

Chapter 2

Linkage Analysis

Jennifer H. Barrett and M. Dawn Teare

Abstract

Linkage analysis is used to map genetic loci using observations on relatives. It can be applied to both major gene disorders (parametric linkage) and complex diseases (model-free or non-parametric linkage), and it can be based on either a relatively small number of microsatellite markers or a denser map of single nucleotide polymorphisms (SNPs). We describe the methods commonly used to map loci influencing disease susceptibility or a quantitative trait. Application of these methods to simulated but realistic datasets is illustrated in some detail using the program Merlin. We provide some guidance on the best methods to use in different situations and on the interpretation of output.

Key words: Linkage analysis, genetic mapping, parametric, non-parametric, Merlin, quantitative traits.

1. Introduction

Genetic linkage analysis observes the segregation of alleles at meiosis to infer the distances between genetic loci. In the context of disease or trait gene mapping, a panel of established genetic markers (i.e. ones where the chromosomal location is known) are used to effectively label the participants' genomes so that the segregation of genetic material can be observed or followed (*see Chapter 1* for more discussion on genetic markers). In the genetic analysis of quantitative traits, the families may be a random population sample. However, in most disease mapping studies, the families have been selected due to the presence of the trait of interest occurring in at least one of the family members. The linkage analysis consists of studying the pattern of co-inheritance of marker alleles and the presence/absence or quantitative measure of a phenotype. The evidence against the null hypothesis (that the risk locus is unlinked to this position) is then reported at

each examined position in the genome. In the context of linkage analysis the term genome does not include mitochondrial DNA.

In this chapter, two forms of linkage analysis are presented: parametric (or model based) and non-parametric (model-free). Parametric linkage analysis requires the investigator to specify the genetic model in advance of mapping the locus. For simple fully penetrant Mendelian traits (such as recessive or dominant traits), there is frequent evidence from clinical experience to support the assumed mode of inheritance. In more common diseases where only a component of the disease risk is attributable to genetic causes, the genetic model can be estimated through segregation analyses or familial aggregation studies. Before its characterisation it is usual to assume that the genetic risk locus will have only two alleles: the common normal form or wild type (denoted by **d**) and the mutant risk-associated form (**D**). There are four model parameters which must be specified. These are the population allele frequency of the **D** allele, and the probabilities of being affected (penetrances) for each of the three genotypes: **dd**, **Dd**, and **DD**. For a simple dominant fully penetrant disorder the respective penetrances would be 0, 1, and 1. A dominant disorder with incomplete penetrance of 80% would have penetrances specified by 0, 0.8, and 0.8. The allele frequency parameter can be estimated from population incidence. In complex disease it is common to allow for a sporadic rate of disease, which means that there is risk of disease in the **dd** genotype, so the corresponding penetrance is greater than zero. In addition, the risk for the **Dd** and **DD** genotypes is generally relatively low, and the penetrance of the **DD** genotype is often higher than that of the **Dd** genotype.

In parametric linkage, the alleles present in an individual at the disease locus are not observed directly but can be inferred through the individual's observed phenotype, the genetic risk model, and the pattern of inheritance of phenotypes in the family. In its simplest form, genetic linkage analysis or mapping attempts to estimate the genetic distance between a pair of loci by studying informative meioses and counting recombinations. An informative meiosis requires heterozygous genotypes at both marker and disease locus. The recombination fraction represents the probability of a gametic recombination of alleles occurring between a pair of loci during meiosis. If the recombination fraction equals $\frac{1}{2}$, then recombinants and non-recombinants are equally likely and the two loci are said to be unlinked. The smaller the recombination fraction the closer two loci are located along the same chromosome.

When considering several loci along the same chromosome, it is more common to see genetic distance reported than the recombination fraction. The recombination fractions between each pair of loci can be transformed to the additive genetic distance (and vice versa) by an appropriate mapping function such as Haldane or

Kosambi (2). Genetic distance is reported in centiMorgans (cM), where a recombination fraction of 0.01 is equivalent to 1 cM.

The statistical evidence for linkage is traditionally reported as a maximum LOD (logarithm of the odds) score. The LOD score function summarises the statistical evidence resulting from the joint analysis of all of the families. It compares the evidence that the disease risk locus resides at a specific location with the evidence that it resides at an unlinked part of the genome (i.e. on a different chromosome or a very distant part of the same chromosome). Parametric linkage analysis is very powerful for mapping rare risk loci with strong effects (3). It can also be successful in detecting linkage when there is locus heterogeneity, i.e. when distinct genes can independently give rise to the same phenotype. To allow for heterogeneity in the parametric framework, an additional parameter is estimated, which is the proportion of families linked (α). The evidence for linkage is assessed by computation of a heterogeneity LOD score (or HLOD). The value of the LOD score captures the evidence of linkage: a high positive score is evidence for linkage and a large negative score is evidence against. For the early years when genetic markers were not so numerous, the threshold of '+3' was required to be reached to declare evidence of linkage. Since the availability of high-density genomic markers, several authors reviewed the thresholds for a variety of study designs (4, 5). Empirical p values associated with a peak LOD score can now be calculated post-linkage analysis by simulation.

Though a parametric heterogeneity LOD score analysis can detect linkage if a small proportion of families are unlinked to the candidate region, a dramatic loss in power is seen as the unlinked proportion increases. Once substantial genetic heterogeneity is suspected (generally after parametric linkage analysis has been unsuccessful), a non-parametric or model-free approach is favoured. In this context non-parametric means that a genetic model is not specified. The model-free approach studies the pattern of alleles shared identical by descent (IBD) between pairs or groups of relatives. Relatives may carry copies of alleles that are descended from recent common ancestors, and such alleles are said to be IBD. Assuming neutral alleles and random mating, the expected IBD sharing probabilities for any pair (or group) of relatives can be calculated (6). Non-parametric linkage requires at least two affected relatives per family, the hypothesis being that, if there is an inherited genetic component to the disease, increased IBD sharing between affected relatives will be seen in the genomic location of the risk loci. In the model-free context, the evidence to support linkage can also be reported with LOD scores, but often a Z -score is reported. It is important to be clear as to whether one is reporting a LOD score or a Z -score, as the thresholds corresponding to the same p values are different.

2. Materials

There are many available software packages for linkage analysis, most of which are freely available. These include Linkage (7), one of the earliest linkage analysis programs, with a faster adaptation in Fastlink (8), Genhunter (9), which is widely used for non-parametric linkage analysis, and Allegro (10) (a faster and modified version of this), Morgan (11), which uses Monte Carlo Markov Chain methods and is suited to handle large complex pedigrees, Solar (12) for quantitative trait linkage analysis and Merlin (1), which can cope with very large numbers of marker loci. A more comprehensive list can be found on the Web site <http://www.nslj-genetics.org/soft/>. We have chosen to use Merlin to illustrate the methods outlined in this chapter, since Merlin is simple to use, has good documentation, can be used in a variety of computing environments, is very fast, and can be used to carry out a wide range of different analyses. The principles we discuss are general, and many of the methods can be implemented similarly in other software; although each program has its own particular features, there is also a degree of consistency in file formats used.

The methods behind Merlin are described in detail elsewhere (1) but the program uses a fast algorithm based on sparse trees to represent the flow of genes through pedigrees. The methodology enables Merlin to handle large numbers of markers such as are found in more recent linkage analyses based on single nucleotide polymorphisms (SNPs) (*see* **Section 3.3**). Besides parametric and non-parametric linkage analysis of binary traits (**Sections 3.4** and **3.1**, respectively), Merlin can be used for quantitative trait linkage analysis (**Section 3.2**) and for simulation and has additional capabilities not discussed here such as error detection and haplotype estimation.

The methods described here are illustrated by applying them to a simulated dataset derived from a large affected sibling pair study of cardiovascular disease (13). The study design was to collect pairs of siblings both of whom were affected by cardiovascular disease before age 66 years. Altogether over 4000 individuals were collected in 1933 families; parents were not genotyped. Some families had three or more affected siblings, some siblings were found to be half-siblings and recoded as such, and a small number of multi-generational families were identified. For the purposes of this analysis, data were used from chromosome 10, including the marker map and independently ascertained allele frequencies for the 20 microsatellite markers used in the original study. Using the simulation facility in Merlin, a disease locus was simulated at position 120 cM, between markers 13 and 14, with a genetic relative

risk of 2 (penetrances assumed to be 0.001, 0.002 and 0.004 for the three genotypes) and allele frequency of 0.1. Genotypes were simulated conditional on the observed affection status of individuals in the family, preserving the family structures and pattern of missing genotype data in the original study. For the analysis described in **Section 3.2**, Merlin was again used to simulate a quantitative trait using the same family structures and patterns of missing data. We assumed that 20% of the variance of the trait was accounted for by a SNP, with minor allele frequency 0.3, again at position 120 cM, with polygenic effects explaining in total 30% of the variance (The locus-specific effect is actually much stronger than we might expect to find, but quantitative trait linkage analysis based on this study design would have low power to detect

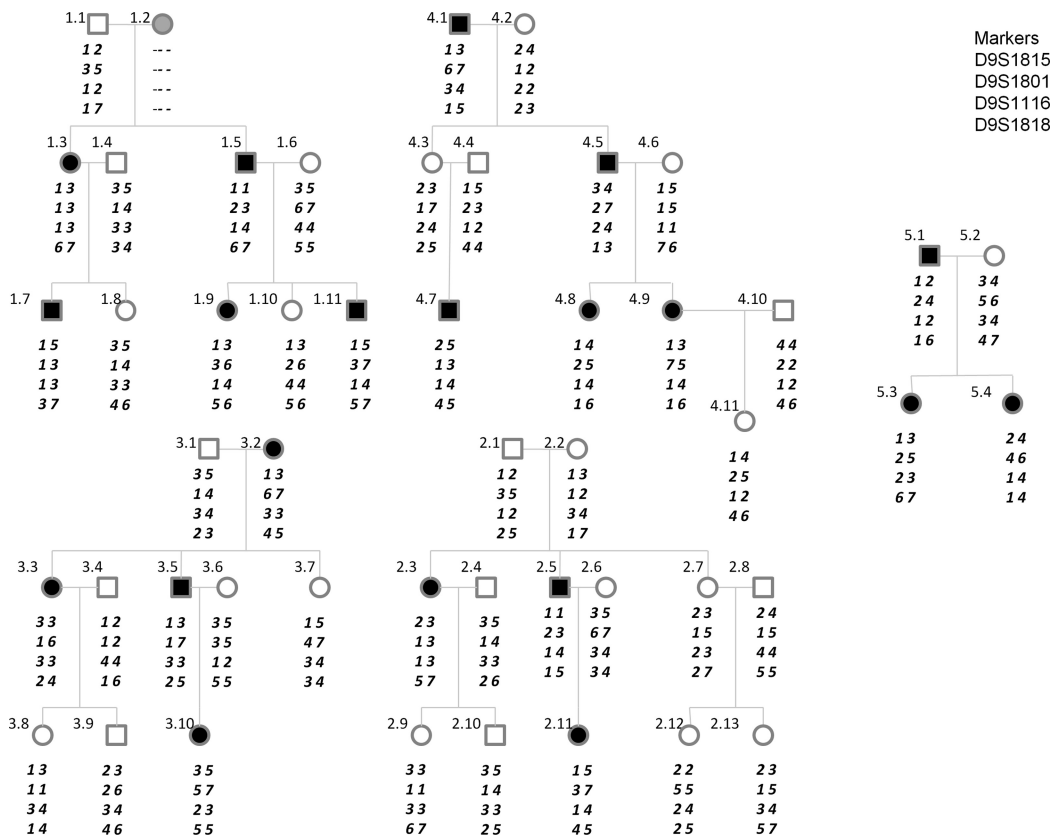


Fig. 2.1. The figure displays five families segregating a rare dominant disorder labelled ACC. *Squares* represent males and *circles* represent females. *Black-filled shapes* indicate the person is affected with the disorder. *Non-filled shapes* indicate the person is not affected, and grey shading indicates the status with respect to the phenotype is unknown. Below each person is a column of genotypes for the chromosome 9 markers D9S1815, D9S1801, D9S1116 and D9S1818. Information about these and other microsatellite markers can be found at <http://research.marshfieldclinic.org/genetics>. The index number to the *top left* of each individual is of the form 'x.y', where 'x' is the family ID and 'y' is the individual ID within the pedigree. The person labelled 1.2 is of unknown phenotype and unknown genotype. She is still required to be entered into the analysis so that 1.3 and 1.5 are correctly analysed as full siblings.

a weaker effect.). The simulated datasets are available for download (details can be found at http://limm.leeds.ac.uk/research_sections/epidemiology_and_biostatistics/groups/barrett.htm).

The dataset simulated for the parametric example (**Fig. 2.1**) consists of five pedigrees segregating a rare autosomal dominant trait. Members of the families have been genotyped at four microsatellite markers on chromosome 9. The column of numbers below each person lists the genotype observed at each marker.

3. Methods

3.1. Non-parametric Linkage Analysis

Non-parametric or *model-free* linkage analysis refers to the investigation of linkage without specification of a disease model. As explained earlier, regardless of the mode of inheritance, pairs of relatives affected by the disease are expected to show greater sharing of haplotypes that are IBD in the region of the disease gene. Various methods test whether IBD sharing at a locus is greater than expected under the null hypothesis of no linkage. Non-parametric linkage analysis requires family data where more than one individual in the family is affected, but it does not require large pedigrees. It is suitable for complex diseases where risk is influenced by a number of genes, usually in addition to environmental factors.

The simplest approach is to study sibling pairs, both of whom are affected. At any locus, according to the null hypothesis of no linkage, the number of alleles shared IBD by a pair of siblings is 0, 1 or 2, with probabilities 0.25, 0.5 and 0.25, respectively. If IBD sharing in the families is known, evidence for excess sharing at any locus can be tested by comparing the observed proportions with these expectations. In practice, IBD sharing usually has to be estimated, either because parental genotypes are unknown or because the markers are not sufficiently polymorphic.

There are several steps to carrying out linkage analysis:

1. Create files for analysis

The key information is contained in a pedigree file, which includes the pedigree structure and all individual genotypes. The file contains one row for every individual in the pedigrees, including those with no genotype information but whose offspring are included. The format of the file is common to most linkage analysis software and contains the following fields: *Pedigree*, *Individual*, *Father*, *Mother*, *Sex*, followed by affection status and genotype information.

The first four columns contain identifiers for the family, the individual, their father and their mother. For founders

(first generation in the pedigree), the parental codes are set to 0. The following field by convention contains the individual's sex (1 for male, 2 for female and 0 for unknown), generally followed by a field for affection status (1 for unaffected, 2 for affected and 0 for unknown). The genotype information follows, recording the two alleles for each marker, either in two separate columns or concatenated (e.g. "1/3"). An example of such a file can be found in *chr10-grr2.ped*.

Merlin also requires a file containing information about the structure of the pedigree file. This allows for more flexibility so that, for example, quantitative phenotypes can be included as well as or instead of affection status and the order of the fields can be specified (*see Section 3.2*). This basic data file contains descriptions of the fields in the pedigree file; the first five columns up to and including sex are taken as read. Each line begins either "A" (for affection), "T" (for quantitative trait) or "M" (for marker) and is followed by the name of the disease, trait or marker respectively. An example of such a file is *chr10-grr2.dat*, which describes a file such as *chr10-grr2.ped*, with 20 markers and affection status (named "disease").

Information may also be required about the markers genotyped, in the form of a genetic map and allele frequencies. In Merlin the marker map can consist of three columns listing for each marker the chromosome, the marker name and the position (in cM), headed "CHR," "MARKER" and "POS" (*see chr10-replicate.map*).

2. Check data integrity

Prior to running analyses it is advisable to check that the data are as expected. This can be done using the *pedstats* program (14) (available with Merlin):

```
>pedstats -d chr10-grr2.dat -p chr10-grr2.ped
```

The program output confirms that there are 8352 individuals, consisting of 4071 founders and 4281 non-founders, in 1983 pedigrees, mainly (99.7%) of 2 generations (the remainder with 3 generations), and with family size ranging from 3 to 9. Most pedigrees (81.4%) are of size 4, and 4223 individuals (50.6%) are affected, reflecting the affected sibling pair ascertainment scheme. For each marker the proportion of subjects genotyped is given, together with the proportion of founders genotyped. Finally heterozygosity is calculated, which reflects the degree of polymorphism and hence informativeness of the marker.

3. Analysis options

Before carrying out the linkage analysis various choices are to be made regarding method of analysis:

a. Single-point or multipoint analysis

In single-point analysis each marker is analysed separately. It is generally preferable to use multipoint analysis, which makes use of the marker map and the information from all markers in the region to estimate IBD sharing, potentially more accurately (*see Note 1*). For multipoint analysis the points along the chromosome at which IBD sharing is estimated can be specified either as the number of points between markers (`--steps n`) or as equally spaced points (`--grid n` for every n cM).

b. Allele frequency estimation

Allele frequencies can either be obtained from an independent source and supplied in a file or be estimated from the dataset (*see Note 2*). The allele frequency file contains two rows for each marker: one giving the marker name preceded by “M” and one listing the frequencies in order preceded by “F” (*see chr10.freq*). An alternative format, which may be preferable for highly polymorphic markers, is illustrated for the same markers in *chr10-grr2.freq*. If this information is not provided then estimates are obtained by counting across all individuals (the default), across all founders (`-ff` option) or by maximum likelihood (`-fm`).

c. Statistical analysis

The original idea behind non-parametric linkage analysis is to take pairs of related affected individuals and compare their (estimated) IBD sharing with the expected distribution under the null hypothesis. Each pedigree is assigned a score that measures IBD sharing, and the test for linkage is based on comparing this score with the expected score according to the null hypothesis (combining over pedigrees) (*see Note 3*). If IBD information is complete, then, under the null hypothesis, the resulting non-parametric-linkage test statistic is normally distributed with mean of 0 and variance of 1, for large enough samples sizes. In the absence of complete information, the test as initially proposed is conservative. In response to this, the approach has been modified to provide accurate likelihood-based tests which are implemented in Merlin (*see Note 4*).

4. Running the analysis

Once the input files have been correctly constructed, carrying out the analysis is simple and fast. The command below, for example, carries out a multipoint non-parametric linkage analysis based on the IBD sharing among all affected individuals in each family, using allele frequencies specified in *chr10-grr2.freq*:

```
merlin -d chr10-grr2.dat -p chr10-grr2.ped -m chr10-  
grr2.map -f chr10-grr2.freq --npl --steps 3 --tabulate --pdf  
--prefix grr2
```

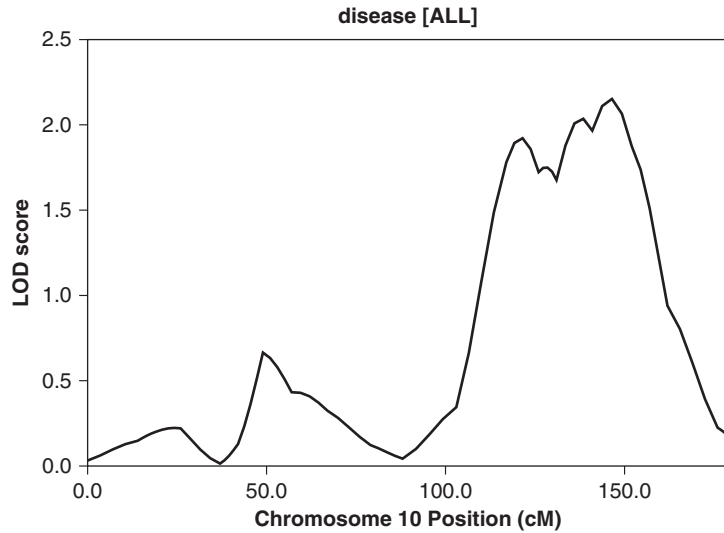



Fig. 2.2. Output from Merlin showing LOD score along the chromosome from non-parametric linkage analysis of a binary disease trait.

The main output from the program is now available in a file named *grr2-nonparametric.tbl* and the graphical representation in **Fig. 2.2** is saved to file *grr2.pdf*.

5. Interpretation of output

At each point along the chromosome at which analysis has been requested (e.g. in the above example at each marker and 3 points between them, as specified by the **step 3** option), four measures are given: the Z-score, delta, LOD score, and *p* value. The Z-score is based on the non-parametric linkage statistic as proposed by Whittemore and Halpern (15) which should follow a standard normal distribution under the null hypothesis of no linkage. Delta is the parameter of interest in the allele-sharing model proposed by Kong and Cox (16) (*see Note 4*), taking the value 0 under the null hypothesis; and the LOD score and *p* value refer to tests of this hypothesis.

For the example above, **Fig. 2.2** shows that the linkage peak is quite broad. The maximum LOD score of 2.15 ($p = 0.0008$) is attained at position 146.5 cM, between markers D10S1693 and D10S587 (*see Note 5*). This is simulated data, and in fact the true location of the marker is at 120 cM, where the evidence for linkage is slightly weaker (LOD score ~ 1.9 , $p \sim 0.002$). It is common to find in linkage analysis that the highest signal occurs some distance from the disease locus (17).

The conclusion of this analysis would be that there is some suggestive evidence of linkage in the broad region covered by the

peak in **Fig. 2.2**, although the evidence is not significant at the genome-wide level (*see Note 6*).

3.2. Quantitative Traits

Analysis of a quantitative trait can also be carried out without assuming a genetic model, and various different approaches to this have been proposed and are implemented in Merlin. The variance components approach is applicable to general pedigrees; the trait covariance matrix between relatives in the pedigree is modelled as a component due to a specific chromosomal region (on the basis of estimated IBD sharing at the locus) and a component due to other unlinked genes (on the basis of degree of kinship, see for example Almasy and Blangero (12)).

For sibling pairs, the classical Haseman–Elston method was based on regression of the squared difference in trait values between the two siblings on the estimated proportion of alleles they share IBD at a locus (18); numerous extensions and variations of this method have been proposed, and Merlin includes a separate regression program (Merlin-Regress) that implements one of these (19). In this approach, multivariate regression is used to regress the estimated IBD sharing among all pairs of relatives in the pedigree on the squared difference and the squared sum of trait values of the relative pairs.

To carry out linkage analysis of a quantitative trait, most of the steps are similar to those described above for binary traits. The pedigree file will now contain the following fields: *Pedigree*, *Individual*, *Father*, *Mother*, *Sex*, *Trait*, in addition to genotype information and possibly affection status. As before another file must also be constructed describing the pedigree file; the file *chr10-qt1.dat* for example describes the structure of *chr10-qt1.ped*. Data integrity can be checked as before using *pedstats*, which now also reports the minimum, maximum, mean and variance of the trait values in the pedigree file and an estimate of correlation between siblings.

For the statistical analysis, variance components analysis can be selected using the *--vc* option:

```
merlin -d chr10-qt1.dat -p chr10-qt1.ped -m chr10-qt1.map -f chr10-qt1.freq --vc --steps 3 --tabulate --pdf --prefix traitvc
```

The on-screen output reports that overall heritability of the trait is estimated to be 33.1% (true value in simulation 30%). At each analysis point four measures are estimated: H², which is an estimate of locus-specific heritability, the chi-squared statistic, corresponding LOD score and associated *p* value (for a one-sided test of heritability greater than zero). From **Fig. 2.3** it can be seen that the linkage peak is at 117 cM, quite close to the true locus at 120 cM; the estimated locus-specific heritability here is 15.7% (true value 20%), *p* = 0.004.

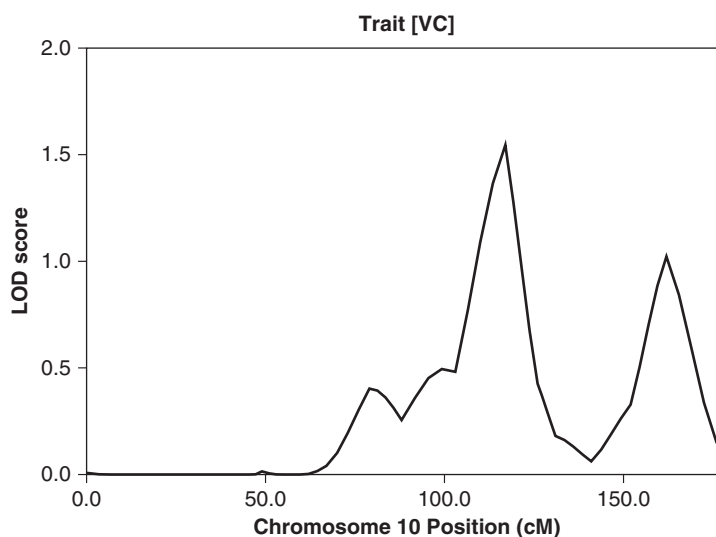


Fig. 2.3. Output from Merlin showing LOD score from quantitative trait linkage analysis.

A disadvantage of the variance components approach is that the method is inappropriate for samples selected on the basis of their phenotype; the method implemented in Merlin-regress by contrast is robust to sample selection. Using the same data files, the analysis can be run using the following command:

```
merlin-regress -d chr10-qtl.dat -p chr10-qtl.ped -m chr10-qtl.map  
-f chr10-qtl.freq --steps 3 --randomSample --tabulate --pdf --prefix  
traitregress
```

For this dataset of unselected samples and sibling pairs, the methods give almost identical results, but the regression analysis has the advantage of being many times faster.

For a sample selected on phenotype, “--randomSample” would be replaced by estimates of the mean, variance and heritability of the trait in the population, e.g.:

```
merlin-regress -d chr10-qtl.dat -p chr10-qtl.ped -m chr10-qtl.map -f  
chr10-qtl.freq --steps 3 --mean0.0 --var1.0 --her 0.3 --tabulate --pdf  
--prefix traitregress
```

The choice of appropriate method may also be influenced by the distribution of the trait (*see* **Note 7**).

3.3. Linkage Analysis Using Single Nucleotide Polymorphisms

Linkage analysis is now often carried out using a dense set of SNPs instead of a smaller number of microsatellite markers (20). Although the basic methods are no different, one complication that arises is that the SNP markers are likely to be in linkage disequilibrium (LD). It has been shown that ignoring the LD between SNPs can lead to false positive linkage signals (21).

In Merlin, this problem is approached by creating clusters of contiguous correlated markers and then assuming that there is no recombination within the clusters and no LD between clusters. These assumptions will only be approximately correct; data which violate the assumptions, such as obligatory recombination events within the clusters, can be analysed by setting such genotypes to missing.

Clusters can be created either based on distance (i.e. any markers within a very small distance of each other are formed into one cluster, and the map file is adjusted so that these markers are mapped to the same location) or on the r^2 measure of LD between them. The clusters can be defined independently or generated from the data by Merlin.

3.4. Parametric Linkage Analysis

As in the examples above, the genetic markers are assumed to be mapped accurately, so the order and distances between each locus are specified in the ‘.map’ file and the evidence for linkage is computed at many locations within the range of the genetic markers. The format of the pedigree (.ped) file is the same as before. The individuals are coded as affected (2), unaffected (1) and unknown (0). Here it is assumed that the marker allele frequencies have been independently estimated, but in this example estimating the alleles from the data will make very little difference as only one person is not directly genotyped. The parametric model is assumed to be rare dominant (labelled ‘ACC’ in .model file) with allele frequency 0.005. There is no sporadic rate, and carriers have incomplete penetrance of 90%. The analysis is performed with the command below:

```
merlin --d rd5.dat --p rd5.ped --m rd5.map --f rd5.freq --model rd5.model --step 3
```

The output is listed on the screen. The default is to list both LODs and HLODs. The ‘**Step 3**’ option delivers 3 evenly spaced Scores per marker. The -perFamily option prints the LOD scores by family to a file.

Parametric Analysis, Model rare-dom

Position	LOD	Alpha	HLOD
99.400	-11.483	0.150	0.243
103.718	-3.177	0.299	0.523
108.035	-2.206	0.386	0.819
112.353	-2.061	0.429	1.082
116.670	-3.642	0.448	1.316
120.133	1.414	0.726	1.967

123.595	2.238	0.789	2.632
127.058	2.622	0.796	3.071
130.520	2.478	0.780	3.393
135.620	2.945	0.832	3.178
140.720	2.644	0.848	2.807
145.820	1.851	0.816	2.105
150.920	-3.713	0.174	0.111

The maximum LOD score of 2.945 is reported at position 135.620. When allowing for heterogeneity, the maximum HLOD is reported one step away at 130.520. Allowing for heterogeneity, there appears to be marginally significant evidence of linkage to this region. The maximum likelihood estimate of the proportion linked is 78%, and the strongest evidence is coincident with marker D9S1116. The true location of the risk locus in the simulations was between the two markers D9S1801 and D9S1116 (position 125.00), and two of the five families were unlinked to this locus. Examining the individual by family LOD scores at this position, it can be seen which families provide the evidence for and against linkage. Examining the LOD scores, post analysis, by family is useful for identifying those who are currently not very informative and may benefit from more genotyping at intervening markers. In this dataset, family 5 shows strong evidence against linkage as the two affected offspring do not share any alleles IBD (Fig. 2.1).

4. Notes

1. In the Introduction linkage analysis is described with respect to two loci, but it is now usually performed in multipoint form (multiple markers per chromosomal region) as this provides much better information on the origin of each chromosomal segment segregating through a family, and much better estimates of IBD sharing. Multipoint linkage analysis is quite sensitive to the correctness of the map on which it is based, so if there are doubts about the accuracy of the map it may be wise to compare results from multipoint and single-point analyses.
2. If good estimates of allele frequencies, applicable to the population from which the families are drawn, can be obtained independently, then these should be used. In the absence of such estimates, frequencies can be estimated from the dataset itself, especially if the sample size is large. However this can

lead to a conservative test for linkage, since the frequencies of any alleles associated with disease will tend to be overestimated.

3. When larger numbers of affected relatives are included within a family, a more powerful alternative to pairwise analysis has been proposed, which considers the sharing of alleles among all affected relatives in the family. This is based on a score that increases more sharply as the number of affected members sharing the same allele IBD increases (15).
4. Kong and Cox (16) proposed an alternative approach to overcome the conservative nature of the “NPL” scores in the presence of incomplete genotype data. For any of the proposed scores, a one-parameter model can be constructed, the free parameter (δ) of which is chosen such that $\delta = 0$ under the null hypothesis of no linkage and $\delta > 0$ in the presence of linkage. The test of $\delta = 0$ is carried out by a likelihood ratio test, and can be converted to a traditional \log_{10} LOD score, for comparability with parametric methods. Two versions of the model are proposed, known as the linear and the exponential models. In most situations very similar results are obtained from the two approaches and the linear model would be used; the exponential model allows δ to take large values, and may be preferable given a small number of pedigrees with extreme IBD sharing.
5. Slightly different results are obtained from the Z -score, where the peak is at 140 cM with a Z -score of 2.51 (p value 0.006). In general, the LOD score analysis is to be preferred.
6. As mentioned in the Introduction, evidence for linkage can now be evaluated by calculating empirical p values, which avoids the need to agree LOD score thresholds for declaring significant evidence of linkage. Data are simulated under the null hypothesis of no linkage preserving the original data structure, and the LOD score is compared with the distribution of LOD scores obtained from the analysis of many such simulated datasets to obtain an empirical p value.
7. Although both the above methods assume normality of the trait, the regression method is more robust than variance components methods to departures from normality. An alternative approach also available in Merlin is an extension of the non-parametric methods for binary data and is based on comparing the IBD sharing among individuals at the tails of the trait distribution (*see* the `--qtl` option). This method is less commonly used and has the disadvantage of relatively low power but the advantage of avoiding distributional assumptions. When applied to the example dataset this method provides very little evidence for linkage.

References

1. Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30, 97–101.
2. Ott, J. (1999) *Analysis of Human Genetic Linkage*, Second ed. The Johns Hopkins University Press, Baltimore, MD.
3. Teare, M. D., and Barrett, J. H. (2005) Genetic epidemiology 2 – Genetic linkage studies. *Lancet* 366, 1036–1044.
4. Chiano, M. N., and Yates, J. R. W. (1995) Linkage detection under heterogeneity and the mixture problem. *Annals of Human Genetics* 59, 83–95.
5. Lander, E., and Kruglyak, L. (1995) Genetic dissection of complex traits – guidelines for interpreting and reporting linkage results. *Nature Genetics* 11, 241–247.
6. Cannings, C., and Thompson, E. A. (1981) *Genealogical and Genetic Structure*. Cambridge University Press, Cambridge.
7. Lathrop, G. M., and Lalouel, J. M. (1984) Easy calculations of lod scores and genetic risks on small computers. *American Journal of Human Genetics* 36, 460–465.
8. Cottingham, R. W., Idury, R. M., and Schaffer, A. A. (1993) Faster sequential genetic-linkage computations. *American Journal of Human Genetics* 53, 252–263.
9. Kruglyak, L., Daly, M. J., ReeveDaly, M. P., and Lander, E. S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* 58, 1347–1363.
10. Gudbjartsson, D. F., Jonasson, K., Frigge, M. L., and Kong, A. (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* 25, 12–13.
11. Wijsman, E. M., Rothstein, J. H., and Thompson, E. A. (2006) Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. *American Journal of Human Genetics* 79, 846–858.
12. Almasy, L., and Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* 62, 1198–1211.
13. The BHF Family Heart Study Research Group (2005) A genomewide linkage study of 1933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) family heart study. *American Journal of Human Genetics* 77, 1011–1020.
14. Wigginton, J. E., and Abecasis, G. R. (2005) PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 21, 3445–3447.
15. Whittemore, A. S., and Halpern, J. (1994) A class of tests for linkage using affected pedigree members. *Biometrics* 50, 118–127.
16. Kong, A., and Cox, N. J. (1997) Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics* 61, 1179–1188.
17. Roberts, S. B., MacLean, C. J., Neale, M. C., Eaves, L. J., and Kendler, K. S. (1999) Replication of linkage studies of complex traits: an examination of variation in location estimates. *American Journal of Human Genetics* 65, 876–884.
18. Haseman, J. K., and Elston, R. C. (1972) Investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* 2, 3–19.
19. Sham, P. C., Purcell, S., Cherny, S. S., and Abecasis, G. R. (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics* 71, 238–253.
20. Schaid, D. J., Guenther, J. C., Christensen, G. B. et al (2004) Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. *American Journal of Human Genetics* 75, 948–965.
21. Huang, Q., Shete, S., and Amos, C. I. (2004) Ignoring linkage disequilibrium among tight linked markers induces false-positive evidence of linkage for affected sib pair analysis. *American Journal of Human Genetics* 75, 1106–1112.



<http://www.springer.com/978-1-61779-175-8>

In Silico Tools for Gene Discovery

Yu, B.; Hinchcliffe, M. (Eds.)

2011, XI, 365 p., Hardcover

ISBN: 978-1-61779-175-8

A product of Humana Press