
CHAPTER 11

Tools for Statistical Analysis of Genetic Data

ARUNA BANSAL¹, PETER R. BOYD² and RALPH MCGINNIS¹

¹*GlaxoSmithKline, Population Genetics
New Frontiers Science Park (North)
Third Avenue, Harlow, Essex CM19 5AW, UK*

²*GlaxoSmithKline, Population Genetics
Medicines Research Centre, Gunnels Wood Road
Stevenage, Herts SG1 2NY, UK*

- 11.1 Introduction
- 11.2 Linkage analysis
 - 11.2.1 Parametric linkage analysis
 - 11.2.2 Non-parametric (model-free) linkage analysis
 - 11.2.3 Example: MAPMAKER/SIBS (Kruglyak and Lander, 1995)
 - 11.2.3.1 Data import
 - 11.2.3.2 NPL analysis for a quantitative trait
- 11.3 Association analysis
 - 11.3.1 Transmission disequilibrium tests
- 11.4 Haplotype reconstruction
 - 11.4.1 Example: EHPLUS and PMPLUS (Zhao *et al.*, 2000)
 - 11.4.1.1 Data import
 - 11.4.2 Estimating haplotype frequencies
 - 11.4.3 Haplotype-based association testing
- 11.5 Linkage disequilibrium
 - 11.5.1 Example: Arlequin (Schneider *et al.*, 2000)
 - 11.5.1.1 Data import
 - 11.5.2 Linkage disequilibrium analysis of genotypes with unknown phase
 - 11.5.3 Linkage disequilibrium analysis of haplotypes
- 11.6 Quantitative Trait Locus (QTL) mapping in experimental crosses
 - 11.6.1 Example: Map Manager QTX (Manley *et al.*, 2001)
 - 11.6.1.1 Data import
 - 11.6.1.2 Single marker association
 - 11.6.1.3 Simple interval mapping

Acknowledgements
References

11.1 INTRODUCTION

The focus of this chapter is on methods that aid in the identification of genetic variants that influence a trait of interest. The trait may be a biological measurement, possibly indicating risk of disease or it may be the response to an environmental stimulus such as a drug. Techniques such as linkage analysis and association analysis are central to the process. These methods are described and corresponding software is reviewed, with worked examples to show how they can be applied. The majority of tools covered may be downloaded, together with full documentation, by following links at <http://linkage.rockefeller.edu>. Web addresses for the few exceptions are provided in the text. Almost all are available free of charge.

11.2 LINKAGE ANALYSIS

Linkage analysis is applied in the early stages of gene localization and is one means by which an initial, often broad, chromosomal interval of interest is defined. It is a process of tracking the inheritance pattern of genetic markers with the inheritance pattern of a disease or trait. Disease linkage manifests as a marker allele being inherited in diseased individuals more often than would be expected under independent assortment.

Linkage analysis may be parametric to test whether the inheritance pattern of the trait fits a specific model of inheritance or it may be non-parametric (model-free). The former is more powerful under a correctly specified model and is most informative for large, multiply affected pedigrees. The latter is more powerful when the mode of inheritance is unknown, as in complex trait analysis for which small pedigrees are often ascertained.

11.2.1 Parametric Linkage Analysis

By the parametric approach (and in certain non-parametric cases), evidence of linkage is measured by the LOD score (Morton, 1955). It proceeds by an assessment of the recombination fraction, often denoted by theta (θ). Theta is the probability of a recombination event between the two loci of interest and as such it is a function of distance. Two unlinked loci are given by $\theta = 0.5$ and the closer a pair of loci, the lower their recombination fraction. The LOD may be expressed as follows, using L to denote likelihood.

$$LOD = \log_{10} \frac{L(\theta = \hat{\theta})}{L(\theta = 0.5)}$$

The likelihood in the numerator is based upon the maximum likelihood estimate of the recombination fraction, derived from the data. It is compared to that calculated under the null hypothesis of no linkage ($\theta = 0.5$). A high LOD score is thus consistent with the presence of linkage. Due to the computational complexity of the likelihood calculation, software for exact parametric linkage analysis is constrained either by pedigree size or by the number of markers included in the calculation.

The software VITESSE (O'Connell and Weeks, 1995) allows rapid, exact parametric linkage analysis of very extended pedigrees. At the expense of some speed, an alternative, FASTLINK (Cottingham *et al.*, 1993), allows the analysis of large pedigrees that also

contain loops (marriages between related individuals). Both VITESSE and FASTLINK are based on an earlier program, LINKAGE (Lathrop *et al.*, 1984) and are available for UNIX, VMS and PC(DOS) systems. Using these pieces of software, analysis is typically conducted by means of a sliding window of one, two or four markers along the chromosome, although larger windows are also possible.

Parametric linkage analysis in more moderately-sized pedigrees is commonly carried out using the software GENEHUNTER (Kruglyak *et al.*, 1996). It is written in C, to be run on UNIX and uses a command-line interface. A major feature of this program is that it allows the rapid, simultaneous analysis of dozens of markers (often an entire chromosome) in a multipoint fashion, thereby providing increased power over single-marker analyses when map positions are known (Fulker and Cardon, 1994; Holmans and Clayton, 1995; Olson, 1995). In order to accommodate uncertainty in marker ordering, an option to perform single marker tests is also available. On most platforms, pedigrees up to size $2n - f = 16$ may be analysed by GENEHUNTER, where n is the number of non-founders (those with parents included in the pedigree), and f is the number of founders. This limit is important to consider, because larger pedigrees are automatically trimmed until they fall within it, leading to possible information loss. Results are stored graphically in postscript files for easy interpretation and presentation.

11.2.2 Non-parametric (Model-free) Linkage Analysis

Non-parametric linkage (NPL) analysis does not allow direct estimation of the recombination fraction, but one source of multiple testing — that derived from examining multiple models — is removed. The general principle is that relatives who share similar trait values will exhibit increased sharing of alleles at markers that are linked to a trait locus (see Holmans (2001) for a review of the method).

Allele sharing may be defined as identical by state (IBS) or identical by descent (IBD). Two alleles are IBS if they have the same DNA sequence. They are IBD if, in addition to being IBS, they are descended from (and are copies of) the same ancestral allele (Sham, 1998). A statistical test is performed to compare the observed degree of sharing to that expected under the assumption that the marker and the trait are not linked. While the test statistic may take the form of a chi-squared, normal or F statistic, often it is transformed to allow it to be expressed in LOD units.

NPL analysis often examines IBD or IBS allele sharing in sets of affected sib-pairs (ASPs), in which both siblings exhibit the trait of interest. In the absence of linkage, ASPs are expected to share zero, one or two alleles IBD, with probabilities 0.25, 0.5 and 0.25 respectively. The presence of linkage to a tested marker leads to a departure from these proportions which may be detected by means of a χ^2 test (Cudworth and Woodrow, 1975). Another model-free test, the mean test, tests the null hypothesis that the proportion of IBD allele-sharing equals 0.5. The latter is implemented in the programs SAGE (1999) and SIBPAIR (Terwilliger, 1996), allowing for larger sibships and cases where IBD status cannot be determined unequivocally.

MAPMAKER/SIBS (Kruglyak and Lander, 1995) is a piece of software widely used to test for linkage in sibling data. It was originally written as a stand-alone program, but its functionality and commands have now also been fully incorporated into GENEHUNTER (Kruglyak *et al.*, 1996) whose algorithms are similar. It accommodates both qualitative and quantitative data for either autosomal or sex-linked chromosomes and again, it allows large numbers of markers to be examined jointly.

For dichotomous trait data, a likelihood ratio (LR) test, analogous to the LOD score above is constructed in MAPMAKER/SIBS. The LR is a test for comparing two models in

which the parameters of one model (the reduced model), form a subset of the parameters of the other (the full model). It has many genetic applications and may be expressed as follows, where L denotes likelihood.

$$LR = 2 \log_e \frac{L_{full}}{L_{reduced}}$$

It is asymptotically distributed as a χ^2 , with degrees of freedom equal to the difference in the number of parameters between the two models. In the current context, the numerator is calculated under maximum likelihood estimates of allele sharing proportions and the denominator is calculated assuming random segregation (Risch, 1990a, b). This LR test is also implemented in other software including SPLINK (Holmans and Clayton, 1995), and ASPEX (Hinds and Risch, 1996).

In the case of quantitative trait (QT) data, a test based on the Wilcoxon rank-sum test is available in MAPMAKER/SIBS. It is broadly applicable, as it makes no assumptions concerning the distribution of phenotypic effects. Alternatively, if the sib-pair QT differences are normally distributed, then the original Haseman–Elston method (1972), also implemented, may be applied with greater power. In this test, the squared QT differences between pairs of siblings are regressed on the proportion of alleles that each pair is estimated to share IBD. It is also implemented in SIBPAL2, part of SAGE (1999).

For pedigrees larger than sibships, there is an ‘NPL’ option in GENEHUNTER, but it was shown to be conservative (Kong and Cox, 1997). Alternatives include the modified version, GENEHUNTER-PLUS (Kong and Cox, 1997) and MERLIN (Abecasis *et al.*, 2002), which also incorporates this modification. The latter is a C++ program for UNIX, again with a command-line interface. It offers further improvements in computational speed and reduction in memory constraints, making it more suited to very dense genetic maps. It has the attractive properties of incorporating error detection routines to improve power, and simulation routines to estimate p -values. Graphical output is not however, currently provided.

For normally distributed quantitative traits (or those capable of being transformed to normality), variance component analysis represents a powerful approach to the study of pedigrees of any size (Amos, 1994; Blangero and Almasy, 1996; Goldgar, 1990). The variance component approach to linkage analysis assumes that the joint distribution of the data for a family depends only on means, variances and covariances. The variance of the phenotype is decomposed into (a) components due to linkage to individual marker locations and (b) residual polygenic and environmental components. Familial covariances are modelled in terms of a maximum of two parameters: an additive genetic-variance component and a dominant genetic-variance component, each estimated from the data. The method is implemented in SOLAR (Blangero and Almasy, 1996), in which the size of each effect may be estimated and tested by an LR test. This is a powerful approach and a major advantage is its scope for incorporating into models the effects of covariates, epistasis and gene–environment interaction. For highly complex problems, Markov Chain Monte Carlo Methods are also available, as implemented for example in LOKI (Heath, 1997) and BLOCK (Jensen *et al.*, 1995). When the parameter set is large however, the computational burden of these methods can be prohibitive.

11.2.3 Example: MAPMAKER/SIBS (Kruglyak and Lander, 1995)

11.2.3.1 Data Import

The current example follows a format originally designed for MAPMAKER/SIBS, but now also accommodated by GENEHUNTER. The input files match rather closely what has

become known as ‘LINKAGE format’ due to the software in which it was first introduced (Lathrop *et al.*, 1984; Terwilliger and Ott, 1994). Two files are required, namely a pedigree file and a map file. In the current example, a genetic trait has been simulated for 200 sibships, and the files have been named *regionA.ped* and *regionA.loc* respectively. For the analysis of a quantitative trait, a third file is also required, called for our purposes, *test.pheno*.

The file *regionA.ped* takes the following form where, for simplicity, only a single marker, genotyped in two families has been presented.

70	8699	0	0	2	0	0	0
70	8698	0	0	1	0	0	0
70	2230	8698	8699	2	2	1	2
70	2231	8698	8699	2	2	2	2
75	8787	0	0	2	0	0	0
75	8786	0	0	1	0	0	0
75	2238	8786	8787	2	2	2	2
75	2239	8786	8787	2	2	2	2

The columns are as follows: kindred ID, individual ID, father’s ID, mother’s ID, sex (1 = male, 2 = female), affection status (1 = unaffected, 2 = affected), genotype. In practice, multiple (paired) columns of genotypes would be included, in map order, for each individual. Missing values are denoted by a zero.

This file therefore provides pedigree structure information, genotypes and, in the case of dichotomous traits, phenotype. For liability class data, an additional liability class column may be included after the affection status column and this is described in more detail in the manual.

For the current example, quantitative trait data is loaded separately using *test.pheno* (not shown). This file contains, on the first line, a count of the number of traits in the file. All subsequent lines take the space-separated form: kindred ID, individual ID and phenotype(s). Only sibling phenotypes should be included.

Lastly, the file *regionA.loc* lists the marker details in map order. Here, the ‘internal’ format is described, but LINKAGE format is also supported. The first line provides a count of the number of markers in the file, and is followed by a blank line. Subsequent lines are in six line blocks as follows. The first line has the marker name and number of alleles; the second has the allele labels; the third has the allele frequencies for each label; the fourth is blank; the fifth is the distance to the next marker; the sixth is blank. If the distances are all below 0.5, they are assumed to be recombination fractions, otherwise they are assumed to be distances in cM. The following is an example of a map file, say *regionA.loc* with just the first two markers shown for the sake of brevity.

```

29

MARKER1 6
1 2 3 4 5 6
.01 .95 .01 .01 .01 .01

5.1

MARKER2 10
1 2 3 4 5 6 7 8 9 10
.114988 .110626 .070579 .218874 .250991 .141158 .028549
.062649 .000793 .000793

```

Note that the program tends to crash if inter-marker distances less than 0.1 cM are provided. This should therefore be used as the lower bound even in the case of apparently recombinationally inseparable markers.

11.2.3.2 NPL Analysis for a Quantitative Trait

The following sequence of UNIX commands may be used.

```
load markers regionA.loc
prepare pedigrees regionA.ped
y
test.pheno
increment step 10
scan
p
nonparametric
1
np.out
np.ps
q
```

The process is as follows. The first step is to import the marker and pedigree data that are stored, respectively in *regionA.loc* and *regionA.ped*. You are then asked whether you wish to import additional phenotypic data. Upon typing *y* (yes), you are prompted for a filename, in this case, *test.pheno*. Increment step 10 specifies that linkage is to be assessed at 10 equally spaced points in each marker interval.

The *scan* command computes the full multi-point probability that two sibs share zero, one or two alleles identical by descent (IBD) with the given map and allele frequencies. You are asked whether to include affected (*a*) or phenotyped pairs (*p*). The latter (*p*) allows NPL analysis to follow. Non-parametric linkage analysis is to be applied to trait 1, with numerical output to be piped to *np.out* and graphical output to be stored in *np.ps* (Figure 11.1, below). Note that if only one trait exists in the phenotype file then the 1 above is not required.

As shown in Figure 11.1, a Z-score provides the measure of linkage and in this case evidence peaks close to marker 22. Localization cannot however be assumed to be precise and separation of at least 10 cM may be seen between studies (Hauser and Boehnke, 1997). It is therefore usual to construct a support interval around a strong linkage signal (Conneally *et al.*, 1985). For example, having converted to LOD units, a 1-unit support interval is the interval that includes all (possibly disjoint) map positions with LOD score less than 1 LOD unit below the peak score. A conservative approach is to adopt a 1.5 to 2 LOD support interval. All points within the support interval are considered to be of interest.

A determination of information content in MAPMAKER/SIBS allows a representation of the amount of IBD information extracted by the genotype data, as plotted along the chromosome. Dips in the graph allow regions to be highlighted in which the typing of additional markers could be beneficial. The following commands are applied.

```
scan
a
infomap
info.out
info.ps
```

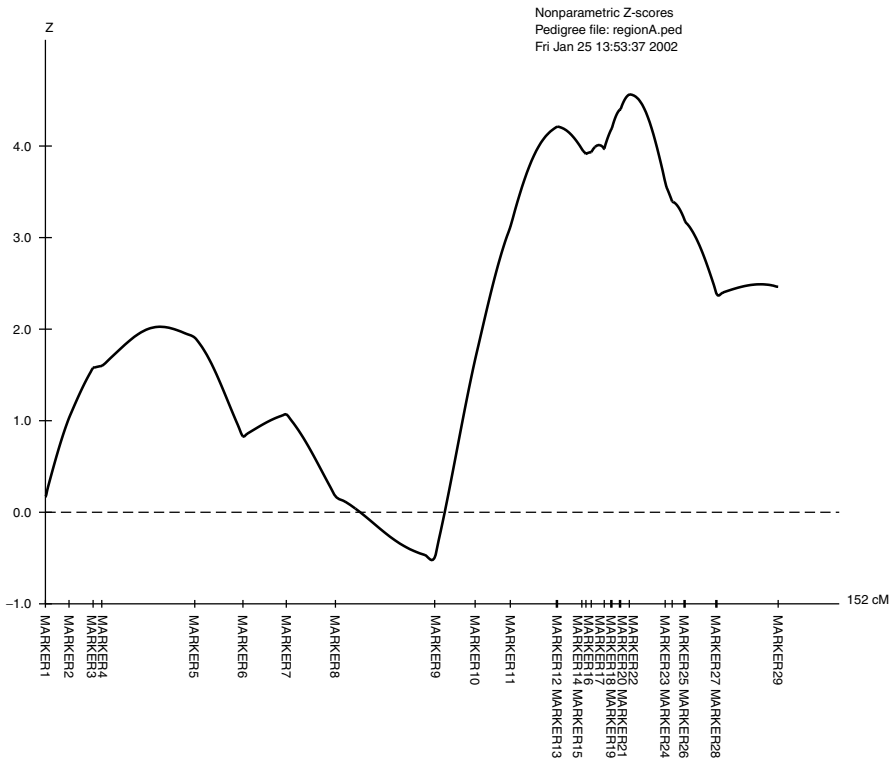


Figure 11.1 Postscript output from MAPMAKER/SIBS. This is *np.ps* from the example run.

A scan of affected pairs (*a*) is conducted and *infomap* is requested. The filenames ensure that numerical output is stored in *info.out* and a graphical representation is saved as *info.ps* (Figure 11.2). In this example the large gaps between markers 4 and 5 and between markers 8 and 9 manifest as troughs in the Information Content graph.

Another useful option (not shown) is that the IBD distribution can be output as a text file using the command *dump ibd*. This is a very rapid means of generating IBD probabilities for sibships and, after re-formatting, the output may be used to generate input files for other software such as QTDT (Abecasis *et al.*, 2000), to be discussed later. Another piece of software, SimWalk2 (Sobel and Lange, 1996) will generate IBD probabilities for a wider range of family structures, but in the case of sibships it is slower than MAPMAKER/SIBS.

11.3 ASSOCIATION ANALYSIS

Association analysis may be regarded as a test for the presence of a difference in allele frequency between cases and controls. A difference does not necessarily imply causality in disease, as many factors, including population history and ethnic make-up may yield this effect. In a well-designed study, however, evidence of association provides a flag

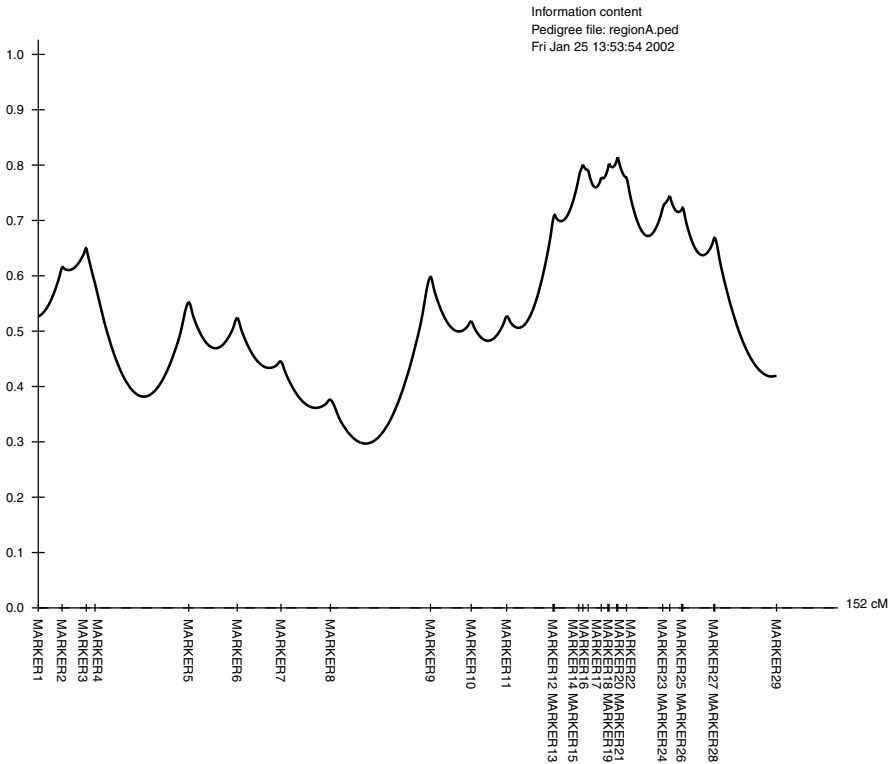


Figure 11.2 Postscript output from MAPMAKER/SIBS. This is *info.ps* from the example run.

for further study. In some instances it is due to the marker being physically close to the causal variant.

Association testing for case–control or population data is often carried out using general (non-genetic) statistical software packages, such as SAS or S-PLUS. A χ^2 test is applied to a contingency table, in which case/control status is tabulated by frequencies of either genotypes or alleles. The test takes the usual form,

$$\chi = \sum \frac{(Obs - Exp)^2}{Exp}$$

where *Obs* and *Exp* are the observed and expected frequencies respectively, and the sum is taken over all cells in the table. The number of degrees of freedom is $(r - 1)(c - 1)$, where r is the number of rows, and c is the number of columns in the table. Equivalently, logistic regression can be applied, using disease status as the dependent variable and alleles or genotypes as the independent variables (see Clayton (2001) for a detailed review of the method). The remaining sections of this chapter all involve applications and extensions of the traditional association test.

11.3.1 Transmission Disequilibrium Tests

In recent years, there has been an upsurge in interest in family-based testing owing to the concern that ethnic mismatching of non-family cases and controls (population stratification) can sometimes yield false positive evidence of association. In particular, the transmission/disequilibrium test or TDT (Spielman *et al.*, 1993) has gained prominence as a test of linkage in the presence of association that does not give false evidence of linkage due to population stratification. The TDT is applied by counting alleles transmitted from heterozygous parents to one or more affected children in nuclear families. The alleles *not* transmitted to affected children may be regarded as control alleles, perfectly ethnically matched to the ‘case’ alleles seen in the affected children. The test takes the form of a McNemar’s test, which, under the null hypothesis of no linkage, follows a χ^2 distribution with one degree of freedom. The TDT is also a valid test for association, but only when applied to alleles transmitted from heterozygous parents to just one affected child per family.

Assuming a diallelic locus, let b denote the counts of heterozygous parent-to-offspring transmissions in which allele 1 goes to an affected child, while allele 2 is not transmitted. Let c denote the counts of transmissions the other way around, in which allele 2 is inherited in an affected child, while allele 1 is not transmitted. The test takes the following form:

$$\chi_1^2 = \frac{(b - c)^2}{(b + c)}$$

A number of groups have focused on generalizing the TDT to quantitative traits or to designs in which parental genotypes are not available. The sib-TDT or S-TDT (Spielman and Ewens, 1998) does not use parental genotypes and, like the original TDT, it is not prone to false positives due to population stratification. For association testing, the S-TDT requires that the data in each family consist of at least one affected and one unaffected sibling, each with different marker genotypes. This test and the original TDT are widely implemented, for example in the Java-based program TDT/S-TDT (Spielman and Ewens, 1996, 1998).

Multi-allelic markers may be tested using ANALYZE (Terwilliger, 1995). This has the advantage of taking LINKAGE format files as input and so provides a natural follow-up to a genome scan. It does however require that LINKAGE (Lathrop *et al.*, 1984) be installed on your system. Other software able to handle multi-allelic markers includes ETDT (Sham and Curtis, 1995) and GASSOC (Schaid, 1996).

For quantitative traits, a major development was the release of QTDT (Abecasis *et al.*, 2000), software which allows TDT testing under a variance components framework. It is applicable to sibships with or without parental genotypes and incorporates a broad range of quantitative trait tests — those proposed by Rabinowitz (1997), Allison (1997), Monks *et al.* (1998), Fulker *et al.* (1999) and Abecasis *et al.* (2000). It is written in C++, to be run on UNIX and has a command-line interface. Its input files are based on LINKAGE format, but in addition, one input file of IBD probabilities must be prepared in advance. QTDT assumes the IBD format generated by the programs SimWalk2 (Sobel and Lange, 1996) and MERLIN (Abecasis *et al.*, 2002). Covariates may also be modelled, but should be kept to a minimum in order to maintain performance.

11.4 HAPLOTYPE RECONSTRUCTION

A haplotype is a string of consecutive alleles lying on the same chromosome. Each individual therefore has a pair of haplotypes for any chromosomal interval—one inherited from the paternal side and one inherited maternally. In statistical genetics, their importance lies in the fact that tests of association may be applied to haplotypes instead of single loci. This may yield increased power if the variant of interest is not being tested directly or if adjacent loci are contributing to a single effect (see Clark *et al.*, 1998; Nickerson *et al.*, 1998). Haplotypes can be inferred from the genotypes of parents or other family members (Weeks *et al.*, 1995) or by laboratory methods (Clark 1990; Nickerson *et al.*, 1998). Often, however, they are estimated by means of the Expectation–Maximization (EM) algorithm (Dempster *et al.*, 1977; Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Little and Rubin, 1987; Long *et al.*, 1995).

The EM algorithm is a method that aims to provide maximum likelihood parameter estimates in the presence of incomplete data. In the case of haplotype frequency estimation, it proceeds as follows (Schneider *et al.*, 2000).

1. An initial set of plausible haplotype frequencies is assigned—for example the product of the relevant allele frequencies may be used.
2. The E-step: assuming Hardy–Weinberg equilibrium, the haplotype frequencies are used to estimate the expected frequencies of ordered genotypes.
3. The M-step: the expected genotype frequencies are used as weights to produce improved estimates of haplotype frequencies.
4. Steps 2 and 3 are repeated until the haplotype frequencies reach equilibrium.

Note that, as with other iterative techniques, it is wise to compare the results of multiple starting points as the EM algorithm may converge to a local, rather than global, optimum. It is not always reasonable to assume that the maximum likelihood haplotype configuration has been reached.

Software written specifically for haplotype analysis includes EHPLUS (Zhao *et al.*, 2000), a reworked and extended version of the earlier program EH (Xie and Ott, 1993). It is written in C and is available in both UNIX and PC versions. EHPLUS can be applied to either case–control data or data assumed to come from a random-mating population. It accommodates large numbers of haplotypes and incorporates a companion program, PMPLUS, which will reformat genotype data ready for use. Estimated haplotypes and their frequencies are output and may be subjected to association tests. Permutation features allow the calculation of empirical *p*-values for these.

Further software for sophisticated haplotype analysis is available from <ftp://ftp-gene.cimr.cam.ac.uk/software/clayton/>. Resources include SNPHAP, a program that uses the EM algorithm to estimate haplotype frequencies for large numbers of diallelic markers using genotype data. Another program, TDTHAP (Clayton and Jones, 1999) allows the TDT to be applied to extended haplotypes. STATA routines to aid SNP selection by haplotype tagging (Johnson *et al.*, 2001) are available in <ftp://ftp-gene.cimr.cam.ac.uk/software/clayton/stata/htSNP/>.

Haplotype reconstruction from family data can be achieved by using SimWalk2 (Sobel and Lange, 1996). The derived haplotypes may then be imported to a pedigree-drawing package such as Cyrillic (Chapman, 1990) for viewing recombinants in positional cloning for example. MERLIN (Abecasis *et al.*, 2002) and GENEHUNTER (Kruglyak *et al.*, 1996) also output haplotypes estimated from family data. Another piece of software,

TRANSMIT (Clayton, 1999) allows association testing of family-based haplotypes. All of these programs allow for missing parental genotypes.

11.4.1 Example: EHPLUS and PMPLUS (Zhao *et al.*, 2000)

11.4.1.1 Data Import

PMPLUS requires two input files, namely a parameter file and a data file. The data file contains for each individual, subject ID, subject status (0 = control, 1 = case) and genotypes listed as either pairs of numbered alleles or as numerical genotype codes. A data file with three markers takes the following form:

```
[Subject ID] [Status] [1a] [1b]    [2a] [2b]    [3a] [3b]
or
[Subject ID] [Status] [1] [2] [3]
```

where [1a] and [1b] are the alleles of the first genotype or, alternatively, [1] alone represents the first genotype. Currently, the compiled limits are 15 alleles, 30 markers and 800 subjects. Note also that whereas subject IDs with a decimal point (e.g. '20.1') work well, more complex IDs containing several dashes and decimal points may lead to erroneous output.

The parameter file consists of five lines of space-delimited integer values, and it defines the tests to be carried out. The following parameter file (*hapfrest.par*) may be used to estimate haplotype frequencies:

```
3 0 0 0
2 2 2
0 0
1 1 1
1 1 1
```

The four values on line 1 specify the number of markers in the data file, whether to perform a marker–marker or case–control analysis (0 or 1, respectively), whether case–control status is to be permuted (0 = no, 1 = yes) and the number of permutations to perform. Line 2 gives the number of alleles for each marker in the data file. The first value on line 3 specifies whether genotypes in the data file are pairs of alleles or numbered genotypes (0 or 1, respectively), while the second value specifies whether screen output is suppressed or shown (0 or 1). Line 4 has a 1 for each marker to be included in the analysis; zero otherwise. Line 5 assigns each marker to one of two blocks (0 or 1), if required in a marker–marker analysis.

11.4.2 Estimating Haplotype Frequencies

Firstly, PMPLUS is run by typing the following:

```
>pmplus hapfrest.par hapfrest.dat hapfrest.out
```

Here *hapfrest.out* is an output file named by the user and created by PMPLUS to record chi-squared statistics and associated *p*-values for the specified analysis. A second output file named *ehplus.dat* is also generated, in which the contents of *hapfrest.dat* have been converted into EHPLUS format ready for estimation of haplotypes.

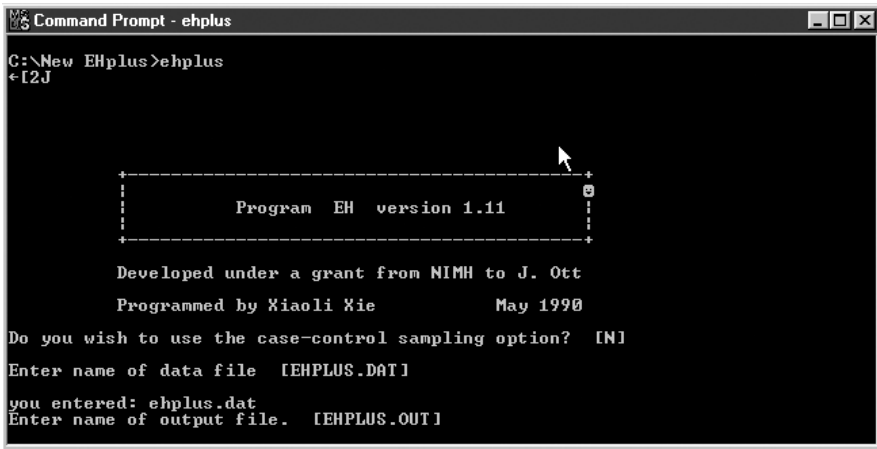


Figure 11.3 EHPLUS interface, as used for estimating haplotype frequencies.

```
-----
# of Typed Individuals: 300

There are 8 Possible Haplotypes of These 3 loci.
They are Listed Below, with their Estimated Frequencies:
```

Allele			Haplotype		Frequency	Observed	Expected	Freeman-Tukey Z:
at	at	at	Independent	w/Association				
Locus 1	Locus 2	Locus 3						
1	1	1	0.253194	0.220146	88.06	101.28	-1.33	
1	1	2	0.011931	0.044928	17.97	4.77	4.11	
1	2	1	0.248181	0.259925	103.97	99.27	0.49	
1	2	2	0.011694	0.000001	0.00	4.68	-3.42	
2	1	1	0.229081	0.239855	95.94	91.63	0.47	
2	1	2	0.010794	0.000071	0.03	4.32	-3.09	
2	2	1	0.224544	0.235074	94.03	89.02	0.46	
2	2	2	0.010501	0.000000	0.00	4.23	-3.23	

```
-----
# of Iterations = 20

                                df      Ln(L)      Chi-square
-----
H0: No Association                3      -474.91      0.00
H1: Allelic Associations Allowed  7      -466.63      16.57
```

Figure 11.4 Haplotype frequency output from EHPLUS — named *ehplus.out*.

Haplotype estimation is carried out by typing `>ehplus` to invoke the program and then pressing the `<CarriageReturn>` three times to accept the default options provided. The process, as seen using the PC(DOS) version, is shown in Figure 11.3. The output file, *ehplus.out*, shown in Figure 11.4, contains the estimated haplotype frequencies (see column labelled *w/Association*) as well as log likelihoods for the null and alternative hypotheses of *No Association* between the markers and *Allelic Associations Allowed* between the markers. Such inter-marker association is termed linkage disequilibrium and shall be the topic of the next section of this chapter.

11.4.3 Haplotype-based Association Testing

In order to test 2-point haplotypes for association to a disease, the parameter file, *hapfrest.par* was modified to produce the following parameter file (*cscentcom.par*):

```

3 1 1 100
2 2 2
0 0
1 1 0
0 0 0
.01 0 1 1

```

The second, third and fourth entries on line 1 of this parameter file now specify that a case–control analysis is to be performed and significance determined by permuting case–control status 100 times. Furthermore, other entries specify that only the first two markers are to be included (line 4), and that genotypes are *not* to be permuted (line 5). This time a sixth line is included, applicable only to case–control analyses. This line contains four possibly non-integer values that define a model of the mode of inheritance of the trait. The first value specifies the assumed disease allele frequency; the following three are penetrant estimates, resulting from zero, one or two copies of the disease allele respectively. In the current example, the model assigns a 0.01 allele frequency and a fully penetrant, pure dominant mode of inheritance. The new data file (*csntcom.dat*) contains the simulated genotypes of both cases and controls, formatted as described previously. PMPLUS is executed as follows:

```
>pmpplus csntcom.par csntcom.dat csntcom.out.
```

When PMPLUS is instructed to perform permutations, EHPLUS is automatically invoked at the end of each PMPLUS run (i.e. following each permutation of the dataset) and thus program control passes back and forth between the two programs until the permutations are complete. Since PMPLUS permutes the data via the *ehplus.dat* input file, the final EHPLUS output file (*ehplus.out*) does not have meaningful haplotype frequency estimates. These are based on permuted, rather than real data.

The output produced by PMPLUS (in this case *csntcom.out*) contains the key analysis results. These are the χ^2 values and permutation-derived *p*-values obtained under five sets of assumptions as follows: (1) under the user-specified disease model, (2) under a Mendelian recessive model, (3) under a Mendelian dominant model, (4) by maximizing the log likelihood ratio over multiple disease models and (5) by a non-parametric ‘homogeneity’ test, to compare log likelihoods calculated from pooling cases and controls and considering them separately. The fifth test is completely non-parametric, while the others are constrained by the population prevalence of disease implied by the user-specified disease model. Figure 11.5 shows the results of evaluating the 2-point haplotype for association with the simulated disease. Note that *p*-values below 0.0001 are rounded down to zero.

11.5 LINKAGE DISEQUILIBRIUM

Linkage disequilibrium (LD) is a lack of independence, in the statistical sense, between the alleles at two loci. LD exists between two linked loci when particular alleles at these loci occur on the same haplotype more often than would be expected by chance alone. This phenomenon can provide valuable information in locating disease variants from marker data, as a marker in LD with the causal variant provides a flag for its location. LD information also provides a means by which the efficiency of high-density marker maps can be increased. If markers are in strong LD with each other, there is an argument for genotyping only a subset of them.

```

Chi-squared statistic for user-specified model = 23.76, df=3, p=0.0000
Chi-squared statistic for recessive model      = 20.58, df=3, p=0.0001
Chi-squared statistic for dominant model       = 23.76, df=3, p=0.0000
Chi-squared statistic for model-free analysis = 23.76, df=4, p=0.0001
Chi-squared statistic for heterogeneity model = 20.38, df=3, p=0.0001

```

```

Random number seed = 3000
Number of replicates = 100

```

```

User-specified model chi-squared statistic (23.76) was reached 0 times
Recessive model chi-squared statistic (20.58) was reached 0 times
Dominant model chi-squared statistic (23.76) was reached 0 times
Model-free chi-squared statistic (23.76) was reached 0 times
Heterogeneity model chi-squared statistic (20.38) was reached 0 times

```

Empirical p-values for these statistics are as follows:

```

T1 - User specified model:      P-value = 0.0000
T2 - Mendelian recessive model: P-value = 0.0000
T3 - Mendelian dominant model:  P-value = 0.0000
T4 - Model-free analysis:      P-value = 0.0000
T5 - Heterogeneity model:      P-value = 0.0000

```

Figure 11.5 Output of haplotype-based association testing in EHPLUS.

The extent of pair-wise LD may be measured by the value D , as follows (Lewontin, 1964). Assume two diallelic loci are linked and let p_{ij} be the proportion of chromosomes that have allele i at the first locus and allele j at the second locus. For example, p_{12} is the frequency of the haplotype with allele 1 at the first locus and allele 2 at the second locus. The disequilibrium coefficient D is the difference between the observed haplotype frequency p_{12} and the haplotype frequency expected under linkage equilibrium, the latter being the product of the two allele frequencies, say p_{1+} and p_{+2} . It may be written as follows:

$$D = p_{12} - p_{1+}p_{+2}$$

Another commonly quoted measure of LD is D' (Lewontin, 1964). This is a normalized form, with numerator equal to D and denominator equal to the absolute maximum D that could be achieved given the allele frequencies at the two loci. Many other valid measures of pair-wise LD exist and have been reviewed elsewhere (Devlin and Risch, 1995; Hedrick, 1987).

As noted above, EHPLUS can perform tests of LD among a group of markers. The complete set of pair-wise tests for the group, together with D and D' values, can be achieved in a single step using software such as Arlequin (Schneider *et al.*, 2000). This is a C++ program available for PC(Win), Linux and MacOS systems. The statistical significance of observed LD is estimated for phase known (haplotype) data by means of a Fisher's Exact Test. For phase unknown data, a likelihood ratio test is applied. An alternative tool is GDA (Lewis and Zaykin, 2001), the PC(Win) companion program to the book, *Genetic Data Analysis II* (Weir, 1996). Both are well documented and perform a broad range of population genetic tests.

The software, GOLD (Abecasis and Cookson, 2000), available for PC(Win), is another program that will calculate D and D' , and it is noteworthy in that it can output them in graphical form. For each marker pair, the pair-wise disequilibrium statistics are colour

coded (bright red to dark blue) and plotted. The output is valuable for presentation purposes and provides a useful summary of the properties of dense maps. The software takes haplotype estimates as input and, in the case of family data, these must be reconstructed using software such as SimWalk2 (Sobel and Lange, 1996) prior to use. Case-control data is not well supported by GOLD, which relies for this purpose upon a limited interface to the software, EH (Xie and Ott, 1993).

Other methods of estimating LD include the Moment Method, applicable to newly-formed populations under certain assumptions concerning the evolutionary process (Hastabacka *et al.*, 1992; Kaplan *et al.*, 1995; Lehesjoki *et al.*, 1993). Maximum likelihood methods have also been explored (Hill and Weir, 1994; Kaplan *et al.*, 1995). Composite likelihood methods were proposed to evaluate the information from multiple pairs of loci simultaneously. Examples of software for the composite likelihood approach include DMAP (Devlin *et al.*, 1996) and ALLASS (Collins and Morton, 1998). The latter uses the Malecot isolation by distance equation and has the advantage of accommodating multiple founder mutations. Each method however relies upon population assumptions and may suffer reduced power when these are not met.

11.5.1 Example: Arlequin (Schneider *et al.*, 2000)

11.5.1.1 Data Import

Arlequin categorizes data into five groups, namely DNA sequences, RFLP data, microsatellite data, allele frequency data and standard data. The latter assumes that different alleles are mutationally equidistant from each other, as is the case with SNP data. Data can be loaded in two ways, by importing a project file, or by using the Project Wizard, to guide you through the creation of a project. Figure 11.6 shows the Arlequin interface in Windows NT, having selected the import screen. As shown, a number of data formats may be read in, and converted by selecting Arlequin as the Target format. LINKAGE format is not however, supported.

With the objective of testing for LD between five markers, the current example may be regarded as a Standard data project. The data and the parameters of the project are shown below in an Arlequin format, for which the filename extension *.arp* is required. The first [Profile] section describes the data before it is listed in the second, [Data] section. Comments are included, preceded by '#' and these are ignored by the program.

```
[Profile]                # first describe the data for this project

Title = 'Simulated data for five genetic markers'
NbSamples = 1           # Number of study populations in the project.
DataType = STANDARD
GenotypicData = 1      # 1= yes; 0 = no (i.e. haplotypic)
LocusSeparator = WHITESPACE
GameticPhase = 0       # 1 = yes; 0 = no (i.e. phase unknown)
RecessiveData = 0      # 1 = yes; 0 = no (i.e. codominant alleles)
RecessiveAllele = null # because RecessiveData = 0
MissingData = '.'      # the missing data code

[Data]                  # next list the data points

[[Samples]]

SampleName = 'Simulation 1'
SampleSize = 200 #200 individuals are in the study set
SampleData = {
```

```

CONFIG1 34 1 1 1 1 2 # The first genotype combination is labelled CONFIG1
          2 1 2 1 2 # 34 individuals have this set of five genotypes
CONFIG2 14 2 1 1 1 2
          2 1 1 1 2
CONFIG3 9 1 1 1 1 2 # 9 individuals have this set of five genotypes
          1 2 2 1 2

```

Subsequent lines of data follow the same paired format and the final line consists of a ‘}’ symbol. This project file is specific to the problem in hand, namely phase-unknown genotype data. Variations exist for other data types and are described in detail in the user manual. It can be seen that genotypes are written with one allele directly below the other allele. This allows a mechanism for inputting phase-known data, for which each line represents a haplotype. In our case, the phase is unknown, so the relative orderings of the alleles are ignored.

Upon successful import, a ‘Project’ is created by Arlequin. It is remembered by the system and can be recalled at a later date. Its details can be viewed by selecting the menu

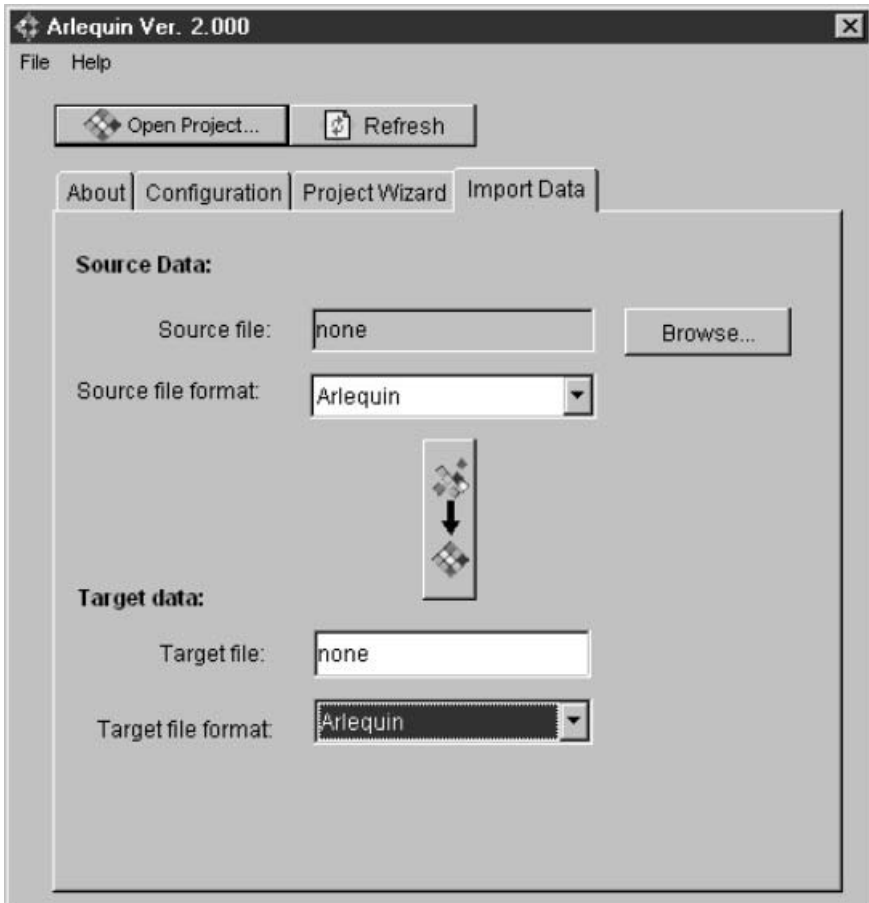


Figure 11.6 Arlequin Screen. Initiating an analysis run.

items Project>View Project Info. The following analyses are then performed by making selections in the launch pad dialogue box.

11.5.2 Linkage Disequilibrium Analysis of Genotypes with Unknown Phase

An LR test statistic, denoted by S , is used to test for LD between a pair of loci when phase is unknown (Slatkin and Excoffier, 1996). It compares the likelihood of a model assuming linkage equilibrium to that of a model allowing linkage disequilibrium. Asymptotically, this statistic follows a χ^2 distribution, but to allow for small sample size or the study of markers with large numbers of alleles, Arlequin also uses a permutation procedure to test for significance.

The analysis screen is given in Figure 11.7. The procedure is as follows:

1. Click on the *Calculation Settings* tab
2. Click to the left of the folder *Linkage disequilibrium*

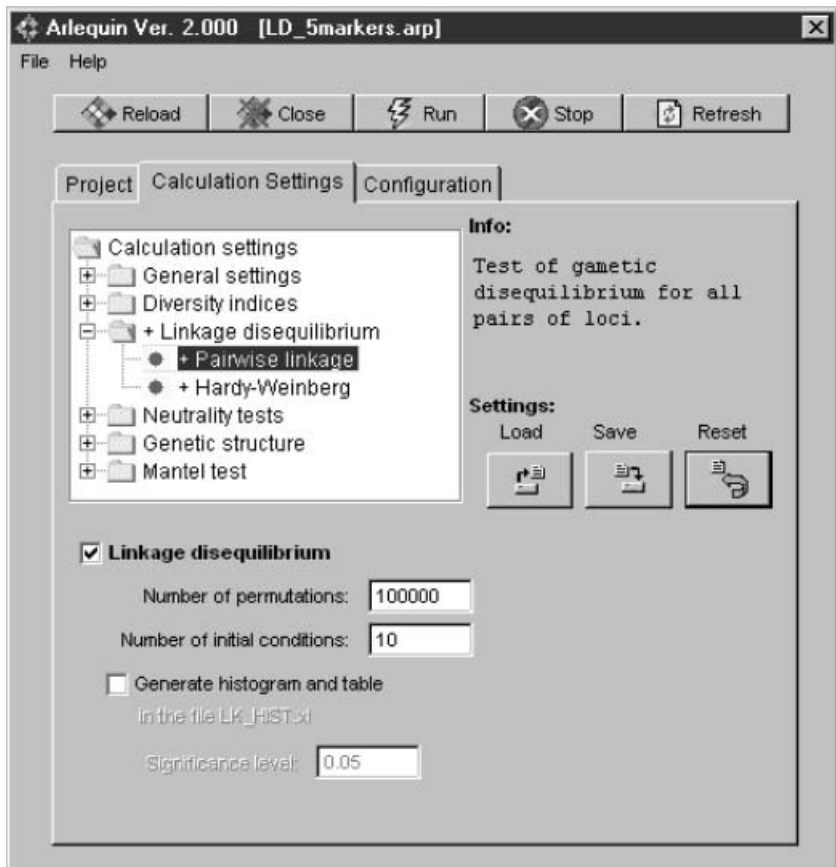


Figure 11.7 Arlequin Screen. Setting up LD analysis of phase-unknown genotype data.

3. Select *Pairwise linkage*
4. Select the *Linkage Disequilibrium* box below the settings window. A plus sign will appear
5. Input parameter values. Default values are given in Figure 11.7. However in the manual, it is recommended that 16,000 permutations be conducted to establish significance and the EM be applied to at least 100 initial conditions
6. Click on the *Run* button

Detailed output is written to an HTML result file, in a sub-directory of that containing the input. First, parameter settings are stated, then for each locus pair, there is a listing of the log-likelihoods under the null and alternative hypotheses, a p -value determined by permutation, the χ^2 test statistic and its corresponding (asymptotic) p -value. Lastly, a table is provided, in which a '+' sign denotes nominal evidence of a departure from linkage equilibrium. This allows the results to be scanned rapidly by eye. Samples of the output are shown below.

```
Pair(0, 1)
  LnLHood LD: -302.71677      LnLHood LE: -319.36838
  Exact P = 0.00000 +- 0.00000 (16002 permutations done)
Chi-square test value = 33.30322 (P = 0.00000, 1 d.f.)
Pair(0, 2)
  LnLHood LD: -420.27411      LnLHood LE: -420.70319
  Exact P = 0.36164 +- 0.00381 (16002 permutations done)
Chi-square test value = 0.85815 (P = 0.35426, 1 d.f.)
```

(and so on)

Table of significant linkage disequilibrium (significance level = 0.0500):

Locus #	0	1	2	3	4
0	*	+	-	-	-
1	+	*	+	-	-
2	-	+	*	-	-
3	-	-	-	*	-
4	-	-	-	-	*

11.5.3 Linkage Disequilibrium Analysis of Haplotypes

Arlequin uses a modified Fisher's Exact test, as opposed to the LR test, to examine LD in haplotype data. Such data is given by *GameticPhase* = 1. The program employs Markov Chain Monte Carlo sampling to explore the space of different possible contingency tables rather than enumerating all the possible contingency tables. In this case, the LD measures, D and D' may also be generated. The analysis screen reflects these additional options as shown in Figure 11.8.

The process of initiating the analysis is very similar to that described above. This time, the number of steps in the Markov chain must be specified, together with the number of de-memorization steps. Again, the default values are lower than those suggested in the manual, which mentions values of 100,000 and 'a few thousand' respectively. If the D and D' boxes are selected, all pair-wise values are tabulated and output in HTML format as well as in a file called *LD_DIS.XL*, ready for inputting to MS Excel.

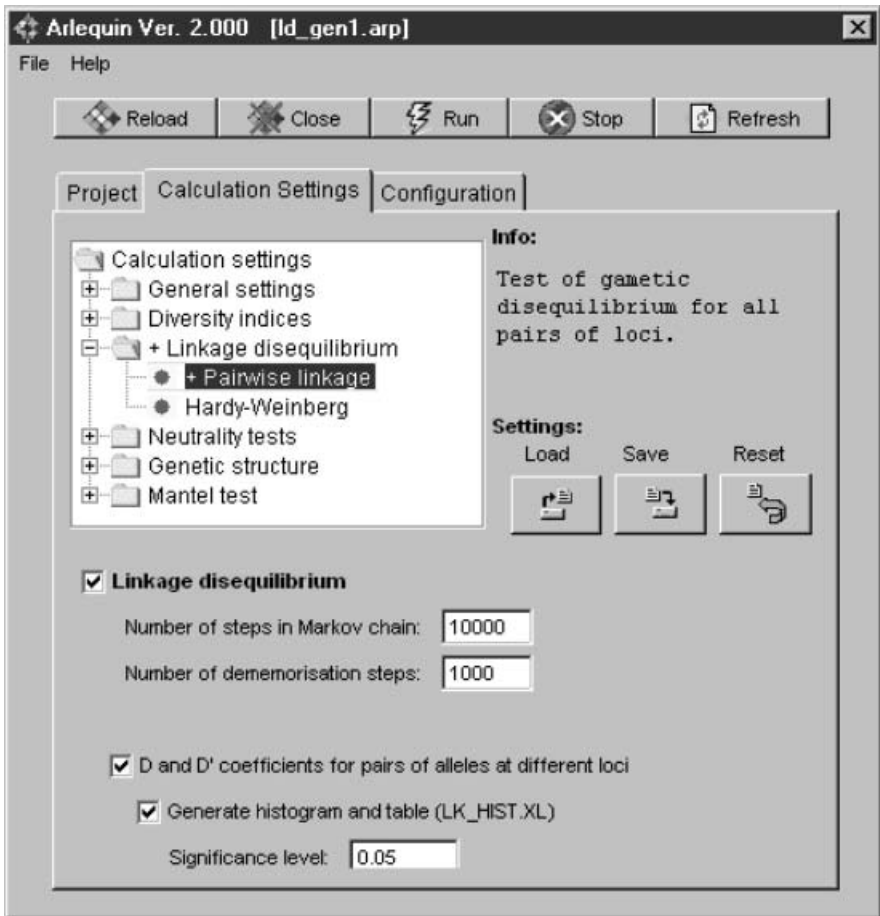


Figure 11.8 Arlequin Screen. Setting up LD analysis of haplotype data.

11.6 QUANTITATIVE TRAIT LOCUS (QTL) MAPPING IN EXPERIMENTAL CROSSES

In contrast to human studies, in which variances of phenotypic differences are used to establish the presence of linkage, QTL mapping in experimental crosses involves comparing means of progeny inheriting specific parental alleles. This is simpler and more powerful (Kruglyak and Lander, 1995). It can be achieved by any of a number of standard statistical methods, such as *t*-tests, analysis of variance (ANOVA), Wilcoxon rank-sum and regression techniques. Again, missing data can be accommodated by an application of the EM algorithm.

Of the very broad array of possible diploid crosses, the following are particularly common. They are derived from a pair of divergent inbred lines in which the genotypes at the majority of loci are homozygous and distinct, say *aa* and *bb* for a particular locus in the two lines respectively. The filial F_1 generation results from crossing these two lines

to produce individuals with heterozygous genotype ab . In the backcross (BC) design, F_1 is crossed with one of the parent strains. For example, in the case of a cross with the aa parent, half the offspring produced are ab and half are aa . In the filial F_2 design, the F_1 is selfed, or two F_1 individuals are crossed so that offspring are aa , ab , and bb in the ratio 1:2:1. Lastly, in the recombinant inbred line (RIL), each F_2 enters individually a single seed descent-inbreeding programme so that all progeny are homozygous for the chosen allele.

The original statistical framework for QTL mapping in experimental crosses was based upon a marker-by-marker analysis. Of particular relevance to sparse maps however, simple interval mapping (IM or SIM) allows the evaluation of any position within a marker interval. The maximum likelihood approach to IM proceeds by calculation of a LOD score (Lander and Botstein, 1989). Similarly, and with lower computational burden, least squares regression achieves the same goal (Haley and Knott, 1992; Martinez and Curnow, 1992). IM may be carried out using a range of software, including MAPMAKER/QTL (Lander *et al.*, 1987). This may appeal to regular users of MAPMAKER/SIBS or GENEHUNTER, as the syntax is similar. It relies upon data pre-processing in MAPMAKER/EXP (Lander *et al.*, 1987) and allows simple graphical output.

Two newer and related methods are Composite Interval Mapping (CIM) and Multiple QTL Mapping (MQM). Both involve performing a genome scan by moving stepwise along the chromosome and testing for the presence of the QTL using a pre-defined set of markers as co-factors (Jansen 1992, 1993; Jansen and Stam, 1994; Kao *et al.*, 1999; Zeng, 1993, 1994; Zeng *et al.*, 1999). In other words, in the sparse map case, interval mapping is combined with multiple regression on markers. This approach allows you to control, to some extent, for effects of other QTLs. Software such as QTL Cartographer (Basten *et al.*, 1994, 1997) and PLABQTL (Utz and Melchinger, 1996; <http://probe.nalusda.gov:8000/otherdocs/jqtl/>) allow the selection of such co-factors by stepwise regression. These programs offer options that will automatically include or exclude background markers according to user-defined criteria.

Lastly, Bayesian methods allow the consideration of multiple QTLs, QTL positions and QTL strengths (Jansen, 1996; Satagopan *et al.*, 1996; Sillanpaa and Arjas, 1998; Uimari *et al.*, 1996). The software Multimapper (Sillanpaa, 1998), for example, allows the automatic building of models of multiple QTLs within the same linkage group. It is designed to work as a companion program to QTL Cartographer (Basten *et al.*, 1994, 1997) and allows a more detailed follow-up of regions of interest. As with other Markov Chain Monte Carlo methods, however, this approach is computer intensive and may suffer from problems of convergence to a local, rather than global, optimum or of lack of convergence if run for a short time.

Ten of the most prominent pieces of software for QTL mapping are reviewed in greater detail by Manley and Olson (1999). The majority will perform IM and CIM for backcross, filial F_2 and recombinant inbred lines. Cordell (2002) provides worked examples of the usage of three of them, MAPMAKER/QTL, QTL Cartographer and another piece of software, MapQTL (van Ooijen and Maliepaard, 1996a, b).

A major limitation of QTL mapping using inbred lines is the broad, ill-defined nature of the resulting linkage peaks, which typically span tens of centiMorgans even if large numbers of progeny are analysed (for example see Farmer *et al.*, 2001). This is a consequence of the multifactorial nature of quantitative traits, which results in an inability to identify unequivocal recombinants that precisely delineate a critical genetic interval, in contrast with monogenic phenotypes. Subsequent attempts to narrow a locus by, for example, successive rounds of backcrossing are often frustrated by the dilution or loss

of unlinked genetic co-factors that are required for trait manifestation. In the future, QTL mapping using genetically heterogeneous stocks may gain in prominence (Mott *et al.*, 2000). Talbot *et al.* (1999) were able to achieve a mapping resolution of less than 1 cM by the study of heterogeneous stocks from eight known inbred mouse progenitor strains that had been intercrossed over 30–60 generations. The group has released software called HAPPY (Mott *et al.*, 2000) which requires knowledge of the ancestral alleles in the inbred founders, together with the genotypes and phenotypes in the final generation. It will then apply variance component methods to test for linkage to the QTL.

11.6.1 Example: Map Manager QTX (Manley *et al.*, 2001)

Map Manager QTX is available for both MacOS and PC(Win). It has no licence fee and was selected here due to the usefulness of its graphic user interface. It has both IM and CIM capability and can reformat data for use in other important software such as QTL Cartographer. Interval mapping is based on the Haley and Knott (1992) procedure, and CIM is achieved by adding background loci. Significance can be assessed by permutation (Churchill and Doerge, 1994).

The genotype data may derive from inbred or non-inbred stock and options are provided for a variety of experimental designs. Extensive documentation can be downloaded in either pdf or Hypertext formats. The *Tutorial* is especially helpful; but readers should be aware that its files are somewhat inconspicuously tucked in with *Sample Data* files, rather than being included in the Map Manager QTX Manual.

For the current example, genotype data was downloaded from the Mouse Genome Database (2001). Specifically, it consists of mouse chromosome 1 genotypes from the Copeland–Jenkins backcross, and a selected subset of 10 markers spanning the entire ~ 100-cM length of the chromosome. Marker *En1* is located near the middle of the chromosome, between markers *Col6a3* and *D1Fcr15*, and it was used to simulate the quantitative trait (QT) for the 193 backcross mice. Homozygotes (denoted as *b*) at *En1* received a QT value of 50 ± 20 (mean \pm SD) while heterozygotes (*s*) at *En1* received a QT value of 100 ± 20 . *En1* was then removed from the dataset and Map Manager QTX was used to analyse QT association with the remaining nine markers as shown below.

11.6.1.1 Data Import

Map Manager QTX is launched by a mouse click on the Map Manager icon (*QTXb13.exe*), thus opening the main menu. The genotype data (alternatively termed ‘Phenotype data’ by Map Manager QTX) is imported by selecting *File>Import>Text*. The name of each marker and the genotypes (phenotypes) of the cross progeny are imported as a single line of text. The marker name is separated from the genotypes by a tab character but the genotypes, each represented as above by a single letter, can be given as either an unbroken string of characters or space-separated. In our case, the first two lines of input therefore took the following form (with missing genotypes given by a hyphen):

```
Actn3<tab>sssbbbbbsbsbsbsssbbsbbsbbbbssbsbbsbsb-bbb-ssss
  <CarriageReturn>
Laf4<tab>-sbbbb- -sb- - - - - -bb- -bsbb-bbsb- -s-bbbbbbsbssb-
  bs<CarriageReturn>
```

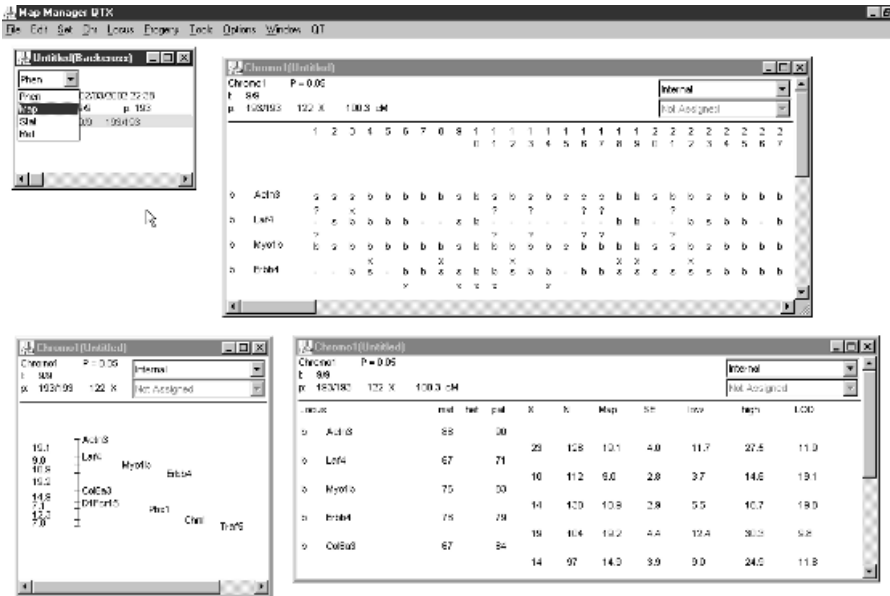


Figure 11.9 Screens in Map Manager QTX. The dataset window (upper left), the Phenotype window (upper right), the Map window (lower left) and the Statistics window (lower right). Genotypes with permission from Mouse Genome Database (2001).

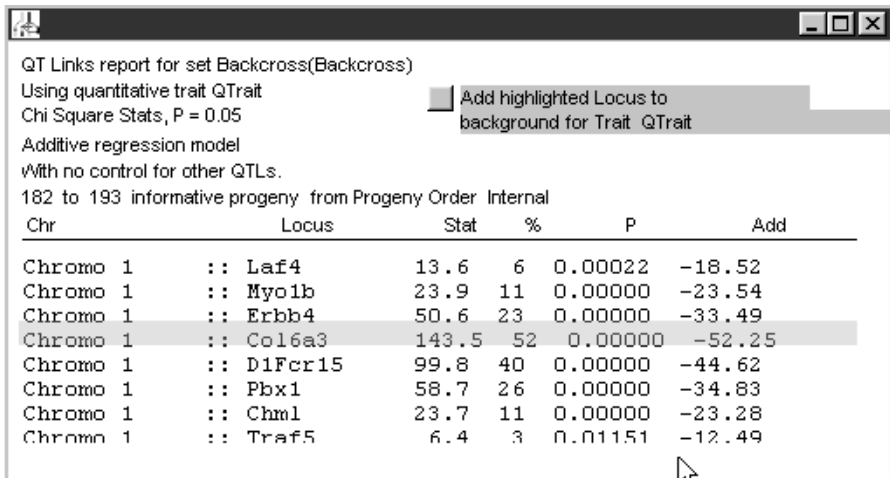


Figure 11.10 Output from single marker association testing in Map Manager QTX: The 'Links Report'. 'Add' denotes the additive regression coefficient for the association. Genotypes with permission from Mouse Genome Database (2001).

Quantitative trait data are then read in from a second text file via *File>Import>Trait Text*. The format is almost identical, except that the name of the trait replaces marker name and the trait value for each mouse must be separated from adjacent values by at least one space. Again, the name of the quantitative trait and all of the values for cross progeny must be in a single line of text.

Successful import of a text genotype file produces a small pop-up window (the *dataset* window), as shown in Figure 11.9, top left. Within it is a menu allowing selection of *Phen*, *Map*, *Stat* or *Ref*. Selecting one of these options and double-clicking on a chromosome name in the *dataset* window, produces the chosen window as shown in Figure 11.9. The *Phenotype* window (top right) displays the marker names on the left side of the window, with one column for each member of the progeny. The body of the *Phenotype* window shows the genotype at each locus and also indicates locations of recombination

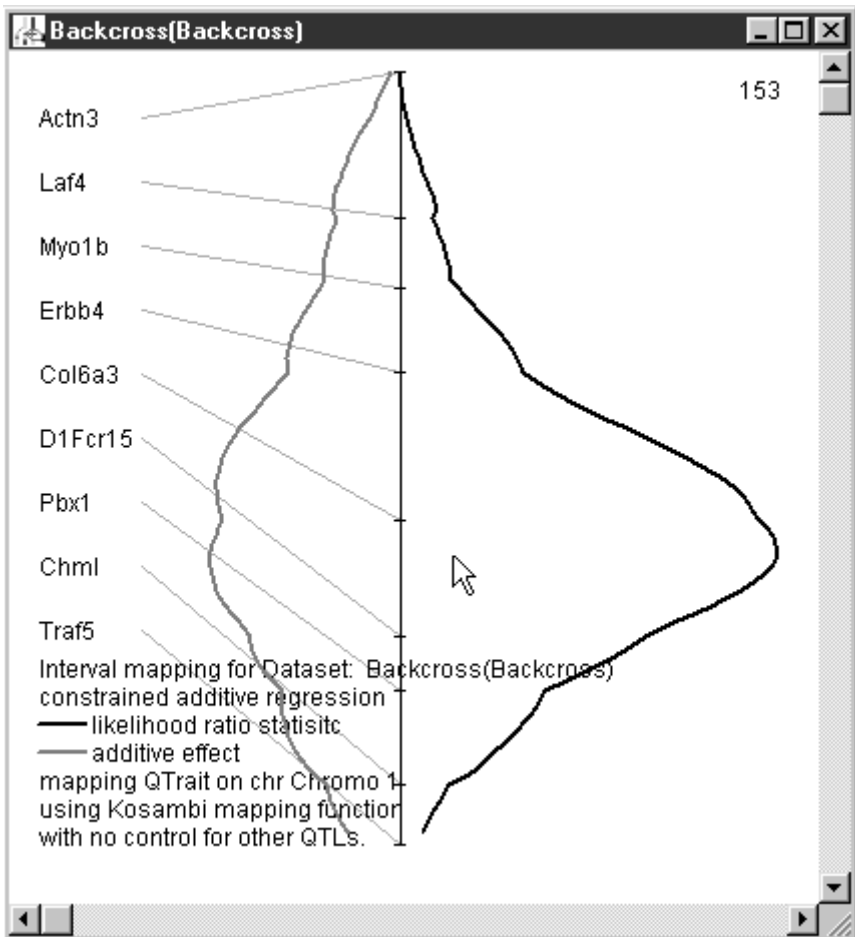


Figure 11.11 Output from Map Manager QTL. Results of interval mapping across nine markers. Genotypes with permission from Mouse Genome Database (2001).

events with an X . Pairs of question marks denote the possible locations of crossovers whose more precise location cannot be specified due to missing genotype data. The *Map* window (bottom left), shows a genetic map with estimated cM distance between markers, and the *Statistics* window (bottom right) summarizes useful numerical information, such as the number of recombination events between adjacent markers and LOD evidence for linkage.

11.6.1.2 Single marker association

Testing for association between an individual marker and a quantitative trait is accomplished by first selecting a p -value cut-off in the *Main* menu under *Options>Search&Linkage criteria*, and then choosing *QT>Links Report* in the *Main* menu. This produces a window allowing the user to select both the name of the quantitative trait to test and the background QTLs to be included in the analysis.

Figure 11.10 shows the table or *Links Report* that was produced by testing each of the nine markers in our panel for association with the simulated trait. Note that only eight markers appear in the table, as one marker did not meet the $p < 0.05$ criterion. Note also that marker *Col6a3* is highlighted as giving the strongest association and therefore as being the best marker to include as a background QTL in analyses of other chromosomal loci.

11.6.1.3 Simple Interval Mapping

Simple interval mapping of a QT across a series of markers is accomplished by choosing *QT>Interval Mapping* from the *Main* menu. This produces a window which again allows the user to specify the trait to be analysed and whether any background QTLs are to be included in the analysis. Once options in this window are specified, Map Manager QTX produces a table and a figure displaying the Interval Mapping results. Figure 11.11 shows the result of interval mapping our simulated trait across the nine markers on mouse chromosome 1. As indicated by the position of the cursor, the peak of the likelihood ratio statistic falls very close to the true location of the simulated QT locus, between markers *Col6a3* and *DIFcr15*.

ACKNOWLEDGEMENTS

The authors wish to thank Heather Cordell, Dmitri Zaykin, Clive Bowman, Meg Ehm and Leonid Kruglyak for helpful discussions, advice and support during the writing of this chapter.

REFERENCES

- Abecasis GR, Cardon LR, Cookson WO. (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**: 279–292.
- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30**: 97–101.
- Abecasis GR, Cookson WOC. (2000). GOLD—Graphical Overview of Linkage Disequilibrium. *Bioinformatics* **16**: 182–183.
- Allison DB. (1997). Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* **60**: 676–690.

- Amos CI. (1994). Robust variance components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* **54**: 535–543.
- Basten C, Weir BS, Zeng Z-B. (1994). Zmap—a QTL cartographer. In Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W, Gibson JP, Kennedy BW, Burnside EB. *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software* vol. 22. pp. 65–66. (On-line publication)
- Basten C, Weir BS, Zeng Z-B. (1997). *QTL Cartographer: A Reference Manual and Tutorial for QTL Mapping*. Department of Statistics, North Carolina State University: Raleigh, NC. (<http://statgen.ncsu.edu/qtcart/>).
- Blangero J, Almasy L. (1996). *SOLAR: Sequential Oligogenic Linkage Analysis Routines*. Technical notes no. 6, Population Genetics Laboratory, Southwest Foundation for Biomedical Research: San Antonio, TX.
- Chapman CJ. (1990). A visual interface to computer programs for linkage analysis. *Am J Med Genet* **36**: 155–160.
- Churchill GA, Doerge RW. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
- Clark AG. (1990). Inference of haplotypes from PCR amplified samples of diploid populations. *Mol Biol Evol* **7**: 111–122.
- Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, *et al.* (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* **63**: 595–612.
- Clayton D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**: 1170–1177.
- Clayton D. (2001). Population association. In Balding DJ, Bishop M, Cannings C. (Eds), *Handbook of Statistical Genetics*. John Wiley: Chichester, pp. 519–540.
- Clayton D, Jones HB. (1999). Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* **65**: 1161–1169.
- Collins A, Morton NE. (1998). Mapping a disease locus by allelic association. *Proc Natl Acad Sci USA* **95**: 1741–1745.
- Conneally PM, Edwards JH, Kidd KK, Lalouel J-M, Morton NE, Ott J, *et al.* (1985). Report of the committee on methods of linkage analysis and reporting. *Cytogenet Cell Genet* **40**: 356–359.
- Cordell HJ. (2002). Diabetes in the NOD mouse. In Camp N, Cox A. (Eds), *Quantitative Trait Loci: Methods and Protocols*. Humana Press: pp. 165–198.
- Cottingham RW Jr, Idury RM, Schaffer AA. (1993). Fast sequential genetic linkage computation. *Am J Hum Genet* **53**: 252–263.
- Cudworth AG, Woodrow JC. (1975). Evidence for HLA-linked genes in ‘juvenile’ diabetes mellitus. *Br Med J* **3**: 133–135.
- Dempster AP, Laird NM, Rubin DB. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc* **B39**: 1–38.
- Devlin B, Risch N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311–322.
- Devlin B, Risch N, Roeder K. (1996). Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**: 1–16.
- Excoffier L, Slatkin M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921–927.
- Farmer MA, Sundberg JP, Bristol IJ, Churchill GA, Li R, Elson CO, *et al.* (2001). A major quantitative trait locus on chromosome 3 controls colitis severity in IL-10-deficient mice. *Proc Natl Acad Sci USA* **98**: 13820–13825.

- Fulker DW, Cardon LR. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *Am J Hum Genet* **54**: 1092–1103.
- Fulker DW, Cherny SS, Sham PC, Hewitt JK. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* **64**: 259–267.
- Goldgar DE. (1990). Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* **47**: 957–967.
- Haley CS, Knott SA. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *J Hered* **69**: 315–324.
- Haseman JK, Elston RC. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**: 3–19.
- Hastabacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. (1992). Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. *Nature Genet* **2**: 204–211.
- Hauser ER, Boehnke M. (1997). Confirmation of linkage results in affected-sib-pair linkage analysis for complex genetic traits. *Am J Hum Genet* **61**: A278.
- Hawley ME, Kidd KK. (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* **86**: 409–411.
- Heath S. (1997). Markov chain segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.
- Hedrick PW. (1987). Gametic disequilibrium measures: Proceed with caution. *Genetics* **117**: 331–341.
- Hill WG, Weir BS. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet* **54**: 705–714.
- Hinds D, Risch N. (1996). The ASPEX package: affected sib-pair mapping <ftp://lahmed.stanford.edu/pub/aspeex>.
- Holmans P. (2001). Nonparametric linkage. In Balding DJ, Bishop M, Cannings C. (Eds), *Handbook of Statistical Genetics*. John Wiley: Chichester, pp. 487–505.
- Holmans P, Clayton D. (1995). Efficiency of typing unaffected relatives in an affected sib-pair linkage study with single locus and multiple tightly-linked markers. *Am J Hum Genet* **57**: 1221–1232.
- Jansen RC. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**: 252–260.
- Jansen RC. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**: 205–211.
- Jansen RC. (1996). A general Monte Carlo method for mapping multiple quantitative trait loci. *Genetics* **142**: 305–311.
- Jansen RC, Stam P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**: 1447–1455.
- Jensen CS, Kong A, Kjaerulff KM. (1995). Blocking–Gibbs sampling in very large probabilistic expert systems. *Int J Hum Computer Studies* **647**–666.
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.
- Kao CH, Zeng Z-B, Teasdale RD. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.
- Kaplan N, Hill WG, Weir BS. (1995). Likelihood methods for locating disease genes in nonequilibrium populations. *Am J Hum Genet* **56**: 18–32.
- Kong A, Cox NJ. (1997). Allele sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* **61**: 1179–1188.

- Kruglyak L, Lander ES. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* **57**: 439–454.
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**: 1347–1363.
- Lander ES, Botstein D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.
- Lander ES, Green P, Abrahamson J, Barlow A, Daly M, Lincoln SE, *et al.* (1987). MAP-MAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**: 174–181.
- Lathrop GM, Lalouel JM, Julier C, Ott J. (1984). Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci USA* **81**: 3443–3446.
- Lehesjoki A-E, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, *et al.* (1993). Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: Linkage disequilibrium allows high resolution mapping. *Hum Mol Genet* **2**: 1229–1234.
- Lewis PO, Zaykin D. (2001). Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from <http://lewis.eeb.uconn.edu/lewishome/software.html>.
- Lewontin RC. (1964). The interaction of selection and linkage I. General considerations; heterotic models. *Genetics* **49**: 49–67.
- Little RJA, Rubin DB. (1987). *Statistical Analysis with Missing Data*. Wiley: New York.
- Long JC, Williams RC, Urbanek M. (1995). An E-M algorithm and testing strategy for multiple locus haplotypes. *Am J Hum Genet* **56**: 799–810.
- Manly KF, Olson JM. (1999). Overview of QTL mapping software and introduction to map manager QT. *Mamm Genome* **10**: 327–334.
- Manly KF, Cudmore RH, Meer JM. (2001). Map Manager QTX, cross-platform software for genetic mapping. *Mamm Genome* **12**: 930–932.
- Martinez O, Curnow RN. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet* **85**: 480–488.
- Monks SA, Kaplan NL, Weir BS. (1998). A comparative study of sibship tests of linkage and/or association. *Am J Hum Genet* **63**: 1507–1516.
- Morton NE. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet* **7**: 277–318.
- Mott R, Talbot C, Turri M, Collins AC, Flint J. (2000). A new method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci USA* **97**: 12649–12654.
- Mouse Genome Database (MGD). (2001). Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine. World Wide Web (URL: <http://www.informatics.jax.org/>).
- Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, *et al.* (1998). DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet* **19**: 233–240.
- O'Connell JR, Weeks DE. (1995). The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genet* **11**: 402–408.
- Olson JM. (1995). Multipoint linkage analysis using sib pairs: an interval mapping approach for dichotomous outcomes. *Am J Hum Genet* **56**: 788–798.
- Rabinowitz D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum Hered* **47**: 342–350.

- Risch N. (1990a). Linkage strategies for genetically complex traits II. The power of affected relative pairs. *Am J Hum Genet* **46**: 229–241.
- Risch N. (1990b). Linkage strategies for genetically complex traits: III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* **46**: 242–253.
- SAGE (1999). *Statistical Analysis for Genetic Epidemiology*, Release 4.0. Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth campus, Case Western Reserve University: Cleveland, OH.
- Satagopan JM, Yandell BS, Newton MA, Osborn TC. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**: 805–816.
- Schaid DJ. (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* **13**: 423–449.
- Schneider S, Roessli D, Excoffier L. (2000). *Arlequin ver. 2.000: A software for population genetics data analysis*. Genetics and Biometry Laboratory, University of Geneva: Geneva, Switzerland.
- Sham PC. (1998). *Statistics in Human Genetics*. Arnold Publishers: London; John Wiley and Sons Inc.: New York.
- Sham PC, Curtis D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* **59**: 323–336.
- Sillanpaa MJ (1998). Multimapper Reference Manual. <http://www.RNL.Helsinki.FI/~mjs/>.
- Sillanpaa MJ, Arjas E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.
- Slatkin M, Excoffier L. (1996). Testing for linkage disequilibrium in genotypic data using the EM algorithm. *Heredity* **76**: 377–383.
- Sobel E, Lange K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am J Hum Genet* **58**: 1323–1337.
- Spielman RS, Ewens WJ. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* **59**: 983–989.
- Spielman RS, Ewens WJ. (1998). A sibship test for linkage in the presence of association: the sib-transmission/disequilibrium test. *Am J Hum Genet* **62**: 450–458.
- Spielman RS, McGinnis RE, Ewens WJ. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *Am J Hum Genet* **52**: 506–516.
- Talbot CJ, Nicod A, Cherny SS, Fulker DW, Collins AC, Flint J. (1999). High-resolution mapping of quantitative trait loci in outbred mice. *Nature Genet* **21**: 305–308.
- Terwilliger JD. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet* **56**: 777–787.
- Terwilliger JD. (1996). Program SIBPAIR—sib pair analysis on nuclear families. <ftp://linkage.cpmc.columbia.edu>.
- Terwilliger JD, Ott J. (1994). *Handbook of Human Genetic Linkage*. Johns Hopkins: Baltimore.
- Uimari P, Thaller G, Hoeschele I. (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* **143**: 1831–1842.
- Utz HF, Melchinger AE. (1996). PLABQTL: a program for composite interval mapping of QTL. *J Quant Trait Loci* **2**, <http://probe.nalusda.gov:8000/otherdocs/jqtl/>.
- van Ooijen JW, Maliapaard C. (1996a). MapQTL version 3.0: software for the calculation of QTL positions on genetic maps. Plant Genome IV abstracts. <http://probe.nalusda.gov:3000/otherdocs/pg/pg4/abstracts/p316.html>.

- van Ooijen JW, Maliepaard C. (1996b). MapQTL version 3.0: software for the calculation of QTL positions on genetic maps. CPRO-DLLO: Wageningen, ISBN90-73771-23-4.
- Weeks DE, Sobel E, O'Connell JR, Lange K. (1995). Computer programs for multilocus haplotyping of general pedigrees. *Am J Hum Genet* **56**: 1506–1507.
- Weir BS. (1996). *Genetic Data Analysis II*. Sinauer Associates Inc Publishers: Sunderland, MA, USA.
- Xie X, Ott J. (1993). Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet* **53**: 1107.
- Zeng Z-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci USA* **90**: 10972–10976.
- Zeng Z-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.
- Zeng Z-B, Kao CH, Basten CJ. (1999). Estimating the genetic architecture of quantitative traits. *Genet Res* **74**: 279–289.
- Zhao JH, Curtis D, Sham PC. (2000). Model-free analysis and permutation tests for allelic associations. *Hum Hered* **50**: 133–139.