimates and their standard errors for each included trait, are only listed in the summary file, not the plot file.) A description of all of the columns of the plot file is in Table 24.2. The main use of the plot file is to ease the production of so-called Manhattan and Q-Q plots by listing all the values to be graphed in adjacent columns.

As explained in detail in the previous section, the keywords `SNP_SAMPLING_INCREMENT` and `KINSHIP_SOURCE` determine how the SNP genotypes are converted to kinship estimates. The default values used here result in good estimates when the provided pedigree structures are reasonably accurate and complete. Finally, and again as described in the previous section, the command `OUTLIERS = True` causes Mendel to search for likely outlier individuals and pedigrees.

If several quantitative traits are listed in the control file along with their accompanying predictors, then by default Mendel will analyze each trait in sequence as a univariate trait. Including several univariate traits in one Mendel run reduces duplicated effort since data initialization and kinship coefficient estimation can be done once rather than multiple times. If you want to analyze all listed quantitative traits simultaneously as a multivariate trait rather than as a sequence of univariate traits, then set `MULTIVARIATE_ANALYSIS = True` in the control file.

The keyword value “all traits” (case insensitive) is useful when a predictor should accompany all listed quantitative traits. For example, the commands

```
QUANTITATIVE_TRAIT = simTrait1
QUANTITATIVE_TRAIT = simTrait2
PREDICTOR = SEX :: all traits
```

```
PREDICTOR = BMI :: simTrait2
```

indicate that sex should be included in the null and all alternate models for both traits but BMI should be included only for simTrait2. If you want to perform univariate analyses on all variables in the input file except those listed as predictors, then invoke the command `QUANTITATIVE_TRAIT = all traits`. In this case, each listed predictor must also be assigned the value "all traits". This construct leads to a concise control file which will analyze a large number of univariate traits under the same model.

In analyzing several univariate traits in a single run, the requested output of each analysis is recorded sequentially in a single standard output file and a single summary file. To avoid the creation of an enormous file, no plot file is produced containing all results for all SNPs. If the keyword `MULTIPLE_PLOT_FILES` is set to true (its default is false), then a separate plot file is produced for each analyzed univariate trait. Each plot file is named using the template `traitname_PLOT_FILE.txt`.

29.4 Examples

We constructed an example data set, 29a, that uses dense SNP data in binary files. To simulate data with realistic linkage disequilibrium (LD) structure, we took advantage of phased sequence data from chromosome 19 on 85 individuals of northern and western European ancestry (originally from the CEPH sample) made publicly available in the 1000 Genomes Project [52]. After we removed markers that were mono-allelic in this set of individuals, 253,141 SNPs remained. Almost half of the SNPs have minor allele frequencies (MAF) below 5%. The haplotype pairs attributed to the 85 CEPH members were reassigned to the 85 founders of 27 pedigree structures selected from the Framingham Heart Study (FHS, http://www.framinghamheartstudy.org). The selected Framingham pedigrees were chosen to reflect the kind of pedigrees commonly collected in family-based genetic studies. The 27 pedigrees encompass 212 people, range in size from 1 to 36 people and from 1 to 5 generations, and contain sibships of 1 to 5 children. The genotypes of non-founders were simulated, using Option 17, conditional on the haplotypes imposed on the founders. All genotypes were recorded as unordered for subsequent analyses.

We next used Option 28 to simulate a univariate trait named simTrait with a major locus at SNP `rs10412915` (position 55,494,740 on chromosome 19; MAF = 0.259). The files we used are included with Mendel as example data set 28d. Examination of the simulation file Simulate28d.in shows that we used a grand mean $\mu = 40$, sex effect $\beta_{\text{sex}} = 6$, major locus effect $\beta_{\text{snp}} = -2$, additive variance $\sigma_a^2 = 4$, and environmental variance $\sigma_e^2 = 2$. (We used no household or dominance variance, $\sigma_h^2 = \sigma_d^2 = 0$.)

Finally, we used the control file shown above at the start of section 29.3 to analyze this family-based data set. We note how fast the analysis completes. In less than three

seconds on an off-the-shelf laptop, Mendel has read 253,141 SNPs over 212 individuals and performed standard quality control checks, which include filtering both SNPs and individuals on genotype success rates. In an additional three seconds, Mendel has completed a score test on all SNPs and a likelihood ratio test on the top 10.

The first section of Summary29a.out

```
        EFFECT ESTIMATES UNDER NULL MODEL WITH NO SNPS INCLUDED

                    SUMMARY FOR MEAN PARAMETERS

 TRAIT              PARAMETER    PREDICTOR          ESTIMATE        STD ERR

 simTrait               1        GRAND               36.9358         0.2684
 simTrait               2        FEMALE              -3.1415         0.1596
 simTrait               3        MALE                 3.1415         0.1596

                    SUMMARY FOR VARIANCE COMPONENTS

 TRAIT              PARAMETER    VARIANCE           ESTIMATE        STD ERR

 simTrait               4        ADDITIVE             4.9865         1.0055
 simTrait               5        ENVIRONMENTAL        1.8435         0.5524

 TRAIT              TOTAL VARIANCE

 simTrait               6.8300

 TRAIT              HERITABILITY    STD ERR

 simTrait               0.7301       0.0641
```

shows the parameter estimates when no SNPs are included in the model. We already see that under this null model, the model parameters, except for the untested major locus, have been recovered quite accurately. It certainly helps that we specified the same set of components for the model as were used to construct the simulation.

The remaining sections of the summary file show the results of testing the SNPs. The (slightly reformatted) first 10 columns of the next section of Summary29a.out

```
      EFFECT ESTIMATES UNDER ALTERNATE MODEL FOR THE TOP SNPS

  IN THIS RUN, P-VALUES HAVE BEEN CALCULATED FOR 219206 SNPS.
  THUS, THE BONFERRONI SIGNIFICANCE THRESHOLD IS 0.22810E-06,
  WHICH ON THE -LOG10 SCALE IS 6.64188.
```

```
   IN THIS RUN, THE ANALYSIS USED 212 INDIVIDUALS.

   THE GENOMIC CONTROL VALUE LAMBDA IS 1.06666.
```

PREDICTOR NAME	CHR NAME	POSITION IN BP	MARGINAL P-VALUE	MARGINAL -LOG10(P-VAL)	EFFECT ESTIMATE	EFFECT STANDARD ERROR	FRACTION OF VAR. EXPLAINED	REGRESS ALLELE	MINOR ALLELE FREQ
rs10412915	19	55494740	0.17066E-08	8.76787	-1.83499	0.28988	0.18915	2=Major	0.25882
rs55799796	19	55493441	0.25161E-07	7.59926	1.69992	0.29286	0.15741	2=Minor	0.24706
rs3826883	19	55494188	0.28821E-07	7.54029	1.76543	0.30614	0.16134	2=Minor	0.22941
rs34804158	19	55494651	0.37076E-07	7.43091	1.68033	0.29328	0.15380	2=Minor	0.24706
rs59004768	19	55488728	0.12852E-06	6.89104	-1.64111	0.30059	0.14190	2=Major	0.23529
rs11667481	19	55497826	0.29125E-06	6.53573	-1.61686	0.30557	0.13533	2=Major	0.22941
rs11672206	19	55495215	0.29125E-06	6.53573	1.61686	0.30557	0.13533	2=Minor	0.22941
rs7253480	19	55498744	0.29125E-06	6.53573	-1.61686	0.30557	0.13533	2=Major	0.22941
rs11671837	19	55498129	0.29125E-06	6.53573	-1.61686	0.30557	0.13533	2=Major	0.22941
rs35121332	19	55487042	0.33104E-06	6.48011	-1.60399	0.30461	0.13076	2=Major	0.22353

shows that an alternate model that includes the causal SNP, rs10412915, is well supported. It is the highest ranked of the 219,206 SNPs that passed the quality control procedure. The estimate for the effect of the SNP on the trait is also a good fit for the true simulation parameter. The above snippet shows that four other SNPs also pass the Bonferroni significance threshold, which for convenience is listed at the top of this output section. Also listed when analyzing a univariate trait is the genomic control value λ [7]. The top four SNPs are within 1300 base-pairs of the major locus, and obviously in linkage disequilibrium (LD). It is instructive to note that further down in the list there is a group of four SNPs that have identical results, except for the minor allele label and the sign of the effect estimate. These four SNPs must be in complete LD, although the allele labels may swap. The sign change in the effect estimate is because that estimate is always relative to the change caused by the second allele (see Table 24.1), which in some of these SNPs is the minor allele and in some the major. We note that the minor allele, minor allele frequency, and Hardy-Weinberg p-value are calculated using only pedigree founders from the data set.

The next section of Summary29a.out

```
 FALSE DISCOVERY RATE THRESHOLDS FOR INDIVIDUAL PREDICTORS:
```

FDR	P-VALUE THRESHOLD	NUMBER OF PASSING PREDICTORS
0.010000	0.12852E-06	5
0.050000	0.27889E-05	29
0.100000	0.13268E-04	32
0.200000	0.36431E-04	49
0.300000	0.94110E-04	75
0.400000	0.00014	79

```
0.500000          0.00021        95
0.600000          0.00073       267
0.700000          0.00105       333
0.800000          0.00471      1292
0.900000          0.05475     13337
```

tallies the p-value thresholds, and number of SNPs meeting that threshold, for various levels of the Simes false discovery rate (FDR) procedure [10]. So, for example, if one is willing to have a 10% chance of a false positive, then one would pursue the regions containing the top 32 SNPs, which in this case is only one region.

Eleven (slightly reformatted) columns of the header line and lines surrounding the major locus in Plot29a.out

```
Predictor  ,Chr, Base-Pair,     P-Value,  Effect, Std_Err,  MA_Freq, HW_Pval,  QQ_Obs,  QQ_Exp, Group

rs55799796 , 19,  55493441, 0.25161E-07,  1.69992, 0.29286, 0.24706, 0.91260, 7.59926, 5.16476,   -
rs10403648 , 19,  55493651,     0.04011,       - ,      - ,      - ,      - , 1.39680, 1.33903, NLRP2
rs17699678 , 19,  55493728,     0.62600,       - ,      - ,      - ,      - , 0.20342, 0.19335, NLRP2
rs56073572 , 19,  55493847,     0.85881,       - ,      - ,      - ,      - , 0.06610, 0.06353, NLRP2
rs193065675, 19,  55494071,     0.38238,       - ,      - ,      - ,      - , 0.41750, 0.40268, NLRP2
rs4306647  , 19,  55494157,     0.04011,       - ,      - ,      - ,      - , 1.39680, 1.33903, NLRP2
rs3826883  , 19,  55494188, 0.28821E-07,  1.76543, 0.30614, 0.22941, 0.77141, 7.54029, 4.94291, NLRP2
rs147207921, 19,  55494320,     0.52477,       - ,      - ,      - ,      - , 0.28003, 0.26733, NLRP2
rs113204023, 19,  55494540,     0.76269,       - ,      - ,      - ,      - , 0.11765, 0.11311, NLRP2
rs3745905  , 19,  55494632,     0.22339,       - ,      - ,      - ,      - , 0.65093, 0.62486, NLRP2
rs34804158 , 19,  55494651, 0.37076E-07,  1.68033, 0.29328, 0.24706, 0.91260, 7.43091, 4.79678, NLRP2
rs10412915 , 19,  55494740, 0.17066E-08, -1.83499, 0.28988, 0.25882, 0.69472, 8.76787, 5.64188, NLRP2
rs145361990, 19,  55494852,     0.69748,       - ,      - ,      - ,      - , 0.15647, 0.14845, NLRP2
rs11672113 , 19,  55494881,     0.00019,       - ,      - ,      - ,      - , 3.73143, 3.41400, NLRP2
rs7254951  , 19,  55495106,     0.59498,       - ,      - ,      - ,      - , 0.22550, 0.21374,   -
rs11672206 , 19,  55495215, 0.29125E-06,  1.61686, 0.30557, 0.22941, 0.77141, 6.53573, 4.52794,   -
```

illustrates the format of the plot file. Only the top SNPs are subject to the likelihood ratio test. Therefore, effects size estimates and associated standard errors are missing for the majority of SNPs, since score tests do not provide effect size estimates. Table 24.2 gives a description of all plot file columns. We again note that here the minor allele, minor allele frequency, and Hardy-Weinberg p-value are calculated using only pedigree founders from the data set. In addition, we only calculate these values for the top SNPs. The Q-Q observed and expected values are calculated for all SNPs to allow easy graphing of the associated Q-Q plot.

Finally, the standard output file, Mendel29a.out, lists the results of the outlier analysis

```
                POSSIBLE OUTLIER INDIVIDUALS

 PEDIGREE     PERSON      QUADRATIC FORM      P-VALUE
 2            6952              9.24          0.00237
```

```
 TRAIT       OBSERVED     PREDICTED    STANDARD DEVIATION
 simTrait     48.06         41.41           2.187


               POSSIBLE OUTLIER PEDIGREES

            DEGREES OF    QUADRATIC
 PEDIGREE     FREEDOM        FORM         P-VALUE

 2               8           15.93        0.04339
```

This indicates that the individual named 6952 in the pedigree named 2, and that entire pedigree, may be outliers. Normally, one would double check these suspicious values. Of course, if one decides to remove these outliers, then one must report this action and its rationale when reporting the results.

The second example data set, 29b, carries out pedigree-based GWAS on the same data and using the same model as 29a. However, Control29b.in has the commands

```
KINSHIP_SOURCE = SNPs_using_everyone
SNP_SAMPLING_INCREMENT = 1
```

which cause Mendel to use all the SNPs while estimating the global kinship coefficients for all pairs of individuals.

In comparison to the 29a analysis, one first notices that the total run time increases from six seconds to nine seconds. Of course, this is due to the larger number of SNPs and pairs of individuals used in the kinship estimation. Second, the genomic control λ decreases from 1.067 to 1.048, which may be an indication that many of the results are slightly more accurate. Finally, since the pedigrees defined in the input files are probably quite complete, the parameter estimates and p-values do not change much from 29a to 29b. In particular, the null model analysis results are very similar in 29a and 29b. The (slightly reformatted) first 10 columns of the alternative analyses in Summary29b

```
       EFFECT ESTIMATES UNDER ALTERNATE MODEL FOR THE TOP SNPS


   IN THIS RUN, P-VALUES HAVE BEEN CALCULATED FOR 219206 SNPS.
   THUS, THE BONFERRONI SIGNIFICANCE THRESHOLD IS 0.22810E-06,
   WHICH ON THE -LOG10 SCALE IS 6.64188.

   IN THIS RUN, THE ANALYSIS USED 212 INDIVIDUALS.

   THE GENOMIC CONTROL VALUE LAMBDA IS 1.04835.
```

PREDICTOR NAME	CHR NAME	POSITION IN BP	MARGINAL P-VALUE	MARGINAL -LOG10(P-VAL)	EFFECT ESTIMATE	EFFECT STANDARD ERROR	FRACTION OF VAR. EXPLAINED	REGRESS ALLELE	MINOR ALLELE FREQ
rs10412915	19	55494740	0.47926E-08	8.31943	-1.82343	0.29709	0.18740	2=Major	0.25882
rs55799796	19	55493441	0.96138E-07	7.01711	1.65955	0.29924	0.15052	2=Minor	0.24706
rs34804158	19	55494651	0.15194E-06	6.81832	1.63949	0.30056	0.14691	2=Minor	0.24706
rs3826883	19	55494188	0.18790E-06	6.72608	1.69073	0.31306	0.14847	2=Minor	0.22941
rs59004768	19	55488728	0.10876E-05	5.96354	-1.55703	0.31000	0.12816	2=Major	0.23529
rs71367132	19	55497506	0.18932E-05	5.72281	1.48404	0.30185	0.12037	2=Minor	0.24706
rs2116886	19	55497039	0.18932E-05	5.72281	-1.48404	0.30185	0.12037	2=Major	0.24706
rs1036231	19	55497843	0.18932E-05	5.72281	-1.48404	0.30185	0.12037	2=Major	0.24706
rs1036232	19	55497943	0.18932E-05	5.72281	1.48404	0.30185	0.12037	2=Minor	0.24706
rs172006	19	55484559	0.21122E-05	5.67527	1.62368	0.33121	0.12393	2=Minor	0.20000

show that again the causal SNP is the highest ranked. The p-values become somewhat less significant in 29b compared to 29a. Indeed, the causal SNP is now the only SNP to pass the 0.01 FDR threshold

FALSE DISCOVERY RATE THRESHOLDS FOR INDIVIDUAL PREDICTORS:

FDR	P-VALUE THRESHOLD	NUMBER OF PASSING PREDICTORS
0.010000	0.47926E-08	1
0.050000	0.10876E-05	5
0.100000	0.66717E-05	17
0.200000	0.12969E-04	29
0.300000	0.12969E-04	29
0.400000	0.54944E-04	32
0.500000	0.00013	63
0.600000	0.00013	63
0.700000	0.00022	69
0.800000	0.00028	79
0.900000	0.00037	91

This example illustrates the tremendous amount of information inherent in dense SNP genotypes. Even ignoring all stated pedigree structures, one can quickly estimate enough relationship information to perform proper association studies on members of extended families.

Our last example showcases Mendel's capability to perform pedigree-based GWAS on multivariate traits, which can be used to detect pleiotropic effects. We use the same pedigree structure and SNP definitions and genotypes as in example data set 29a. We again used Option 28 to simulate trait values based on the same major locus used in data set 29a, `rs10412915`. Here we simulated two correlated quantitative traits, `simTrait1`

and simTrait2. The files we used are included with Mendel as example data set 28e. Examination of the simulation file Simulate28e.in shows that for simTrait1 we used grand mean $\mu_1 = 40$, sex effect $\beta_{\text{sex},1} = 6$, major locus effect $\beta_{\text{snp},1} = -1.5$, additive variance $\sigma^2_{a1} = 4$, and environmental variance $\sigma^2_{e1} = 2$. For simTrait2 we used $\mu_2 = 20$, $\beta_{\text{sex},2} = 4$, $\beta_{\text{snp},2} = -1.5$, $\sigma^2_{a2} = 4$, and $\sigma^2_{e2} = 2$. The covariances between the traits are $\sigma^2_{a1,a2} = 1$ and $\sigma^2_{e1,e2} = 0$. Compared to the univariate trait used in examples 29a and 29b, SNP effects are reduced for each trait while variance components are held fixed.

We note that even for this multivariate analysis, the total run time was under eight seconds on a normal laptop computer. In the relevant section of Control29c.in

```
ANALYSIS_OPTION = ped_GWAS
QUANTITATIVE_TRAIT = simTrait1
QUANTITATIVE_TRAIT = simTrait2
PREDICTOR = SEX :: all traits
COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
DESIRED_PREDICTORS = 10 :: LRT
MULTIVARIATE_ANALYSIS = True
```

since the keywords controlling the kinship estimation procedure are not listed, their default values are used, which are the same values as in data set 29a. We note the use of the shorthand “all traits” to indicate that sex should be used as a predictor for each of the traits.

The first section of Summary29c.out

```
      EFFECT ESTIMATES UNDER NULL MODEL WITH NO SNPS INCLUDED

                  SUMMARY FOR MEAN PARAMETERS

TRAIT            PARAMETER    PREDICTOR          ESTIMATE        STD ERR

simTrait1            1        GRAND              37.5552         0.2622
simTrait1            3        FEMALE             -3.2888         0.1560
simTrait1            4        MALE                3.2888         0.1560
simTrait2            2        GRAND              17.8913         0.2673
simTrait2            5        FEMALE             -2.1842         0.1556
simTrait2            6        MALE                2.1842         0.1556

                  SUMMARY FOR COVARIANCE PARAMETERS

TRAIT            TRAIT              VARIANCE          ESTIMATE      STD_ERR_1     STD_ERR_2

simTrait1        simTrait1          ADDITIVE            4.5992        1.1135        1.1176
simTrait2        simTrait1          ADDITIVE            1.2789        0.7701        0.7754
simTrait2        simTrait2          ADDITIVE            4.7825        1.0408        1.0588
simTrait1        simTrait1          ENVIRONMENTAL       1.8369        0.6396        0.6444
simTrait2        simTrait1          ENVIRONMENTAL      -0.0876        0.4294        0.4360
simTrait2        simTrait2          ENVIRONMENTAL       1.7279        0.5764        0.5985
```

```
 TRAIT             TRAIT           TOTAL VARIANCE OR COVARIANCE

 simTrait1         simTrait1                6.4361
 simTrait2         simTrait1                1.1913
 simTrait2         simTrait2                6.5103
```

shows the parameter estimates under the null model, that is, without a major locus. Again, the fits are reasonable.

The (slightly reformatted) first 10 columns of the SNP table of Summary29c.out

```
      EFFECT ESTIMATES UNDER ALTERNATE MODEL FOR THE TOP SNPS

   P-VALUES HAVE BEEN CALCULATED FOR 219206 SNPS.
   THUS, THE BONFERRONI SIGNIFICANCE THRESHOLD IS 0.22810E-06,
   WHICH ON THE -LOG10 SCALE IS 6.64188.
```

PREDICTOR NAME (Trait)	CHR NAME	POSITION IN BP	MARGINAL P-VALUE	MARGINAL -LOG10(P-VAL)	EFFECT ESTIMATE	EFFECT STANDARD ERROR	FRACTION OF VAR. EXPLAINED	REGRESS ALLELE	MINOR ALLELE FREQ
rs10412915	19	55494740	0.99098E-09	9.00394				2=Major	0.25882
simTrait1					-1.34295	0.29812	-		
simTrait2					-1.57523	0.28925	-		
rs1036231	19	55497843	0.16238E-07	7.78948				2=Major	0.24706
simTrait1					-1.27097	0.29154	-		
simTrait2					-1.41487	0.28699	-		
rs1036232	19	55497943	0.16238E-07	7.78948				2=Minor	0.24706
simTrait1					1.27097	0.29154	-		
simTrait2					1.41487	0.28699	-		
rs2116886	19	55497039	0.16238E-07	7.78948				2=Major	0.24706
simTrait1					-1.27097	0.29154	-		
simTrait2					-1.41487	0.28699	-		
rs71367132	19	55497506	0.16238E-07	7.78948				2=Minor	0.24706
simTrait1					1.27097	0.29154	-		
simTrait2					1.41487	0.28699	-		
rs11667481	19	55497826	0.24745E-07	7.60651				2=Major	0.22941
simTrait1					-1.37505	0.30213	-		
simTrait2					-1.38360	0.30106	-		
rs11671837	19	55498129	0.24745E-07	7.60651				2=Major	0.22941
simTrait1					-1.37505	0.30213	-		
simTrait2					-1.38360	0.30106	-		
rs11672206	19	55495215	0.24745E-07	7.60651				2=Minor	0.22941
simTrait1					1.37505	0.30213	-		
simTrait2					1.38360	0.30106	-		
rs7253480	19	55498744	0.24745E-07	7.60651				2=Major	0.22941
simTrait1					-1.37505	0.30213	-		
simTrait2					-1.38360	0.30106	-		
rs3826883	19	55494188	0.25905E-07	7.58661				2=Minor	0.22941
simTrait1					1.43403	0.30696	-		
simTrait2					1.36414	0.30771	-		

illustrates that despite the reduction in SNP effect sizes, testing both traits simultaneously boosts power. We again see the causal SNP is well supported and highest ranked. Note

that at each of the top SNPs an effect estimate and associated standard error is provided for each individual trait. At the major locus, the effect estimates for each trait recovers the true simulation parameters quite well. Summary29c.out shows that 18 SNPs pass the Bonferroni significance threshold. All 18 are within 6 Kb of the major locus.

Finally, we note that Mendel's multivariate analysis can include any number of linear constraints on the parameters. The syntax for creating these constraints is described in Section 0.10.2. For example, if in the current analysis we want to constrain the sex effects to be equal on the two traits (which is common when the two traits are two time points in a longitudinal study), then we would put a command such as

```
PARAMETER_EQUATION = PAR03 - PAR05 :: 0.0
```

in the control file. This command forces parameter 3, which corresponds to the female predictor for trait 1, minus parameter 5, which corresponds to the female predictor for trait 2, to be zero. The parameter numbers are found in the null model analysis output in the summary file and, when the ECHO keyword is at least partially turned on, in the standard output file.

29.5 Germane Keywords

```
ANALYSIS_OPTION = PED-GWAS
COVARIANCE_CLASS
DESIRED_PREDICTORS
KINSHIP_METHOD
KINSHIP_SOURCE
MIN_MINOR_ALLELE_COUNT
MIN_SAMPLED_SNPS
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
MODEL
MULTIPLE_PLOT_FILES
MULTIVARIATE_ANALYSIS
MULTIVARIATE_NORMAL
OUTLIERS
PEDIGREE_CUT_POINT
PERSON_CUT_POINT
PLOT_FILE
PREDICTOR
QUANTITATIVE_TRAIT
SNP_DOMINANCE_MODEL
SNP_SAMPLING_INCREMENT
TRANSFORM
```

30 Analysis Option 30: QMFG LRT

30.1 Background

Quantitative trait maternal-fetal genotype (QMFG) incompatibility represents an interaction between the genes of a mother and her fetus at a particular locus that ultimately affects the quantitative trait values of the offspring. One can view the QMFG test more generally as testing the joint effects of maternal and offspring genotypes on a quantitative trait (Clark et al., 2015). The QMFG option of Mendel is an application of the variance components model used in other options of Mendel (Options 19, 20, and 29). In the current option, joint maternal and offspring effects including MFG incompatibilities are included as fixed effects and the variance components allow for residual covariation between family members. The QMFG test can use nuclear and extended pedigrees simultaneously and can incorporate additional covariates into the analysis.

Option 30 tests for significant associations using likelihood ratio tests (LRTs). An advantage of the QMFG LRT is that, in maximizing the likelihood, parameter estimates are calculated. Additionally, complex null and alternative hypotheses are easily handled. Because iterative maximization of the likelihood under both the null and alternative models is required, this approach can be computationally burdensome when using many extended pedigrees or a large numbers of markers. For that reason, a separate option (Option 31) runs the QMFG score test, in which key quantities can be pre-computed once under the null and reused for the analysis of each SNP. This pre-computing makes the calculation of the score test statistic for each SNP rapid and makes Option 31 a valuable tool to quickly screen markers for associations with MFG incompatibility. Option 30 has the limitation that the kinship matrix is calculated assuming the pedigree structure is known exactly whereas Option 31 has the added benefit of allowing the user to choose either this theoretical kinship matrix or to use kinships calculated from a genetic relationship matrix using genome-wide SNP data.

30.2 Appropriate Problems and Data Sets

Pedigrees of any shape or size are admissible. The user should be aware that the number of variance components (effectively nuisance parameters in this analysis) that can be estimated depends on the pedigree structure. For example, if the data consist of only trios or mother-child pairs, then only the environmental variance can be specified. Because the QMFG test requires genotype data from mothers, quantitative trait data from founders (including singletons) are not used. Additionally, usable offspring are restricted to those who are genotyped at the locus of interest, have a quantitative trait value, and have mothers who are also genotyped at the locus of interest. If a Mendelian error exists between a

mother-offspring pair (an `A/A` mother with `B/B` offspring or vice versa), an error message is produced. To use these offspring, these errors must be resolved or the genotypes for the offending markers removed for these individuals. Preprocessing the pedigrees using Option 5 (mistyping) can be used to facilitate this step. Only biallelic loci will be analyzed. Analysis Option 16 (combining alleles) can be used to reduce multi-allelic loci to two alleles.

Table 30.1: Examples of MFG Incompatibility

Maternal Genotype	Offspring Genotype	QMFG factor	Parameter labels based on input file order	RHD incompatibility effect	NIMA & offspring genotypic effects
A/A	A/A	00	PAR02	Ref	Ref
A/A	A/B	01	PAR03	Ref	Heter
A/B	A/A	10	PAR04	Ref	NIMA
A/B	A/B	11	PAR05	Ref	Heter
A/B	B/B	12	PAR06	Ref	Homoz
B/B	A/B	21	PAR07	RHD	Heter
B/B	B/B	22	PAR08	Ref	Homoz

Option 30 takes text-based files only. Analysis Option 25 (file conversion) can be used to extract specified loci from binary files. For the user-specified locus of interest, Mendel calculates the variant allele counts for mother and offspring from the genotypes in the pedigree file. For this option, the reference allele is the first allele listed in the definition file. If alleles are not listed in the definition file, Mendel will set the reference allele to the first allele consistent with their lexicographic order. Once variant (non-reference) allele counts are determined, the QMFG LRT option internally includes a QMFG factor into the first factor spot of the pedigree. An empty column must therefore be included in the pedigree file (see accompanying sample input files for Option 30). This new factor should have seven possible values, representing the seven possible maternal-fetal gene-gene combinations. The seven values of the QMFG factor are `00`, `01`, `10`, `11`, `12`, `21`, and `22` where the first digit represents the number of variant alleles in the mother's genotype and the second digit represents the number of variant alleles in the offspring's genotype (see Table 30.1). The user can define a specific model for their data by listing predictors, variance components, and parameter constraints on the QMFG factor to test specific forms of MFG incompatibility.

There are two well-known examples of MFG incompatibility, RHD incompatibility and non-inherited maternal antigens (NIMA) (described in section 26.2; also see Table 30.1). Although these two examples are often framed in terms of disease, they could be analyzed

using quantitative phenotypes. For instance, RHD incompatibility, can lead to hemolytic disease of the newborn (Levine et al., 1941), which is associated with high levels of bilirubin resulting from the breakdown of the fetus's red blood cells (Lee et al., 2009). Table 30.1 shows how the RHD incompatibility parameter corresponds to the QMFG parameters. For this example, reference allele `A` corresponds to the antigen coding allele (often coded as `D`) and allele `B` corresponds to the null allele (often coded as `d`). The model therefore includes a parameter for the `B/B - A/B` genotype combination between mother and offspring. The other six maternal-offspring gene combinations make up the reference group. For the second prototypical example, at HLA-DRB1, NIMA appears to increase the risk of rheumatoid arthritis (RA) in offspring. As an example of an associated quantitative trait, anti-CCP antibodies are markers for diagnosis and prognosis for RA (Visser et al., 2002, Silveira et al., 2007). In this case, there is an MFG parameter when the mother has one variant allele and the offspring has none (the `A/B - A/A` genotype combination). As shown in Table 30.1, two offspring genotype effect parameters are also included in the model: one for when the offspring is heterozygous and another for when the offspring is homozygous for the variant allele, regardless of the mother's genotype. The sample input files demonstrate how models consistent with these two scenarios of MFG incompatibility could be fit using Option 30.

When null model is no genetic effect at the locus, it should be run by including the QMFG predictor but setting all QMFG parameters equal to one another (see the control file, Control30b.in, as an example). By using Option 30 to run the null model, rather than some other option or program, ensures the same offspring are used when calculating the loglikelihoods for both the null and alternative models. To compare two sets of nested restrictions, for example NIMA and offspring effects compared with offspring effects alone, the user must run Mendel twice, once under the alternative hypothesis and once under the null model. The reported loglikelihoods can then be used to calculate the LRT and p-value.

30.3 Input Files

Option 30 uses standard definition, map, and pedigree files. The control file, Control30a.in, illustrates the basic commands needed to run QMFG LRT:

```
! Input and Output Files
!
MAP_FILE = Map30a.in
DEFINITION_FILE = Def30a.in
PEDIGREE_FILE = Ped30a.in
SUMMARY_FILE = Summary30a.out
OUTPUT_FILE = Mendel30a.out
ECHO = SOME
```

```
!
! Analysis Parameters
!
ANALYSIS_OPTION = QMFG_LRT
QMFG_LOCUS = 1
QUANTITATIVE_TRAIT = Var1
PREDICTOR = QMFG :: Var1
PREDICTOR = Sex :: Var1
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
!
! QMFG Model Equations
!
PARAMETER_EQUATION = PAR02 - PAR03 :: 0
PARAMETER_EQUATION = PAR03 - PAR04 :: 0
PARAMETER_EQUATION = PAR04 - PAR05 :: 0
PARAMETER_EQUATION = PAR05 - PAR06 :: 0
PARAMETER_EQUATION = PAR06 - PAR08 :: 0
```

The keyword `QMFG_LOCUS` allows the user to select which SNP in the pedigree file to test for an MFG effect. Here `QMFG_LOCUS = 1` tells Mendel to create a QMFG factor based on the first SNP in the pedigree, Marker1. The QMFG factor created based on the specified locus is internally inserted into the first factor of the pedigree file and thus there must be a field in the pedigree file for this purpose. Using the `PREDICTOR` keyword in the Control30a.in, QMFG and Sex are included as variables in the model. Variance components, such as additive and environmental variances, are included in the model using the `COVARIANCE_CLASS` keyword. The list of possible covariance classes can be found in Table 19.1. By imposing parameter constraints via the `PARAMETER_EQUATION` keyword (see section 0.10.2 for syntax details), specific MFG incompatibility scenarios can be tested.

Because QMFG is the first predictor put into the model in Control30a.in, the seven parameters of the QMFG factor are numbered 2–8 (parameter 1 is automatically the grand mean). These parameter values are used when specifying restrictions in the parameter values using `PARAMETER_EQUATION`. The numbers corresponding to the QMFG parameters may change depending on which order QMFG and other covariates listed in the control file. For example, if sex had been listed first in the control file then the QMFG factors would start with `PAR03`. This factor must also be defined in the definition file.

In the definition file, Def30a.in, this factor and its seven levels are specified:

```
QMFG, Factor, 7
00
01
10
```

```
11
12
21
22
```

Here, the name QMFG is used, but this factor can be given any name by the user as long as the exact same name is used in the control and definition files.

30.4 Examples

All seven examples for Option 30 use the same set of simulated data with three unlinked markers (Marker1, Marker2, Marker3) and three variables (Var1, Var2, Var3). In the control file, Control30a.in, `ANALYSIS_OPTION = QMFG_LRT` fits a variance components model, calculates parameter estimates, and outputs the loglikelihood. Setting `QMFG_LOCUS = 1` tells Mendel to internally put the maternal-offspring genotype combinations for Marker1 into the QMFG factor as `00, 01, 10, 11, 12, 21,` or `22`. Control30a.in describes an example of fitting a model consistent with RHD incompatibility on the first quantitative trait, Var1. The `PARAMETER_EQUATION` commands

```
!
! QMFG Model Equations
!
PARAMETER_EQUATION = PAR02 - PAR03 :: 0
PARAMETER_EQUATION = PAR03 - PAR04 :: 0
PARAMETER_EQUATION = PAR04 - PAR05 :: 0
PARAMETER_EQUATION = PAR05 - PAR06 :: 0
PARAMETER_EQUATION = PAR06 - PAR08 :: 0
```

declare that QMFG parameters 2, 3, 4, 5, 6, and 8 (corresponding to QMFG values 00, 01, 10, 11, 12, and 22) are set equal to one another and that the RHD incompatibility effect, i.e., `QMFG:21`, will be estimated separately from the other six mother-offspring genotype combinations.

Summary30a.out contains the output

```
                        POLYGENIC-QTL OPTION

                    SUMMARY FOR MEAN PARAMETERS

 TRAIT              PARAMETER    PREDICTOR          ESTIMATE        STD ERR

 Var1                     1      GRAND              40.0177         0.0580
 Var1                     2      QMFG:00            -0.0755         0.0261
```

Var1	3	QMFG:01	-0.0755	0.0261
Var1	4	QMFG:10	-0.0755	0.0261
Var1	5	QMFG:11	-0.0755	0.0261
Var1	6	QMFG:12	-0.0755	0.0261
Var1	7	QMFG:21	0.4528	0.1563
Var1	8	QMFG:22	-0.0755	0.0261
Var1	9	FEMALE	-0.0910	0.0533
Var1	10	MALE	0.0910	0.0533

SUMMARY FOR VARIANCE COMPONENTS

TRAIT	PARAMETER	VARIANCE	ESTIMATE	STD ERR
Var1	11	ADDITIVE	0.7966	0.2971
Var1	12	ENVIRONMENTAL	4.9482	0.3198

TRAIT	TOTAL VARIANCE
Var1	5.7447

TRAIT	HERITABILITY	STD ERR
Var1	0.1387	0.0449

For this example, RHD incompatibility is estimated to increase an offspring's phenotype by 0.5283 units (0.4528 – (-0.0755)) compared to any other mother-offspring genotype combination. The maximum loglikelihood for this model given the Mendel30a.out file is -2744.18.

To calculate the likelihood ratio test statistic with a null model of no genetic effect, the null model is run on the same offspring. In this case this is accomplished by adding an additional constraint, `PARAMETER_EQUATION = PAR06 - PAR07 :: 0`, thus keeping all QMFG parameters equal. This was done in control file Control30b.in. Mendel30b.out shows the calculated loglikelihood for the null model of no QMFG effect is -2748.37. The likelihood ratio test statistic is twice the difference of the loglikelihoods (in this case 2(-2744.18 – (-2748.37)) = 8.37) so the one degree-of-freedom p-value for an RHD effect is 0.0038.

The Control30c.in file fits a model consistent with NIMA and offspring genotypic effects for Marker2 to the trait Var2 using the following model equations:

```
! QMFG Model Equations
!
PARAMETER_EQUATION = PAR03 - PAR05 :: 0
```

```
PARAMETER_EQUATION = PAR05 - PAR07 :: 0
PARAMETER_EQUATION = PAR06 - PAR08 :: 0
```

Here, the effects of parameter 2 (QMFG:00) and the parameter 4 (QMFG:10) are not constrained and are therefore estimated separately. Parameters 3, 5, and 7 correspond to QMFG levels 01, 11, and 21. These three levels capture the mother-offspring genotype combinations in which the offspring is heterozygous regardless of the mother's genotype. The parameter equations above set these three parameters equal to each other, thus estimating the effect of one variant allele in the offspring's genotype. The effect of offspring homozygous for the variant allele is captured by setting parameters 6 and 8 equal (i.e., QMFG:12 and QMFG:22 are constrained to be the same). Thus parameters 6 and 8 estimate the effect of offspring with two copies of the variant allele in their genotype. The parameter estimates for this model are seen in Summary30c.out

POLYGENIC-QTL OPTION

SUMMARY FOR MEAN PARAMETERS

TRAIT	PARAMETER	PREDICTOR	ESTIMATE	STD ERR
Var2	1	GRAND	40.1340	0.0313
Var2	2	QMFG:00	-0.1257	0.0483
Var2	3	QMFG:01	0.0184	0.0216
Var2	4	QMFG:10	0.0587	0.0444
Var2	5	QMFG:11	0.0184	0.0216
Var2	6	QMFG:12	0.0059	0.0389
Var2	7	QMFG:21	0.0184	0.0216
Var2	8	QMFG:22	0.0059	0.0389
Var2	9	FEMALE	0.0299	0.0202
Var2	10	MALE	-0.0299	0.0202

SUMMARY FOR VARIANCE COMPONENTS

TRAIT	PARAMETER	VARIANCE	ESTIMATE	STD ERR
Var2	11	ADDITIVE	0.9348	0.0644
Var2	12	ENVIRONMENTAL	0.0140	0.0384

TRAIT	TOTAL VARIANCE
Var2	0.9488

TRAIT	HERITABILITY	STD ERR

```
 Var2                  0.9853       0.0266
```

Thus, one copy of the variant allele B increases an offspring's phenotype on average by 0.1441 units (0.0184 – (-0.1257)) compared to offspring with no copies of the variant allele. Two copies of the variant allele is estimated to increase an offspring's phenotype by 0.1316 units (0.0059 – (-0.1257)) compared to offspring with no copies of the variant allele. Comparing the estimate for parameter 4 to the estimate for parameter 2, the NIMA effect is estimated to be 0.1844 units (0.0587 – (-0.1257)). This indicates that even if the offspring does not have a copy of the variant allele in his own genotype, if his mother does, his phenotype may be increased. The estimates are shown in Table 30.2.

Table 30.2: Estimates for NIMA and Offspring Genotypic Effect Model

Effect	Estimate (SE)	P-value
Grand mean	39.9784 (0.0610)	<0.0001
A/B offspring	0.1441 (0.0529)	0.006
B/B offspring	0.1316 (0.0620)	0.03
NIMA	0.1844 (0.0656)	0.005
Female	0.0598 (0.0286)	0.04

If the user instead wants to test for a NIMA effect in the presence of an additive offspring effect, the alternative and null models can be run using the parameter constraints provided in Control30d.in and Control30e.in, respectively. In Control30d.in, corresponding to our alternative model, the QMFG model equations are specified as

```
! QMFG Model Equations
!
PARAMETER_EQUATION = PAR03 - PAR05 :: 0
PARAMETER_EQUATION = PAR05 - PAR07 :: 0
PARAMETER_EQUATION = PAR06 - PAR08 :: 0
PARAMETER_EQUATION = PAR06 - 2*PAR03 + PAR02 :: 0
```

There are various ways in which the additive offspring effect can be defined. It is important to note that because QMFG:00 is the reference group, to estimate an additive offspring effect the parameter equations must constrain the difference of PAR06 and PAR02 to be two times the difference of PAR03 and PAR02. In this example, we do this using the command

```
PARAMETER_EQUATION = PAR06 - 2*PAR03 + PAR02 :: 0
```

For this model, the result of the loglikelihood calculation is shown in Mendel30d.out to be -740.61. The control file Control30e.in corresponds to the null hypothesis of no NIMA effect given an additive offspring effect. To fit this model the constraint

```
PARAMETER_EQUATION = PAR02 - PAR04 :: 0
```

is needed in addition to the constraints in Control30d.in to ensure that the estimate for the NIMA parameter is equal to that of the reference group. The loglikelihood for this model is seen in Mendel30e.out to be -742.98, giving a likelihood ratio test statistic of 4.74. A p-value of 0.029 for this one degree-of-freedom test suggests that there is a significant NIMA effect in the presence of additive offspring effects.

To further demonstrate the flexibility of this option, we compare the fully generalized model to a model that accounts for maternal genetic effects. In the control file for the general model, Control30f.in, no constraints on the QMFG parameters are included. We want to test the significance of the six-parameter general model compared to a two-parameter maternal effect model. The control file Control30g.in contains the parameter equations for the maternal effect model

```
! QMFG Model Equations
!
PARAMETER_EQUATION = PAR02 - PAR03 :: 0
PARAMETER_EQUATION = PAR04 - PAR05 :: 0
PARAMETER_EQUATION = PAR05 - PAR06 :: 0
PARAMETER_EQUATION = PAR07 - PAR08 :: 0
```

Constraining parameters 2 and 3 to be equal forces the reference group to consist of offspring whose mothers are homozygous for the reference allele (`QMFG:00` and `QMFG:01`). To estimate the effect of a heterozygous mother compared to the reference group, parameters 4, 5, and 6 (corresponding to QMFG factors 01, 11, and 12) are set equal. The effect of a mother with two variant alleles is estimated separately, by constraining the difference between parameter 7 and 8 to be zero, indicating that the effects of `QMFG:21` and `QMFG:22` are equal. The loglikelihood for the fully generalized model is -746.26 (from Mendel30f.out) and the loglikelihood for the null model of a maternal genotypic effect is -756.80 (from Mendel30g.out). The test statistic for this three degrees-of-freedom LRT (21.08) has a p-value of 0.0001, suggesting that the fully generalized model is a better fit for these data than the maternal effects model.

30.5 Germane Keywords

```
ANALYSIS_OPTION = QMFG_LRT
COVARIANCE_CLASS
```

```
COVARIANCE_FACTORS
DEVIANCES
GRID_INCREMENT
GROUP_FACTOR
MAX_ADJUSTED_MEIOSES
MULTIVARIATE_NORMAL
OUTLIERS
PEDIGREE_CUT_POINT
PERSON_CUT_POINT
PREDICTOR
PROBAND
PROBAND_FACTOR
QMFG_LOCUS
QUANTITATIVE_TRAIT
TRANSFORM
```

Random Quotes

To the congressman who declared his preference to be right rather than President: The gentleman need not be disturbed; he will never be either.

Thomas Reed, Speaker of the US House of Representatives in the 1890's

Everything That Rises Must Converge
A Good Man is Hard to Find

two book titles by *Flannery O'Connor*

I married beneath me; all women do.

Mary Astor

Where there's marriage without love, there will be love without marriage.

Benjamin Franklin

We think about sex obsessively except during the act, when our minds tend to wander.

Howard Nemerov

The second night in Talkingham, Hazel Motes walked along down town close to the store fronts but not looking in them. The black sky was underpinned with long silver streaks that looked like scaffolding and depth on depth behind it were thousands of stars that all seemed to be moving very slowly as if they were about some vast construction work that involved the whole order of the universe and would take all of time to complete. No one was paying any attention to the sky.

Flannery O'Connor in *Wise Blood*

31 Analysis Option 31: QMFG Score

31.1 Background

As described above in Analysis Option 30 (QMFG LRT), joint maternal and offspring genetic effects, including interactions, can impact offspring disease susceptibility and their associated traits. These effects on quantitative traits can be modeled using the quantitative trait maternal-fetal genotype (QMFG) test (Clark et al., 2015). Analysis Option 31 uses the variance components model used in other options of Mendel (Options 19, 20, 29, and 30) where the genotypes of the offspring, mother, and their interactions are included as fixed effects (predictors) and residual correlations are taken into account via variance components (covariance classes). The QMFG test can use nuclear and extended pedigrees simultaneously and can incorporate additional predictors into the analysis. Unlike Option 30, which requires iterative maximization of the likelihood for each SNP to calculate parameter estimates, Option 31 runs the QMFG score test. The QMFG score test precomputes key quantities once during the analysis of the null model and then uses those values in the analysis of each SNP. This makes the calculation of the score test statistic for each SNP very rapid. Thus, Mendel's QMFG score test can be a valuable tool to quickly screen data sets for associations with MFG incompatibility, even dense, genome-wide data sets. Once markers are screened, Option 30 (QMFG LRT) can be used to refine results and estimate parameters for the most significant SNPs.

31.2 Appropriate Problems and Data Sets

Pedigrees of any shape or size are admissible. Because the QMFG test requires genotype data from mothers, quantitative trait data from founders (including singletons) are not used. Additionally, usable offspring are restricted to those who are genotyped, have a quantitative trait value, and have mothers who are also genotyped. If a mother or offspring is not genotyped, the entire SNP will be removed from the analysis, thus we recommend imputing missing genotypes. If a Mendelian error exists between a mother-offspring pair (an `A/A` mother with `B/B` offspring or vice versa), the SNP will not be used in the analysis. To include these SNPs in the analysis, these errors must be resolved, for example by using an imputation method that accounts for pedigree relationships or by excluding the offspring. Only biallelic loci will be analyzed. Analysis Option 16 (combining alleles) can be used to reduce multi-allelic loci to two alleles.

The QMFG score test can use either text-based files or the binary SNP data files as described in Section 0.6. The current option uses algorithms first introduced for Option 29, which implements a fast score test for GWAS on pedigree data with quantitative traits (Zhou et al., 2015). Here, the user-specified QMFG model is fit to all SNPs included in the pedi-

gree file unless there is a mother-offspring Mendelian error detected or a missing genotype. By default, the score test statistics, p-values, and additional information such as minor allele frequency for the top 20 most significant SNPs are output to the summary file. Unlike Option 29 however, the user must follow up Option 31 manually with Analysis Option 30 (QMFG LRT) to get parameter estimates for these top significant hits. The number of SNPs shown in the summary file can be changed using the keyword `DESIRED_PREDICTORS`. The plot file contains the score test statistics, p-values, and additional information for all SNPs analyzed. Like Option 26 (MFG), there are three models available to users, which are repeated here in Table 31.1. In the table, allele `A` represents the reference allele. For Option 31, the major allele is the default reference allele. Alternatively, adding the command `QMFG_REF_ALLELE = FIRST` to the control file will set the reference allele to the first allele listed in the definition file, or if not listed, the first with respect to lexicographic order.

Table 31.1: MFG Factor Levels for the QMFG Score Test

Mother	Offspring	Model 1: RHD incompatibility	Model 2: NIMA & offspring effects	Model 3: General model
A/A	A/A	Ref	Ref	Ref
A/A	A/B	Ref	Heter	U_{01}
A/B	A/A	Ref	NIMA	U_{10}
A/B	A/B	Ref	Heter	U_{11}
A/B	B/B	Ref	Homoz	U_{12}
B/B	A/B	RHD	Heter	U_{21}
B/B	B/B	Ref	Homoz	U_{22}

There are two model shortcuts included for the convenience of the user. These two sets of constraints correspond to the well-known examples of MFG incompatibility, RHD incompatibility and non-inherited maternal antigens (NIMA) (described in Sections 26.3 and 30.2). When `MODEL = 1` is set in the control file, Mendel calculates the score test statistic and p-value of RHD-type incompatibility effects. Here, allele `A`, the reference allele, corresponds to the antigen coding allele (often coded as `D`) and allele `B`, the variant allele, corresponds to the null allele (often coded as `d`). The model therefore includes a factor level indicating the `B/B - A/B` genotype combination between mother and offspring. The other six maternal-offspring gene combinations make up the reference factor level. The score test statistic in this case follows a chi-square distribution with one degree-of-freedom.

When `MODEL = 2`, Mendel fits a model consistent with NIMA and offspring genotypic effects. In this case there is a factor level that corresponds to the mother having one variant allele and the offspring having none (the `A/B - A/A` genotype combination). There

are also two offspring factor levels: one for when the offspring is heterozygous and another for when the offspring is homozygous for the variant allele, regardless of the mother's genotype. The score test statistic in this case follows a chi-square distribution with three degrees-of-freedom.

When `MODEL = 3`, Mendel performs the most general score test for maternal-offspring genotype effects. As seen in Table 31.1, there are seven factor levels, however there are at most six distinct factor levels because the `A/A - A/A` genotype combination is equivalent to no deviation from the mean. Here U_{ij} denotes the factor level when there are i variant alleles in the mother's genotype and j variant alleles in the offspring's genotype. This model is useful when the user has no prior hypothesis of mechanism of maternal-fetal genotype incompatibility. Under this general model, the score test statistic follows a chi-square distribution with six degrees-of-freedom. Model 3 is also useful if the user has a specific hypothesis for MFG incompatibility that differs from RHD (Model 1) or NIMA and offspring genotypic effects (Model 2) because the user can define linear relationships among the levels, $U_{01}, \ldots, U_{22}$, to regroup the genotype data using the `QMFG_LEVEL` keyword described below.

31.3 Input Files

The common keywords needed to run Option 31 are shown in the control file Control31a.in

```
! Input and Output Files
!
MAP_FILE = Map31a.in
DEFINITION_FILE = Def31a.in
PEDIGREE_FILE = Ped31a.in
SUMMARY_FILE = Summary31a.out
OUTPUT_FILE = Mendel31a.out
PLOT_FILE = Plot31a.out
ECHO = Some
!
! Analysis Parameters
!
ANALYSIS_OPTION = QMFG_SCORE
MODEL = 1
QUANTITATIVE_TRAIT = Var1
PREDICTOR = Sex :: Var1
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
DESIRED_PREDICTORS = 10 :: SCORE
```

The command ANALYSIS_OPTION = QMFG_SCORE instructs Mendel to perform the score test for all SNPs. The command MODEL = 1 indicates that the effect tested should be RHD incompatibility using the quantitative trait, Var1, that is specified by the keyword QUANTITATIVE_TRAIT. Additionally, sex is included in the model as a covariate using the command PREDICTOR = Sex :: Var1. The inclusion of the COVARIANCE_CLASS commands indicates that additive and environmental variances are included in the model. By default, the score test statistics, p-values, and additional information such as minor allele frequency for the top 20 most significant SNPs are output to the summary file. As shown in Control31a.in, the number of significant SNPs that appear in the summary file can be changed to a user specified value, for example to 10 SNPs, with a command such as DESIRED_PREDICTORS = 10 :: SCORE. However, note that, unlike Option 29 (PedGWAS), in Option 31 the only meaningful argument for this keyword is SCORE.

In addition to the above keywords, QMFG_LEVEL can be used to test for a specific scenario of MFG incompatibility outside of RHD incompatibility and NIMA with offspring effects. Unlike PARAMETER_EQUATION, QMFG_LEVEL partitions the design matrix of fixed effects. To test specific hypotheses regarding the data that are special cases of the default models, the user invokes the keyword QMFG_LEVEL. For example, in the control file Control31c.in, MODEL = 2 defines a test of a NIMA effect and genotypic offspring effects. Thus, the design matrix for that run has three QMFG columns that are 0/1 indicators for the absence/presence of the NIMA mother-offspring genotype combination (column 1), the heterozygous offspring genotype (column 2), and the homozygous variant offspring genotype (column 3). If the user instead desires to test a NIMA effect along with an additive offspring effect, then the number of levels needs to be reduced from three to two. The two levels can be specified using

```
! QMFG Level Constraints
!
QMFG_LEVEL = U_10 :: L1
QMFG_LEVEL = U_01 + U_11 + U_21 + 2*U_12 + 2*U_22 :: L2
```

Here, QMFG_LEVEL = U_10 :: L1 indicates that the first level (L1) represents that mother-offspring pairs with genotype combination A/B - A/A (coded as 1) have a different effect on the phenotype than any other mother-offspring genotype pair (coded as 0). The second expression indicates a second grouping, L2, which counts the number of B alleles in the offspring regardless of the mother's genotype. Here heterozygous offspring (U_{01}, U_{11}, and U_{21}) are counted once and homozygous offspring (U_{12} and U_{22}) are counted twice. Note that once the keyword QMFG_LEVEL is invoked, any genotype combinations that are not specified with a specific QMFG_LEVEL are grouped into the reference category. The reference category always includes offspring with no variant alleles who also have a mother with no variant alleles, that is, the A/A - A/A genotype combination. Thus, the second

column takes the value 0 for all other QMFG levels (U_{00} and U_{10}). Therefore it follows that the L2 expression defines a second column with entries that are equivalent to the number of variant alleles in each offspring's genotype (0, 1, and 2), alleles acting in an additive manner in the offspring.

In summary, for each of the offspring the possible entries in the design matrix are (0,0) corresponding to `A/A - A/A`, (1,0) corresponding to `A/B - A/A`, (0,1) corresponding to `A/A - A/B`, `A/B - A/B` or `B/B - A/B`, and (0,2) corresponding to `A/B - B/B` or `B/B - B/B`. For example, the QMFG portion of the design matrix for the offspring in the pedigree

```
2:1     ,  100    ,          ,          ,  M      ,          ,A/A
2:1     ,  101    ,          ,          ,  F      ,          ,A/B
2:1     ,  200    ,          ,          ,  F      ,          ,A/B
2:1     ,  203    ,          ,          ,  M      ,          ,B/B
2:1     ,  201    ,  100     ,  101     ,  M      ,          ,A/A
2:1     ,  202    ,  100     ,  101     ,  F      ,          ,A/B
2:1     ,  300    ,  201     ,  200     ,  M      ,          ,A/A
2:1     ,  301    ,  203     ,  202     ,  F      ,          ,B/B
```

has the form $\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 2 \end{pmatrix}$ where column 1 corresponds to `L1`, column 2 corresponds to `L2`, and the rows correspond to offspring 201, 202, 300, and 301, respectively.

31.4 Examples

In this option a score test with the named quantitative trait is run for each SNP in the data set. In our first example, the control file Control31a.in shown above tests whether there is an RHD effect on trait Var1 with Marker1, Marker2, and Marker3. The first part of Summary31a.out gives the results for the null model with no MFG effects

```
   QUANTITATIVE MATERNAL-FETAL GENOTYPE INCOMPATIBILITY SCORE TEST ANALYSIS OPTION


        EFFECT ESTIMATES UNDER NULL MODEL WITH NO SNPS INCLUDED

                     SUMMARY FOR MEAN PARAMETERS

 TRAIT              PARAMETER    PREDICTOR          ESTIMATE        STD ERR

 Var1                   1        GRAND               39.9933         0.0574
 Var1                   2        FEMALE              -0.0893         0.0535
```

```
 Var1                 3        MALE                 0.0893         0.0535

                    SUMMARY FOR VARIANCE COMPONENTS

 TRAIT             PARAMETER    VARIANCE           ESTIMATE        STD ERR

 Var1                 4        ADDITIVE             0.7773         0.2981
 Var1                 5        ENVIRONMENTAL        4.9902         0.3217

 TRAIT             TOTAL VARIANCE

 Var1                 5.7676

 TRAIT             HERITABILITY    STD ERR

 Var1                 0.1348       0.0451
```

The second part of Summary31a.out then gives the results for the top SNPs

```
          QMFG SCORE TEST RESULTS FOR THE TOP SNPS

PREDICTOR     QMFG         QMFG          SCORE     REGRESSION     MINOR        HARDY-
  NAME      P-VALUE   -LOG10(P-VAL)   STATISTIC     ALLELE      ALLELE       WEINBERG
                                                                FREQUENCY     P-VALUE
                                                              IN FOUNDERS   IN FOUNDERS

Marker1     0.00385      2.41422      8.35192      B=Minor      0.41350       0.74094
Marker3     0.05190      1.28485      3.77905      B=Minor      0.39387       0.36950
Marker2     0.12841      0.89142      2.31168      B=Minor      0.41075       0.68845
```

For MODEL = 1, which tests for an effect of RHD incompatibility, the score test statistic is distributed as a chi-square with one degree-of-freedom. Here, Marker1 is found to have a significant RHD incompatibility effect on Var1 (using a significance cut off of 0.05).

Our second example tests for effects on Var2 consistent with NIMA and offspring genotypic effects by setting `MODEL = 2` in Control31b.in. The score test results from Summary31b.out show a significant effect of either NIMA or offspring genotype for Marker1 and Marker2 on Var2 (using significance cut off of 0.05).

```
          QMFG SCORE TEST RESULTS FOR THE TOP SNPS

PREDICTOR     QMFG         QMFG          SCORE     REGRESSION     MINOR        HARDY-
  NAME      P-VALUE   -LOG10(P-VAL)   STATISTIC     ALLELE      ALLELE       WEINBERG
                                                                FREQUENCY     P-VALUE
```

```
                                                          IN FOUNDERS  IN FOUNDERS

Marker2     0.03625     1.44067     8.52920     B=Minor     0.41075      0.68845
Marker3     0.45114     0.34569     2.63642     B=Minor     0.39387      0.36950
Marker1     0.83771     0.07690     0.84900     B=Minor     0.41350      0.74094
```

Our third example uses `MODEL = 3` and specifies factor levels using the `QMFG_LEVEL` keyword. This runs a specific MFG incompatibility scenario outside the basic RHD and NIMA models. The control file Control31c.in also specifies the use of a binary file to input the SNP genotype data

```
! Input and Output Files
!
DEFINITION_FILE = Def31c.in
PEDIGREE_FILE = Ped31c.in
SNP_DEFINITION_FILE = SNP_def31c.in
SNP_DATA_FILE = SNP_data31c.bin
SUMMARY_FILE = Summary31c.out
OUTPUT_FILE = Mendel31c.out
PLOT_FILE = Plot31c.out
ECHO = SOME
!
! Analysis Parameters
!
ANALYSIS_OPTION = QMFG_SCORE
MODEL = 3
QMFG_REF_ALLELE = FIRST
QUANTITATIVE_TRAIT = Var3
PREDICTOR = Sex :: Var3
COVARIANCE_CLASS = Additive
COVARIANCE_CLASS = Environmental
KINSHIP_SOURCE = Pedigree_Structure
!
! QMFG Level Constraints
!
QMFG_LEVEL = U_10 :: L1
QMFG_LEVEL = U_01 + U_11 + U_21 + 2*U_12 + 2*U_22 :: L2
```

In contrast to the previous example in which no relationship was imposed between the effect of a heterozygous offspring and a homozygous offspring, here we hypothesize that the offspring effect on Var3 is additive. The discussed above these `QMFG_LEVEL` equations run the two degrees-of-freedom test for NIMA or additive offspring effects. For text-based files, the kinship coefficients are by default estimated based on the pedigree structures

in the input file, but for binary files the default kinship coefficient estimation uses SNP genotypes for all pairs of individuals within the same pedigree. This default is equivalent to KINSHIP_SOURCE = SNPs_within_pedigrees. More details on the method of kinship coefficient estimation is described in more detail in Section 29.2. Here, since we have so few SNPs in the example data set, we estimate kinship coefficients based on the pedigree structure by using the command KINSHIP_SOURCE = Pedigree_Structure in the control file. Also, we specify that the reference allele is to be the first allele listed in the definition file for each marker rather than the major allele by including the command QMFG_REF_ALLELE = FIRST in the control file. The results can be seen in Summary31c.out

QMFG SCORE TEST RESULTS FOR THE TOP SNPS

PREDICTOR NAME	QMFG P-VALUE	QMFG -LOG10(P-VAL)	SCORE STATISTIC	REGRESSION ALLELE	MINOR ALLELE FREQUENCY IN FOUNDERS	HARDY-WEINBERG P-VALUE IN FOUNDERS
Marker3	0.01398	1.85462	8.54084	A=Major	0.39387	0.36950
Marker2	0.16219	0.78996	3.63792	B=Minor	0.41075	0.68845
Marker1	0.95279	0.02100	0.09671	B=Minor	0.41350	0.74094

There is a significant NIMA or additive offspring effect on Var3 for Marker3. Data set 31c demonstrates the flexibility of this option to test numerous joint offspring and maternal effects of interest using the score test.

31.5 Germane Keywords

ANALYSIS_OPTION = QMFG_SCORE
COVARIANCE_CLASS
DESIRED_PREDICTORS
KINSHIP_METHOD
KINSHIP_SOURCE
MIN_MINOR_ALLELE_COUNT
MIN_SAMPLED_SNPS
MIN_SUCCESS_RATE_PER_INDIVIDUAL
MIN_SUCCESS_RATE_PER_SNP
MODEL
MULTIVARIATE_NORMAL
OUTLIERS
PEDIGREE_CUT_POINT
PERSON_CUT_POINT
PLOT_FILE

```
PREDICTOR
QMFG_LEVEL
QMFG_REF_ALLELE
QUANTITATIVE_TRAIT
SNP_SAMPLING_INCREMENT
```

Random Quotes

I love everything that's old: old friends, old times, old manners, old books, old wine; and I believe, Dorothy, (taking her hand) you'll own that I have been pretty fond of an old wife.

Oliver Goldsmith in *She Stoops to Conquer*

I saw the angel in the marble and carved until I set him free.

Michelangelo

It is quite a three pipe problem, and I beg that you won't speak to me for fifty minutes.

Sherlock Holmes in *The Red-Headed League* by *Arthur Conan Doyle*

I returned and saw under the sun, that the race is not always to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favor to men of skill; but time and chance happeneth to them all.

Ecclesiastes

Objective consideration of contemporary phenomena compels the conclusion that success or failure in competitive activities exhibits no tendency to be commensurate with innate capacity, but that a considerable element of the unpredictable must invariably be taken into account.

George Orwell's parody of the Ecclesiastes quote

Table 32.1: **Mendel's Keywords**

Keyword	Default Value	Type	Short Description	Multi-valued
AFFECTED	2 & AFFECTED	10 Characters	Disease phenotype designator	no
AFFECTED_LOCUS_OR_FACTOR		16 Characters	Factor giving disease status	no
ALL_MAP_ORDERS	True	Logical	Either all or only input order considered	no
ALLELE_COMBINING_THRESHOLD	0.05	Real	Frequency threshold for merging alleles	no
ALLELE_COUNT_FILE		1024 Characters	External expected allele counts file	no
ALLELE_SEPARATOR	/ and \	1 Character	Separates codominant alleles	no
ANALYSIS_OPTION		50 Characters	See Table 0.1	no
BASELINE_CUMULATIVE_HAZARD		50 Characters	Name of variable holding that value	no
BASELINE_HAZARD		50 Characters	Name of variable holding that value	no
BASE_PREDICTORS		16 Characters	Name of base SNP for interaction analysis	yes
BASE_SET		Integer :: String	Size and type of base set for SNP interactions	no
CANDIDATE_ORDERS	10	Integer	Number of marker orders to consider	no
CASE2_CONTROL1	False	Logical	Case and Control labels are 2 and 1	no
CENSORING_VARIABLE		8 Characters	Censoring indicator	no
COEFFICIENT_FILE		1024 Characters	Conditional kinship coefficient file	no
COMPLEMENTARY_TRANSMISSION	False	Logical	Should transmission to unaffecteds be included?	no
COMPLETE_CONVERSION	False	Logical	Should all loci be converted during file conversion?	no
COMPLEXITY_THRESHOLD	5.0E7	Real	Criterion for not analyzing pedigrees	no
CONVERGENCE_CRITERION	0.0001	Real	Convergence criterion	no
CONVERGENCE_TESTS	4	Integer	Times convergence criterion to be met	no
COVARIANCE_CLASS		20 Characters	Covariance class for quantitative traits	yes
COVARIANCE_FACTORS		Integer :: String	# of factors for a covariance class	yes
DEFAULT_LIST_READ	True	Logical	Default list read settings	no
DEFAULT_PENETRANCE	0.0	Real	Default penetrance value	no
DEFINITION_FILE		1024 Characters	Name of definition file	no
DEFINITION_LIST_READ	True	Logical	For list directed read of definition file	no
DELETE_PROBAND_FIELD		16 Characters	Field to delete in probands	yes
DESIRED_PREDICTORS		Integer :: String	Number of predictors of various types in Summary file	yes
DEVIANCE_VARIABLE		8 Characters	Variable ranking deviances	no
DEVIANCES	False	Logical	Compute pedigree deviances	no

continued on next page

Table 32.1: **Mendel's Keywords Continued**

Keyword	Default Value	Type	Short Description	Multi-valued
ECHO	No	8 Characters	Controls echo of input to output: No, Partial, or Yes	no
EQUILIBRIUM_FREQUENCIES	False	Logical	Put equilibrium freq in new definition file	no
ESTIMATE_LINKED_PROPORTION	False	Logical	Estimate proportion in location scores	no
ESTIMATION_THRESHOLD	0.00001	Real	P-value threshold to force regression estimate	no
FEMALE	2 and F	6 Characters	Female designator	no
FEMALE_MUTATION_RATE	0.0	Real :: String	Female mutation rate in risk prediction	yes
FLANKING_DISTANCE	0.5	Real	Flanking distance for location scores	no
FLANKING_DISTANCE_FEMALE	0.5	Real	Flanking female distance for location scores	no
FLANKING_DISTANCE_MALE	0.5	Real	Flanking male distance for location scores	no
FLANKING_POINTS	3	Integer	Flanking analysis points for location scores	no
FLANKING_SNPS	7	Integer	Number of flanking SNPs in imputation window	no
GENDER_NEUTRAL	True	Logical	Forces equal recombination fractions	no
GENE_DROP_OUTPUT	Unordered	16 Characters	Output format style for gene drop pedigrees	no
GENE_FLOW_CUTOFF	1.0E-10	Real	For deleting unlikely descent graphs	no
GENOTYPING_ERROR_RATE	0.01	Real	Probability of genotyping error	no
GENOTYPING_ERROR_THRESHOLD	0.25	Real	Threshold for noting genotyping errors	no
GRID_INCREMENT	0.0	Real	Spacing between map points	no
GROUP_FACTOR		16 Characters	Factor defining groups of people	no
INDICATOR_THRESHOLD	0.0	Real	Threshold to convert variable to indicator	no
INPUT_FORMAT		16 Characters	Input format type for list-directed files	no
INTERACTION_LEVELS	2	Integer	Max. interacting SNPs in SNP Association	no
INTERIOR_POINTS	0	Integer	Points between adjacent markers	no
IMPUTATION_METHOD	HPP	32 Characters	HPP or BHP: Method used in SNP imputation	no
KEEP_FOUNDER_GENOTYPES	False	Logical	Retain founder genotypes during gene dropping	no
KEEP_HIDDEN_FILES	False	Logical	Retain hidden pedigree files	no
KINSHIP_METHOD	GRM	32 Characters	GRM or MOM: Method used to estimate kinship	no
KINSHIP_SOURCE		32 Characters	Source of data to estimate kinship	no
KINSHIP_THRESHOLD	0.05	Real	Kinship threshold for inclusion in pedigree	no
LASSO_PENALTY	False	Logical	Indicates use of the LASSO penalty function	no
LINKED_PROPORTION	1.0	Real	Proportion of linked pedigrees	no

continued on next page

Table 32.1: **Mendel's Keywords Continued**

Keyword	Default Value	Type	Short Description	Multi-valued
MAJOR_LOCUS		16 Characters	Name of locus used to simulate trait values	no
MALE	1 and M	6 Characters	Male designator	no
MALE_MUTATION_RATE	0.0	Real :: String	Male mutation rate in risk prediction	yes
MAP_CONVERSION	1.0	Real	Conversion rate for Mbp to cM map distances	no
MAP_CONVERSION_FEMALE	1.0	Real	Conversion rate for female map distances	no
MAP_CONVERSION_MALE	1.0	Real	Conversion rate for male map distances	no
MAP_DISTANCE_UNITS	RF	32 Characters	Units for reading map distances	no
MAP_FILE		1024 Characters	Name of map file	no
MAP_LIST_READ	True	Logical	For list directed read of map file	no
MARGINAL_ANALYSIS	True	Logical	Perform marginal SNP association analysis	no
MAX_ADJUSTED_MEIOSES	16	Integer	Max adjusted meioses per pedigree	no
MAX_COMBINED_ALLELES	10	Integer	Maximum alleles after allele merging	no
MAX_ITERATIONS	200	Integer	Maximum iterations per search	no
MAX_KINSHIP_PAIRS	25	Integer	Maximum extreme relative pairs	no
MAX_MAF	0.5	Real	Maximum minor allele frequency of SNPs to include	no
MAX_STEP_LENGTH	Infinity	Real	Maximum parameter update in search	no
MAX_STEPS	3	Integer	Maximum step halves per iteration	no
MAX_THREADS	0	Integer	Requested number of threads to use	no
MIN_MAF	0.5	Real	Minimum minor allele frequency of SNPs to include	no
MIN_MINOR_ALLELE_COUNT	3	Integer	Minimum minor allele count included in analysis	no
MIN_SAMPLED_SNPS	5000	Integer	Minimum number of SNPs in global kinship estimates	no
MIN_SUCCESS_RATE_PER_INDIVIDUAL	0.98	Real	Minimum allowed typing success rate per person	no
MIN_SUCCESS_RATE_PER_SNP	0.98	Real	Minimum allowed typing success rate per SNP	no
MISSING_AT_RANDOM	0.0	Real	Rate of randomly missing values for simulation	no
MISSING_DATA_PATTERN	Existing_Data	16 Characters	Output missing data pattern for simulation	no
MISSING_QUANTITATIVE_VALUE	–	4 Characters	String for missing quantitative values	no
MISSING_VALUE		4 Characters	String for missing non-quantitiative values	no
MODEL	1	Integer	Sub-option within an option	no
MULTIPLE_PLOT_FILES	False	Logical	Indicates if a plot file is made for each QT in ped-GWAS	no
MULTIVARIATE_ANALYSIS	False	Logical	Indicates type of analysis in ped-GWAS	no

continued on next page

Table 32.1: **Mendel's Keywords Continued**

Keyword	Default Value	Type	Short Description	Multi-valued
MULTIVARIATE_NORMAL	True	Logical	False for multivariate t	no
NEW_DEFINITION_FILE		1024 Characters	Mendel constructed definition file	no
NEW_MAP_FILE		1024 Characters	Mendel constructed map file	no
NEW_PEDIGREE_FILE		1024 Characters	Mendel constructed pedigree file	no
NEW_PENETRANCE_FILE		1024 Characters	Mendel constructed penetrance file	no
NUMBER_OF_MARKERS_INCLUDED	1	Integer	Size of sliding window of markers	no
ORDERED_ALLELE_SEPARATOR	|	1 Character	Separates ordered codominant alleles	no
OUTLIERS	False	Logical	Flag in standard output any outlier pedigrees and people	no
OUTPUT_FILE	Mendel.out	1024 Characters	Name of standard output file	no
PAIRWISE_ANALYSIS		10 Characters	Specifies pairwise interaction analysis	no
PARAMETER_EQUATION		String :: Real	Linear constraint on parameters	yes
PARAMETER_INITIAL_VALUE		Real :: String	Initial parameter value	yes
PARAMETER_MAX		Real :: String	Parameter upper bound	yes
PARAMETER_MIN		Real :: String	Parameter lower bound	yes
PEDIGREE_CUT_POINT	0.05	Real	Cutpoint for outlier pedigrees	no
PEDIGREE_FILE		1024 Characters	Name of pedigree file	no
PEDIGREE_LIST_READ	True	Logical	For list directed read of pedigree file	no
PEDIGREE_MAX_LINE_LEN	262,144	Integer	Maximum line length in pedigree file	no
PENALIZED_INTERACTION	False	Logical	Perform penalized regression interaction analysis	no
PENALIZED_REGRESSION	False	Logical	Perform penalized regression GWAS	no
PENETRANCE		Real :: String	Penetrance value for specified genotype	yes
PENETRANCE_FILE		1024 Characters	Name of penetrance file	no
PENETRANCE_LIST_READ	True	Logical	For list directed read of penetrance file	no
PENETRANCE_MODEL		String :: String	Specifies penetrance model	yes
PERSON_CUT_POINT	0.01	Real	Cutpoint for outlier people	no
PLOT_FILE	Plot.out	1024 Characters	Name of output plot file	no
POPULATIONS	1	Integer	Number of populations	no
POPULATION_FACTOR		16 Characters	Factor indicating population identity	no
POPULATION_PRIOR_COUNT	0.0	Real :: String	Dirichlet count for a population	yes
PREDICTOR		String :: String	Predictor for quantitative trait	yes

continued on next page

Table 32.1: **Mendel's Keywords Continued**

Keyword	Default Value	Type	Short Description	Multi-valued
PREDICTOR_PENALTY_PROPORTION	1.0	Real	Proportion of weight on individual vs group predictors	no
PRETRIM_PEDIGREES	True	Logical	Pedigree trimming during preprocessing	no
PRINCIPAL_COMPONENTS	0	Integer	Number of principal components to output	no
PROBABILITY_PEDIGREE_LINKED	False	Logical	For computing of linked proportion	no
PROBAND	PROBAND	10 Characters	Proband designator	no
PROBAND_FACTOR		16 Characters	Factor storing proband status	no
PSEUDO_ALLELES	0.0	Real	Number of prior alleles or haplotypes	no
QMFG_LOCUS		Integer	Number of the locus to use for QMFG_LRT	no
QMFG_LEVEL		String	Equations to use in QMFG_Score	yes
QMFG_REF_ALLELE	Major	String	Major or First: Reference allele in QMFG_Score	no
QUANTITATIVE_TRAIT		8 Characters	Name of a trait variable	yes
READ_PEDIGREE_COPIES	False	Logical	For reading of copies per pedigree	no
READ_PEDIGREE_RECORDS	False	Logical	Permits omission of pedigree records	no
REPETITIONS	0	Integer	Repetitions for sharing statistics	no
RESTORE_PROBAND_FIELD		16 Characters	Field to restore in probands	yes
RETAINED_GROUP		16 Characters	Group of predictors to retain in penalized analysis	yes
RETAINED_PREDICTOR		16 Characters	Predictor to retain in penalized analysis	yes
SAMPLES	10000	Integer	Samples in Monte Carlo sampling	no
SAMPLE_SUBSET_FILE		1024 Characters	Name of input sample subset file	no
SEED	0	Integer	Random number generator seed	no
SIMULATION_FILE		1024 Characters	Name of file containing trait simulation parameters	no
SNP_DATA_FILE		1024 Characters	Name of binary input SNP genotype file	no
SNP_DEFINITION_FILE		1024 Characters	Name of input SNP definition file	no
SNP_DOMINANCE_MODEL	Additive	16 Characters	Dominance model in SNP association tests	no
SNP_PHASE_FILE		1024 Characters	Name of binary input SNP phase file	no
SNP_SAMPLING_INCREMENT	100	Integer	Increment between sampled SNPs	no
SNP_SUBSET_FILE		1024 Characters	Name of input SNP subset file	no
SNPS_TYPED	EVERYONE	10 Characters	SNPs typed designator	no
SNPS_TYPED_FACTOR		16 Characters	Factor storing SNPs typed status	no
STANDARD_ERRORS	True	Logical	Set to true for parameter standard errors	no

continued on next page

Table 32.1: **Mendel's Keywords Continued**

Keyword	Default Value	Type	Short Description	Multi-valued
STATS_BY_SEX	False	Logical	Sex-specific descriptive statistics	no
STANDARD_GRID	False	Logical	Standard theta grid for LOD scores	no
SUMMARY_FILE	Summary.out	1024 Characters	Name of output summary file	no
TITLE		128 Characters	Title of run	no
TRANSFORM		String :: String	Quantitative trait transform	yes
TRAVEL	SEARCH	6 Characters	Mode of sampling parameter space	no
TUNING_CONSTANT		Real :: String	Tuning constants for SNP analysis	yes
UNIFORM_WEIGHTS	True	Logical	All predictors get default weight 1.0?	no
VERBOSE	True	Logical	Controls printing of messages to screen	no
ZOOM_SNP		16 Characters	Name of SNP at center of region to be output	no

References

[1] Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2001) Merlin — rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet* 30:97–101. 79

[2] Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19:1655–1664. 189

[3] Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Amer J Hum Genet* 62:1198–1211. 209

[4] Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *Amer J Hum Genet* 54:535–543. 209

[5] Ayers KL, Cordell HJ (2010) SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiology* 34:879–891. 240

[6] Ayers KL, Lange K (2008) Penalized estimation of haplotype frequencies. *Bioinformatics* 24:1596–1602. 234

[7] Bacanu S-A, Devlin B, Roeder K (2000) The power of genomic control. *Amer J Hum Genet* 66:1933–1944. 296

[8] Bauman L, Almasy L, Blangero J, Duggirala R, Sinsheimer JS, Lange K (2005) Fishing for pleiotropic QTLs in a polygenic sea. *Ann Hum Genet* 69:590–611. 214

[9] Bauman LE, Sinsheimer JS, Sobel EM, Lange K (2008) Mixed effects models for QTL mapping with inbred strains. *Genetics* 180:1743–1761. 271

[10] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B* 57:289–300. 252, 297

[11] Blangero J, Almasy L (1997) Multipoint oligogenic linkage analysis of quantitative traits. *Genet Epidemiology* 14:959–964. 209

[12] Boehnke M (1991) Allele frequency estimation from data on relatives. *Amer J Hum Genet* 48:22–25. 118

[13] Boerwinkle E, Chakraborty R, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann Hum Genet* 50:181–194. 219

[14] Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs. *Biometrika* 39:324–345. 132

[15] Bridge PJ (1997) *The Calculation of Genetic Risks: Worked Examples in DNA Diagnostics,* 2nd ed. Johns Hopkins University Press, Baltimore, MD. 128, 129

[16] Cantor RM, Chen GK, Pajukanta P, Lange K (2005) Association testing in a linked region using large pedigrees. *Amer J Hum Genet* 76:538–542. 230

[17] Cavalli-Sforza LL, Bodmer WF (1971) *The Genetics of Human Populations*. Freeman, San Francisco. 9, 138, 141, 189, 209

[18] Chen J, Zheng H, Wilson ML (2009) Likelihood ratio tests for maternal and fetal genetic effects on obstetric complications. *Genet Epidemiol* 33:526–538. 263

[19] Childs EJ, Palmer CGS, Lange K, Sinsheimer JS (2010) Modeling maternal-offspring gene-gene interactions: the Extended MFG Test. *Genet Epidemiol* 34:512–521. 262, 263

[20] Claerbout J, Muir F (1973) Robust modeling with erratic data. *Geophysics* 38:826–844. 240

[21] Clarke CA, Price-Evans DA, McConnell RB, Sheppard PM (1959) Secretion of blood group antigens and peptic ulcers. *Brit Med J* 1:603–607. 121

[22] Coyne JA (1976) Lack of similarity between two sibling species of Drosophila as revealed by varied techniques. *Genetics* 84:593–607. 126

[23] Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory*. Harper and Row, New York. 9, 123, 149, 209

[24] Daiger SP, Miller M, Chakraborty R (1984) Heritability of quantitative variation at the group-specific component (Gc) locus. *Amer J Hum Genet* 36:663–676. 223

[25] Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM (2011) Unifying ideas for non-parametric linkage analysis. *Human Heredity* 71:267–280. 107

[26] Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM (2011) Linkage analysis without defined pedigrees. *Genet Epidemiol* 35:360–370. 143, 144, 147, 290

[27] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genet* 29:229–232. 236

[28] Dobson AJ (2001) *An Introduction to Generalized Linear Models*, 2nd ed. Chapman & Hall, London. 173, 283

[29] Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. *Hum Hered* 21:523–542. 9

[30] Elston RC (1971) The estimation of admixture in racial hybrids. *Ann Hum Genet* 35:9–17. 189

[31] Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3:87–112. 121

[32] Ewens WJ (2004) *Mathematical Population Genetics, I: Theoretical Introduction*. Springer-Verlag, New York. 121

[33] Falk CT (1989) A simple scheme for preliminary orderings of multiple loci: applications to 45 CF families. *In* Multipoint mapping and linkage based upon affected pedigree members: *Genetic Analysis Workshop 6*. Liss, New York, pp 17–22. 86

[34] Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinb* 52:399–433. 209

[35] George VT, Elston RC (1987) Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet Epidemiology* 4:193–201. 219

[36] Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Ducat L, Spence SJ, AGRE Steering Committee (2001) The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Amer J Hum Genet* 69:463–466. 150, 171, 217

[37] Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. *Amer J Hum Genet* 47:957–967. 209

[38] Gonźalez JR, Wang W, Ballana E, Estivill X (2006) A recessive Mendelian model to predict carrier probabilities of DFNB1 for nonsyndromic deafness. *Human Mutation* 27:1135–1142. 130

[39] Haldane JBS (1935) The rate of spontaneous mutation of a human gene. *J Genet* 25:251–255. 129

[40] Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King M-C (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250:1684–1689. 110

[41] Harney S, Newton J, Milicic A, Brown MA, Wordsworth BP (2003) Non-inherited maternal HLA alleles are associated with rheumatoid arthritis. *Rheumatology* 42:171–174. 264

[42] Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M (2004) Ordered subset analysis in genetic linkage mapping of complex traits. *Genet Epidemiology* 27:53–63. 77

[43] Hodge SE, Vieland VJ, Greenberg DA (2002) HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity. *Amer J Hum Genet* 70:556–557. 93

[44] Hoffman GE, Logsdon BA, Mezey JG (2013) PUMA: A unified framework for penalized multiple regression analysis of GWAS data. *PLoS Computational Biology* 9:e1003101. 240

[45] Holloway SM, Smith C (1973) Equilibrium frequencies in X-linked recessive diseases. *Amer J Hum Genet* 25:388–396. 129

[46] Holt SB (1954) Genetics of dermal ridges: bilateral asymmetry in finger ridge-counts. *Ann Eugenics* 18:211–231. 212

[47] Hopper JL, Mathews JD (1982) Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* 46:373–383. 209, 219

[48] Hsieh HJ, Palmer CGS, Sinsheimer JS (2006) Allowing for missing data at highly polymorphic genes when testing for maternal, offspring, and maternal-fetal genotype incompatibility effects. *Hum Hered* 62:165–174. 262, 264

[49] Jacquier M, Arango D, Villareal E, Torres O, Serrano ML, Cruts M, Montañes P, Cano C, Rodriguez MN, Serneels S, Van Broeckhoven C (2001) APOE ϵ4 and Alzheimer's disease: Positive association in a Colombian clinical series and review of the Latin-American studies. *Arq Neuropsiquiatr* 59:11–17. 166

[50] Jin K, Speed TP, Klitz W, Thomson G (1994) Testing for segregation distortion in the HLA complex. *Biometrics* 50:1189–1198. 132

[51] Keavney B, McKenzie CA, Connell JMC, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, (1998) Measured haplotype analysis of the angiotensin-1 converting enzyme gene. *Hum Mol Gen* 11:1745–1751. 101, 136

[52] The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073. 150, 294

[53] Kraft P, Palmer CGS, Woodward JA, Turunen JA, Minassian S, Paunio T, Lonnqvist J, Peltonen L, Sinsheimer JS (2004) RHD Maternal-fetal genotype incompatibility and schizophrenia: extending the MFG test to include multiple siblings and birth order. *Euro J Hum Genet* 12:192–198. 262, 263

[54] Kraft P, Hsieh HJ, Cordell HJ, Sinsheimer JS (2005) A conditional-on-exchangeable-parental-genotypes likelihood that remains unbiased at the causal locus under multiple-affected-sibling ascertainment. *Genet Epidemiol* 29:87–90. 262

[55] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Amer J Hum Genet* 58:1347–1363. 107

[56] Kruglyak L, Lander ES (1998) Faster multipoint linkage analysis using Fourier transforms. *J Computational Biol* 5:1–7. 9, 103

[57] Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 84:2363–2367. 9, 103

[58] Lange EM, Lange K (2004) Powerful allele sharing statistics for nonparametric linkage analysis. *Hum Hered* 57:49–58. 107

[59] Lange K (2002) *Mathematical and Statistical Methods for Genetic Analysis,* 2nd ed. Springer-Verlag, New York. 9, 92, 107, 119, 144, 149, 157, 158, 161, 162, 209, 220, 262, 288

[60] Lange K, Boehnke M (l983) Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. *Amer J Med Genet* 14:513–524. 213

[61] Lange K, Boehnke M, Weeks DE (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet Epidemiology* 5:471–472. 9, 132

[62] Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E (2001) Mendel version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Amer J Hum Genet* 69(Supplement):504. 68

[63] Lange K, Elston RC (1975) Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Hum Hered* 25:95–105. 9

[64] Lange K, Goradia TM (1987) An algorithm for automatic genotype elimination. *Amer J Hum Genet* 40:250–256. 103, 113

[65] Lange K, Papp JC, Sinsheimer JS, Sobel EM (2014) Next-generation statistical genetics: Modeling, penalization, and optimization in high-dimensional data. *Annual Review of Statistics and Its Application* 1:279–300. 240

[66] Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM (2013) Mendel: The Swiss army knife of genetic analysis programs. *Bioinformatics* 29:1568–1570. 68

[67] Lange K, Sinsheimer JS (2004) The pedigree trimming problem. *Hum Hered* 58:108–11. 226

[68] Lange K, Sinsheimer JS, Sobel E (2005) Association testing with Mendel. *Genet Epidemiology* 29:36–50. 173, 219, 283

[69] Lange K, Sobel E (1991) A random walk method for computing genetic location scores. *Amer J Hum Genet* 49:1320–1334. 80, 103, 113, 144

[70] Lange K, Sobel E (2006) Variance component models for X-linked QTLs. *Genet Epidemiology* 30:380–383. 210

[71] Lange K, Westlake J, Spence MA (1976) Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann Hum Genet* 39:485–491. 209

[72] Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci* 81:3443–3446. 92

[73] Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* 48:67–81. 158, 168

[74] Lewis M, Kaita H, Philipps S, Giblet ER, Anderson JE, McAlpine PJ, Nickel B (1980) The position of the Radin blood group in relation to other chromosome 1 loci. *Ann Hum Genet* 44:179–184. 87, 136

[75] Litt M, Kramer P, Browne D, Gancher S, Brunt ERP, Root D, Phromchotikul T, Dubay CJ, Nutt J (1994) A gene for Episodic Ataxia/Myokymia maps to chromosome 12p13. *Amer J Hum Genet* 55:702–709. 99

[76] Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Amer J Hum Genet* 56:799–810. 234

[77] Markianos K, Daly MJ, Kruglyak L (2001) Efficient multipoint linkage analysis through reduction in inheritance space. *Amer J Hum Genet* 68:963–977. 79

[78] McCullagh P, Nelder JA (1989) *Generalized Linear Models*, 2nd ed. Chapman and Hall, London. 173, 283

[79] McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12:153–157. 168

[80] McKeigue PM (2005) Prospects for admixture mapping of complex traits. *Amer J Hum Genet* 76:1–7. 189

[81] Monaco AP, Bertelson CJ, Middlesworth W, Colletti CA, Aldridge J, Fishbeck KH, Bartlett R, Pericak-Vance MA, Roses AD, Kunkel LM (1985) Detection of deletions spanning the Duchenne muscular dystrophy locus using a tightly linked DNA segment. *Nature* 316:842–845. 129

[82] Morris AP, Curnow RN, Whittaker JC (1997) Randomization tests of disease marker associations. *Ann Hum Genet* 61:49–60. 169

[83] Murphy EA, Chase GA (1975) *Principles of Genetic Counseling*. Year Book Medical Publishers, Chicago. 9, 128, 129

[84] Newton JL, Harney SM, Wordsworth BP, Brown MA (2004) A review of the MHC genetics of rheumatoid arthritis. *Genes Immun* 5:151–157. 264

[85] Oehlmann R, Zlotogora J, Wenger DA, Knowlton RG (1993) Localization of the Krabbe disease gene (GALC) on chromosome 14 by multipoint linkage analysis. *Amer J Hum Genet* 53:1250–1255. 104

[86] Ott J (1999) *Analysis of Human Genetic Linkage*, 3rd ed. Johns Hopkins University Press, Baltimore, MD. 9, 262

[87] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge. 86

[88] Roberts DF, Hiorns RW (1965) Methods of analysis of the genetic composition of a hybrid population. *Hum Biol* 37:38–43. 189

[89] Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *Amer J Hum Genet* 73:1402–1422. 189, 190, 194

[90] Schork NJ (1993) Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Amer J Hum Genet* 53:1306–1319. 209

[91] Schrott HG, Goldstein JL, Hazzard WR, McGoodwin MM, Motulsky AG (1972) Familial hypercholesterolemia in a large kindred: Evidence for a monogenic mechanism. *Ann Int Med* 76:711–720. 182

[92] Sham PC (1997) *Statistics in Human Genetics*. Oxford University Press, Oxford. 9

[93] Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-allelic marker loci. *Ann Hum Genet* 59:323–336. 132

[94] Sinsheimer JS, Blangero J, Lange K (2000) Gamete competition models. *Amer J Hum Genet* 66:1168–1172. 132

[95] Sinsheimer JS, McKenzie CA, Keavney B, Lange K (2001) SNPs and snails and puppy dogs' tails: analysis of SNP data using the gamete competition model. *Ann Hum Genet* 65:483–490. 132

[96] Sinsheimer JS, Palmer CGS, Woodward JA (2003) Detecting genotype combinations that increase risk for disease: the maternal-fetal genotype incompatibility test. *Genet Epidemiol* 24:1–13. 262, 263

[97] Sinsheimer JS, Plaisier CL, Huertas-Vazquez A, Aguilar-Salinas C, Tusie-Luna T, Pajukanta P, Lange K (2008) Estimating ethnic admixture from pedigree data. *Amer J Hum Genet* 82:748–755. 189

[98] Sinsheimer J, Elston R, Fu WJ (2010) Gene-gene interaction in maternal and perinatal research. *J Biomed Biotechnol* 2010:article 853612. 262

[99] Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Amer J Hum Genet* 58:1323–1337. 80, 103, 107, 113, 144

[100] Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *Amer J Hum Genet* 70:496–508. 113, 138

[101] Sobel E, Sengul H, Weeks DE (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum Hered* 52:121–131. 144

[102] Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Amer J Hum Genet* 52:506–516. 168

[103] Strachen T, Read AP (1999) *Human molecular genetics*, 2nd ed. John Wiley, New York. 263

[104] Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiology* 28:289–301. 190

[105] Terwilliger JD, Ott J (1992) A haplotype based "haplotype relative risk" approach to detecting allelic associations. *Hum Hered* 42:337–346. 168

[106] van der Horst-Bruinsma IE, Haxes JM, Schreuder GM, Radstake TR, Barrera P, van de Putte LB, Mustamu D, van Schaardenburg D, Breedveld FC, de Vries RR (1998) Influence of non-inherited maternal HLA-DR antigens on susceptibility to rheumatoid arthritis. *Ann Rheum Dis* 57:672–675. 264

[107] Danecek P, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

[108] Watterson GA (1978) The homozygosity test of neutrality. *Genetics* 88:405–417. 121

[109] Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 62:969–978. 262

[110] Weir BS (1996) *Genetic Data Analysis II*. Sinauer, Sunderland, MA. 9, 123, 142

[111] White R (1986) The search for the cystic fibrosis gene. *Science* 234:1054–1055. 129

[112] Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. *Biometrics* 50:118–127. 107

[113] Whittemore AS, Halpern J (2002) Reply to Hodge et al. *Amer J Hum Genet* 70:558–559. 93

[114] Wu B, Liu N, Zhao H (2006) PSMIX: an R package for population structure inference via maximum likelihood method. *BMC Bioinformatics* 7:317. 190

[115] Wu TT, Lange K (2008) Coordinate descent algorithms for lasso penalized regression. *Annals Appl Stat* 2:224–244. 239, 240

[116] Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genomewide association analysis by lasso penalized logistic regression. *Bioinformatics* 25:714–721. 239

[117] Yamagata Z, Asada T, Kinoshita A, Zhang Y, Asaka A (1997) Distribution of apolipoprotein E gene polymorphisms in Japanese patients with Alzheimer's disease and in Japanese centenarians. *Hum Hered* 47:22–26. 166

[118] Xiong M, Jin L (2000) Combined linkage and linkage disequilibrium mapping for genome screens. *Genet Epidemiology* 19:211–234. 230

[119] Zhang T (2010) Analysis of multi-stage convex relaxation for sparse regularization. *J Machine Learning Research* 11:1081–1107. 240

[120] Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26:2375–2382. 239

[121] Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel EM, Lange K (2011) Penalized regression for genome-wide association screening of sequence data. *Pac Symp Biocomput* 2011:106–117. 239

[122] Zhou H, Blangero J, Dyer TD, Chan KHK, Sobel EM, Lange K (2013) Fast genome-wide QTL association mapping with general pedigrees. (Submitted). 288

[123] Zhou JJ, Ghazalpour A, Sobel EM, Sinsheimer JS, Lange K (2012) QTL association mapping by imputation of strain origins in multifounder crosses. *Genetics* 190:459–473. 271, 273

Index