Performing

linkage analysis using MERLIN

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

Overview

- MERLIN and associated programs
- Error checking
- Parametric linkage analysis
- Nonparametric linkage analysis
- Variance components linkage analysis
- Calculating IBD for other programs
- Combination of SNPs in linkage disequilibrium

Goncalo Abecasis's excellent program.

506 journal citations since 2002.

```
[davidD@bioinfo ~]$ merlin
MERLIN 1.0.1 - (c) 2000-2005 Goncalo Abecasis

References for this version of Merlin:

   Abecasis et al (2002) Nat Gen 30:97-101        [original
citation]
   Fingerlin et al (2004) AJHG 74:432-43          [case selection for
association studies]
   Abecasis and Wigginton (2005) AJHG 77:754-67   [ld modeling,
parametric analyses]
```

Genetic linkage analysis: MERLIN

- For linkage analysis of binary or quantitative traits and many markers
- Small to moderately large families

Performs:

- parametric and non-parametric linkage analysis
- variance components linkage analysis of quantitative traits
- regression-based analysis of quantitative traits: MERLIN-REGRESS
- multimarker-based ibd and kinship estimation
- haplotyping
- error detection
- simulation of marker data under null hypothesis of no linkage

Linkage disequilibrium between markers is allowed in models

Limitations

- Uses the Lander-Green approach to multipoint linkage, so not suitable for large pedigrees (>30 “bits”)
- Maximizer for variance components analysis a little slow if multiple fixed effects included, and may sometimes get stuck

All the MERLIN Commands

```
[davidD@bioinfo ~]$ merlin
…
The following parameters are in effect:
                     Data File :      merlin.dat (-dname)
                  Pedigree File :      merlin.ped (-pname)
             Missing Value Code :         -99.999 (-xname)
                      Map File :      merlin.map (-mname)
             Allele Frequencies : ALL INDIVIDUALS (-f[a|e|f|m|file])
                    Random Seed :           123456 (-r9999)

Data Analysis Options
            General : –error, –information, –likelihood, –model
[param.tbl]
          IBD States : –ibd, –kinship, –matrices, –extended, –select
         NPL Linkage : –npl, –pairs, –qtl, –deviates, –exp
          VC Linkage : –vc, –useCovariates, –ascertainment
         Haplotyping : –best, –sample, –all, –founders, –horizontal
       Recombination : –zero, –one, –two, –three, –singlepoint
           Positions : –steps, –maxStep, –minStep, –grid, –start,
–stop
     Marker Clusters : –clusters [], –distance, –rsq, –cfreq
              Limits : –bits [24], –megabytes, –minutes
         Performance : –trim, –noCoupleBits, –swap, –cache []
              Output : –quiet, –markerNames, –frequencies, –perFamily,
–pdf,
                       –prefix [merlin]
          Simulation : –simulate, –reruns, –save
```

The MERLIN pedigree file ".ped"

There are three essential files:

The pedigree file format is consistent with the GAS (Sib-pair) or LINKAGE formats:

```
01927 0192703 x       x       m  0    x         x      1936.3404 1/2 1/1 1/1
01927 0192704 x       x       f  0    x         x      1938.9885 1/2 1/1 1/1
01927 0192701 0192703 0192704 m MZ    2.3300   4.0943  1968.0000 x/x x/x x/x
01927 0192702 0192703 0192704 f MZ    2.1700   5.3293  1968.0000 1/2 1/1 1/1
01927 0192750 0192703 0192704 f  0    x        3.8067  1966.0000 1/2 1/1 1/1
01940 0194003 0       0       1  0    -         -      1918.7665 0   0 0   0 0   0
01940 0194004 0       0       2  0    -         -      1921.4147 1   1 1   1 1   2
01940 0194005 0       0       1  0    -        3.4012  1946.3317 1   1 2   1 1   1
01940 0194001 0194003 0194004 2  0   1.8100    4.2806  1949.0000 1   1 1   1 1   2
01940 0194002 0194003 0194004 2  0   2.7600    5.0599  1949.0000 1   1 1   1 1   1
01940 0194008 0194005 0194001 1  0    -        3.2189  1978.0000 1   1 1   1 1   2
01940 0194009 0194005 0194001 2  0    -        4.4998  1980.0000 1   1 2   1 1   1
```

The first five fields are pedigree_ID, individual_ID, father_ID, mother_ID, sex.

A character string is treated as missing (hence the use of “x” or “-” here). Therefore alleles **must** be numeric. A “0” for a marker or “-99.999” are also missing data tokens.

The slash between alleles at a genotype are optional.

The MZ twin indicator is “MZ” in that column.

The MERLIN locus file ".dat"

```
Z mztwin
T aat
T lige2
C yob
M rs1303
M E366K
M rs17580
M rs6647
S2 rs2753934
M rs1980618
```

The first column defines the variable type: "S" and "S2" skip one or two columns of data, "A" is a binary trait (taking values "[12yn]").

The second column gives the variable name

The MERLIN map file ".map"

```
14 rs1303      93.915
14 E366K       93.915
14 rs17580     93.917
14 rs6647      93.917
14 rs709932    93.919
14 rs17090730 93.920
```

The columns are chromosome, marker_name, map_position (in cM).

A typical call to MERLIN

```
> merlin -d chr1.dat -m chr1.map -p chr1.ped –grid 10 –vc
–pdf

…
Family: 42258 - Founders: 2  - Descendants: 2  - Bits: 2
  Skipping Marker D11S2008_S [BAD INHERITANCE]
  Skipping Marker ATA27C11_M [BAD INHERITANCE]

Family: 99008 - Founders: 2  - Descendants: 2  - Bits: 2
  Skipping Marker D11S1301_S [BAD INHERITANCE]
```

A typical call to MERLIN

```
Phenotype: ige1 [VC] (773 families, h2 = 48.46%)
=================================================================
            Position      H2    ChiSq     LOD  pvalue
               4.939   37.76%    5.91    1.28   0.008
              14.939   31.25%    3.83    0.83    0.03
              24.939   10.59%    0.52    0.11     0.2
              34.939    0.00%    0.00    0.00     0.5
              44.939    0.00%    0.00    0.00     0.5
              54.939    0.00%    0.00    0.00     0.5
              64.939    1.51%    0.03    0.01     0.4
              74.939    3.38%    0.10    0.02     0.4
              84.939    0.00%    0.00    0.00     0.5
```

Columns are:

- map position where lod score evaluated
- heritability due to QTL at that position (AQE model)
- Likelihood Ratio Test Statistic testing H2=0

- LRTS/(2 log(10))
- One-sided P-value (as negative heritability not legal, H_0 on bound)

Error checking (unlikely double recombinants)

Two or more close recombination events on the same chromosome are uncommon, due to the chance distribution of chiasmata, and because of interference. Broman and Weber [2000] suggest the probability of a double recombinant within a 20 cM interval in humans is only 2 in 1000. It is more likely that there is a genotyping error in one of the markers used to infer the location of the recombination event.

```
> merlin -d chr1.dat -m chr1.map -p chr1.ped –error
```

```
Family: 83520 - Founders: 2  - Descendants: 2  - Bits: 2
  D15S205 genotype for individual 8352001 is unlikely
[2.016e-02]
  D15S205 genotype for individual 8352002 is unlikely
[2.016e-02]

Family: 85060 - Founders: 2  - Descendants: 3  - Bits: 4
  D15S978 genotype for individual 8506002 is unlikely
[2.410e-02]
```

Parametric Linkage Analysis

This is a new addition to MERLIN's functionality.

One needs an additional file of penetrances and trait allele frequencies:

```
disease 0.01 0.0,0.5,0.5 dominant
disease 0.10 0.0,0.0,0.5 recessive
```

Each line is trait, allele_frequency, penetrances, model_name.

The job is run as:

```
> merlin -d chr1.dat -m chr1.map -p chr1.ped –model
pkd.model
```

For multipoint analysis in small pedigrees, this is very fast.

Nonparametric Binary Trait Linkage Analysis

MERLIN can carry out the Kong and Cox nonparametric affected-pedigree-member analysis using either the “pairs” or “all” statistics, and using either the linear or exponential models.

The job is run as:

```
> merlin -d chr1.dat -m chr1.map -p chr1.ped –npl –pairs
```

Haplotyping

The default algorithm for this assumes there is no linkage disequilibrium between markers.

MERLIN can be requested to combine adjacent SNPs into haplotypes where the r^2 is above a threshold. This is a necessary preliminary to using closely associated SNPs for linkage analysis, as missing parental genotypes can be wrongly inferred if LD is neglected.