Univariate Linkage in Mx

Boulder, TC 18, March 2005

Posthuma, Maes, Neale

VC analysis of Linkage

Incorporating IBD Coefficients

Covariance might differ according to sharing at a particular locus.
Sharing at a locus can be quantified by the estimated proportion of alleles shared IBD:

$$\hat{\pi} = p_{IBD=2} + 0.5 \times p_{IBD=1}$$

Variance-Covariance Matrix

$$\Sigma_{jk} = \begin{cases} \sigma_q^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 & \text{if } j = k \\ \hat{\pi}\sigma_q^2 + \rho\,\sigma_a^2 + \sigma_c^2 & \text{if } j \neq k \end{cases}$$

Where,

ρ is twice the kinship coefficient, (i.e. twice the probability that two genes sampled at random from a pair of individuals are identical. 1 for MZ twins and 0.5 for DZ twins)

$\hat{\pi}$ depends on the number of alleles shared IBD

j and *k* index different individuals from the same family

q = variation due to the QTL; *a* = polygenetic variation; *e* = individual-specific variation, *c* = shared environmental variation

Alternate hypothesis of linkage for sibpairs (Likelihood function):

$$L_1 = \prod_i (2\pi)^{-1} |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})}$$

$$\Sigma = \begin{bmatrix} \sigma_q^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 & \hat{\pi}_i \sigma_q^2 + \rho_i \sigma_a^2 + \sigma_c^2 \\ \hat{\pi}_i \sigma_q^2 + \rho_i \sigma_a^2 + \sigma_c^2 & \sigma_q^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 \end{bmatrix}$$

Note three uses of 'pi'!

Null hypothesis:

$$L_0 = \prod_i (2\pi)^{-1} |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})}$$

$$\Sigma = \begin{bmatrix} \sigma_a^2 + \sigma_c^2 + \sigma_e^2 & \rho_i \sigma_a^2 + \sigma_c^2 \\ \rho_i \sigma_a^2 + \sigma_c^2 & \sigma_a^2 + \sigma_c^2 + \sigma_e^2 \end{bmatrix}$$

$2(\ln L_1 - \ln L_0)$ is distributed as a χ^2 distribution.
Dividing the χ^2 by $2\ln 10$ (~4.6) gives the LOD-score (for 1 df)

Linkage Analysis: many programs out there..

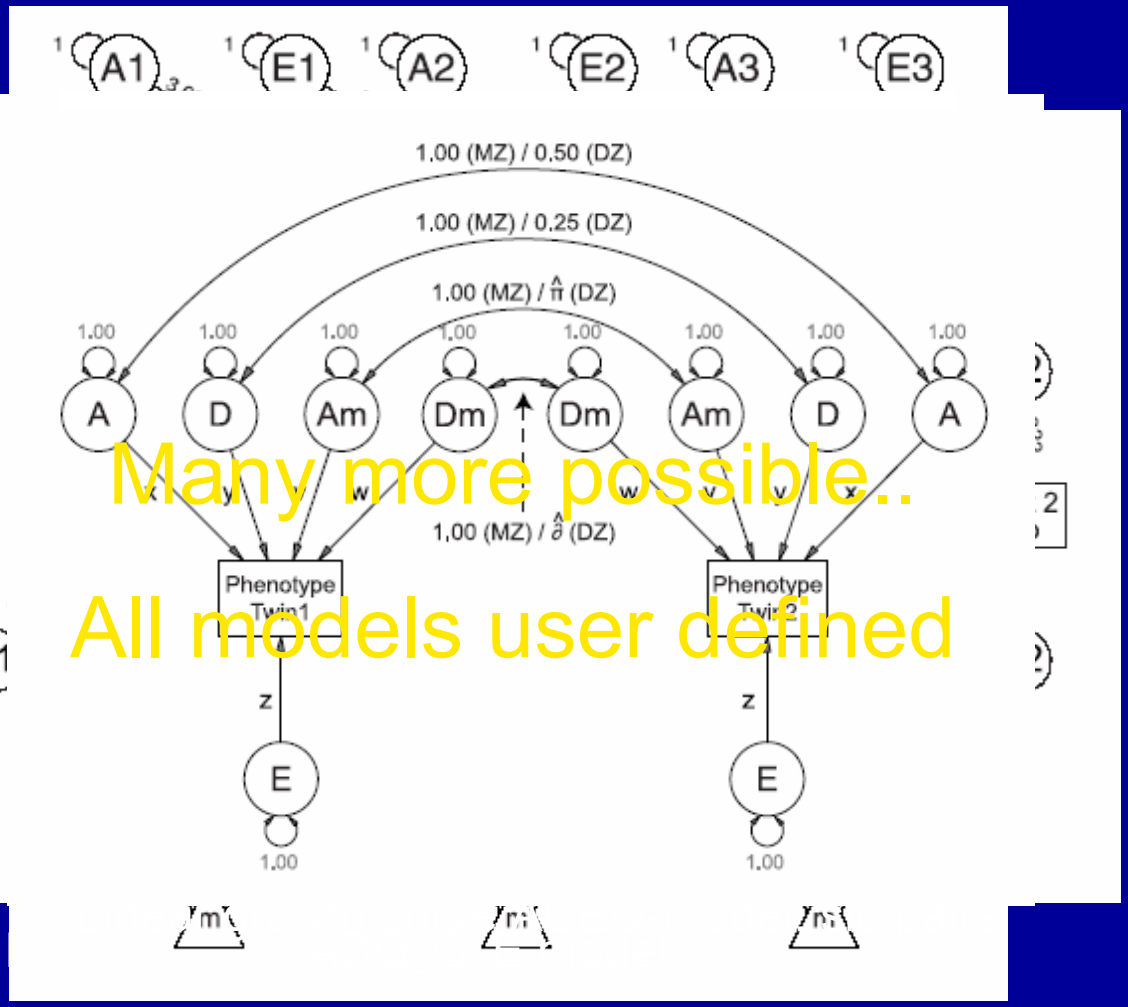
Mx vs *MERLIN*

Mx

- Does not calculate IBDs
- Model specification nearly unlimited
  - multivariate phenotypes
  - Longitudinal modelling
  - Factor analysis
  - Sample heterogeneity testing
  - …
- No Graphical output

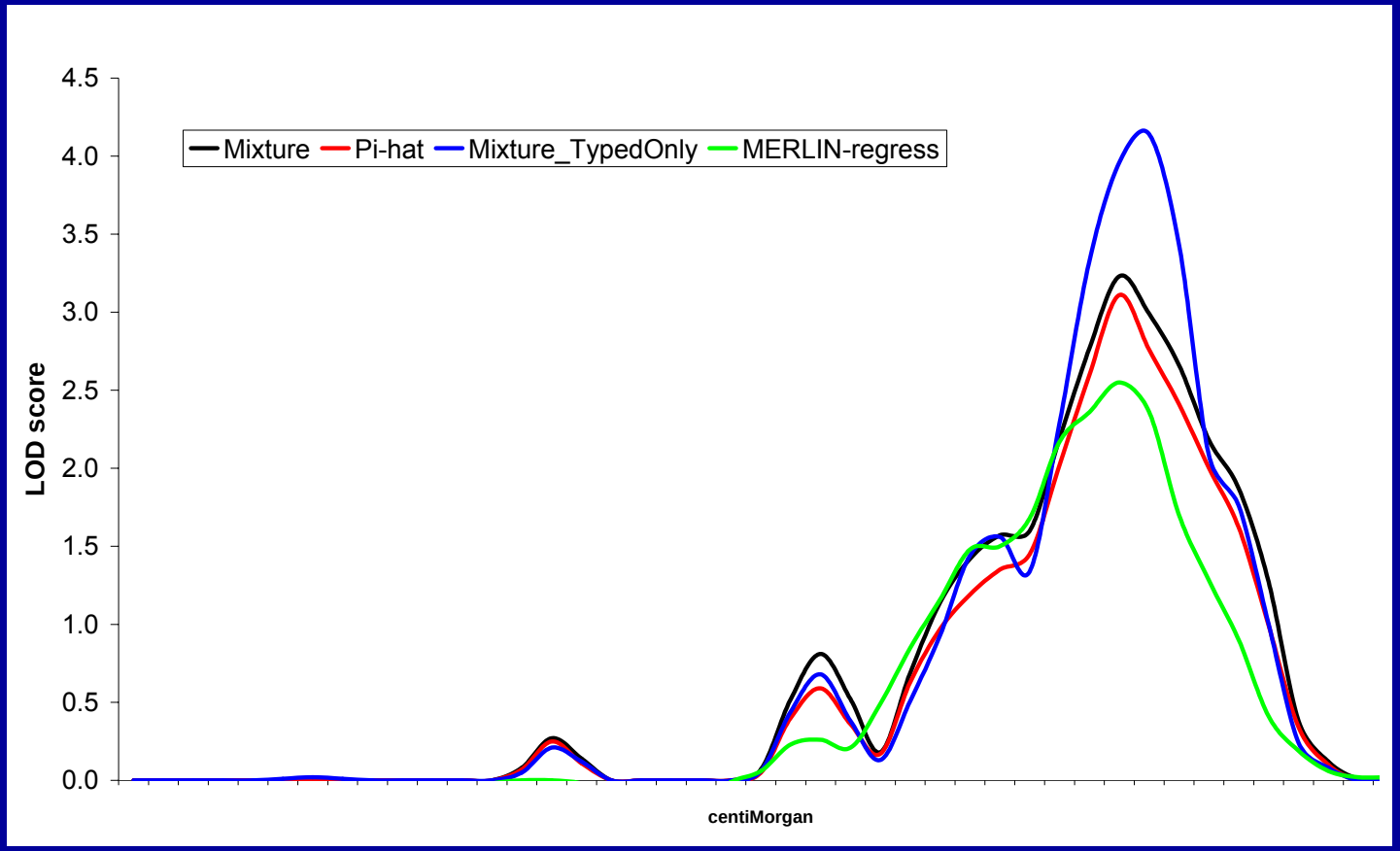
MERLIN

- Calculates IBDs
- Model specification relatively limited
- Some graphical output



Trivariate model including covariates

Example comparing different methods



How to get IBD's estimated with Merlin into Mx

Merlin ibd output:

```
FAMILY ID1 ID2 MARKER P0 P1 P2
60007 31 31 0.000  0.0 0.0 1.0
60007 41 31 0.000  1.0 0.0 0.0
60007 41 41 0.000  0.0 0.0 1.0
60007 1 31 0.000  0.0 1.0 0.0
60007 1 41 0.000  0.0 1.0 0.0
60007 1 1 0.000  0.0 0.0 1.0
60007 2 31 0.000  0.0 1.0 0.0
60007 2 41 0.000  0.0 1.0 0.0
60007 2 1 0.000  0.18899 0.54215 0.26886
60007 2 2 0.000  0.0 0.0 1.0
60007 31 31 10.000  0.0 0.0 1.0
60007 41 31 10.000  1.0 0.0 0.0
60007 41 41 10.000  0.0 0.0 1.0
60007 1 31 10.000  0.0 1.0 0.0
60007 1 41 10.000  0.0 1.0 0.0
60007 1 1 10.000  0.0 0.0 1.0
60007 2 31 10.000  0.0 1.0 0.0
60007 2 41 10.000  0.0 1.0 0.0
60007 2 1 10.000  0.14352 0.59381 0.26267
60007 2 2 10.000  0.0 0.0 1.0
….
```

Alsort.exe

Usage: alsort <inputfile> <outpfile> [-vfpm] [-c] [-i] [-t] [-x <id1> ...]

-v Verbose (implies -vfpm)
-vf Print family ID list
-vp Print marker positions
-vm Print missing p-values
-c Create output file per chromosome
-i Include 'self' values (id1=id2)
-x Exclude list; id-values separated by spaces
-t Write tab as separator character

Practical Alsort.exe

Open a dos prompt, go to the directory where alsort.exe is and type

Alsort test.ibd sorted.txt –c –x 31 41 –t

PRACTICAL: F:\danielle\UnivariateLinkage\pract_alsort

This session's example

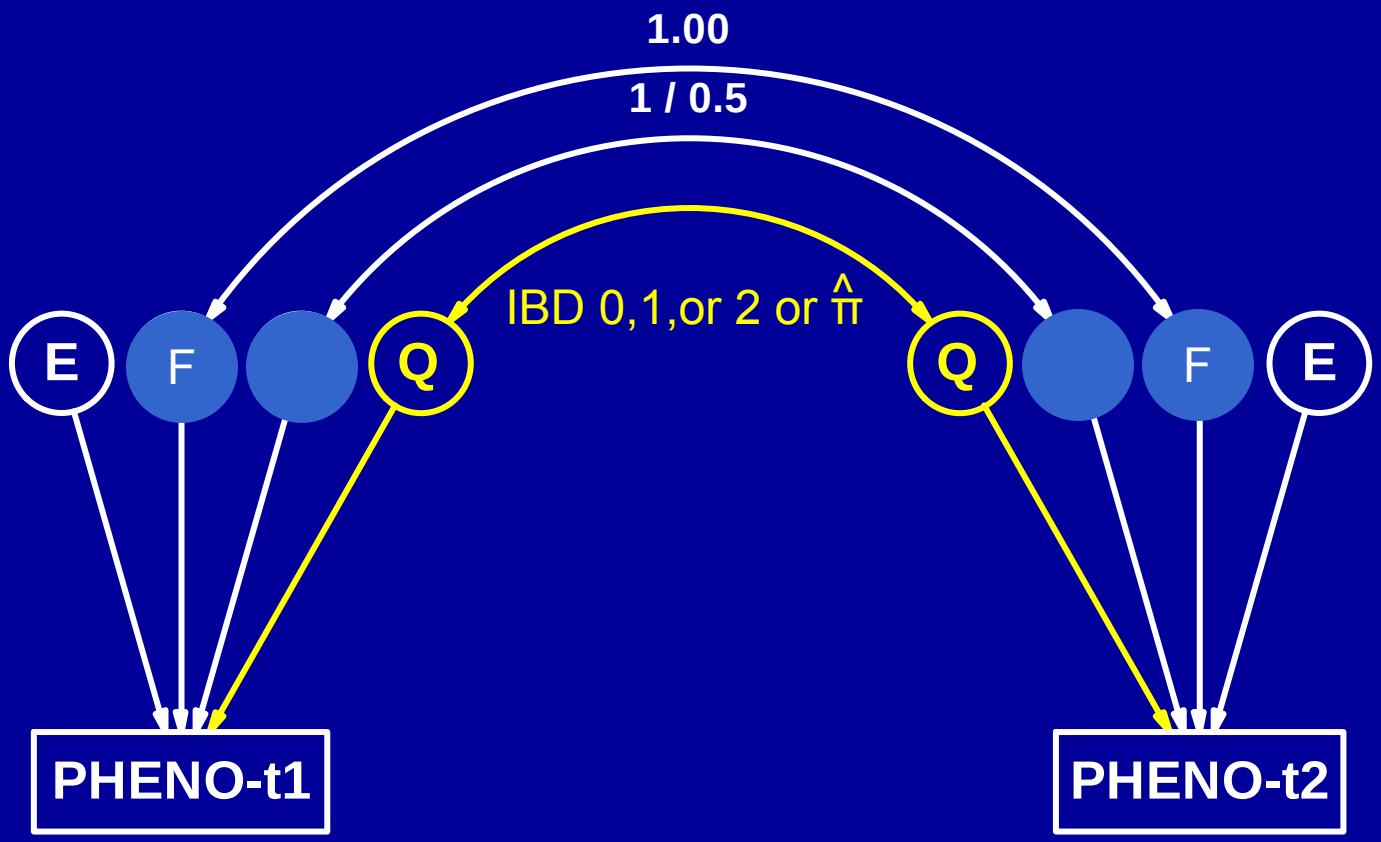
Heritabilities of Apolipoprotein and Lipid Levels in Three Countries

Twin Research April 2002

Marian Beekman[1,2], Bastiaan T. Heijmans[1], Nicholas G. Martin[3], Nancy L. Pedersen[4], John B. Whitfield[5], Ulf DeFaire[6], G. Caroline M. van Baal[7], Harold Snieder[8,9], George P. Vogler[10], P. Eline Slagboom[1], and Dorret I. Boomsma[7]

Lipid data: apoB

FEQ-model



Pi-hat

Pi-hat = 0 x p(ibd=0) +
.5 x p(ibd=1) +
1 x p(ibd=2)

Mx

```
K Full 3 1 fix                    ! ibd0 ibd1 ibd2
J Full 1 3 fix                    ! 0  .5  1

Specify K ibd0m1 ibd1m1 ibd2m1
Matrix J 0 .5  1

 P = J*K;                         ! Calculates pi-hat

Covariance
        F+E+Q              | F+P@Q   _
       F+P@Q               | F+E+Q          ;
```

Practical

- Mx script: pihat1.mx

- Change the script at the ?????
- Choose a position, run

PRACTICAL: F:\danielle\UnivariateLinkage\pract_linkage

Alternative way to model linkage

Rather than calculating pi-hat, we can fit three models (for ibd=0, 1, or 2) to the data and weight each model with its corresponding probability for a pair of siblings:

Full information approach aka Weighted likelihood or Mixture distribution approach

Full information approach

```
K full 3 1                                  ! will contain IBD probabilities
Specify K ibd0m1 ibd1m1 ibd2m1              ! put ibd probabilities in K

Covariance
        F+Q+E           |F_
        F               |F+Q+E_            ! IBD 0 matrix

        F+Q+E           |F+H@Q_
        F+H@Q           |F+Q+E_            ! IBD 1 matrix

        F+Q+E           |F+Q_
        F+Q             |F+Q+E_            ! IBD 2 matrix

Weights     K  ;
```

Practical

- (Adjust mixture1.mx to run in batch mode and) change the ?????'s and run mixture1.mx script for 3 positions
- Run pihat1.mx for 3 positions
- Calculate lod-score and write your results on the board (one for pihat, one for mixture). Null (FE) model -2ll=304.832

PRACTICAL: F:\danielle\UnivariateLinkage\pract_linkage

Pihat vs mixture

Pihat simple with large sibships
Solar, Genehunter etc

Pihat shows substantial bias with missing data

Pihat vs mixture

Example: pihat =.5
May result from

ibd0=0.33
ibd1=0.33
ibd2=0.33

or:

ibd0=0.5
ibd1=0
ibd2=0.5

So mixture retains all information wherea pihat does not

How to increase power?

Large sibships much more powerful
Dolan et al 1999

Expected IBD Frequencies

Sibships of size 2

Type	Configuration	Frequency
1	2	4/16
2	1	8/16
3	0	4/16

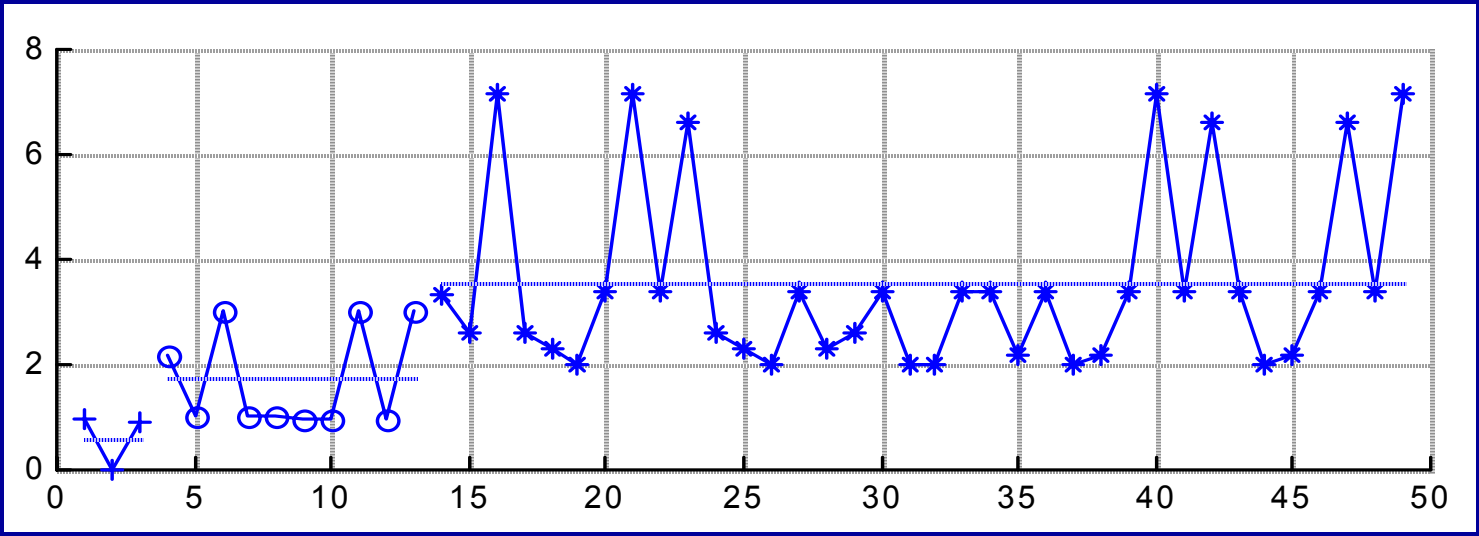
Expected IBD Frequencies

Sibships of size 3

Type	Configuration	Frequency
1	222	4/64
2	211	8/64
3	200	4/64
4	121	8/64
5	112	8/64
6	110	8/64
7	101	8/64
8	020	4/64
9	011	8/64
10	002	4/64

More power in large sibships

Dolan, Neale & Boomsma (2000)



+Size 2 o Size 3 * Size 4

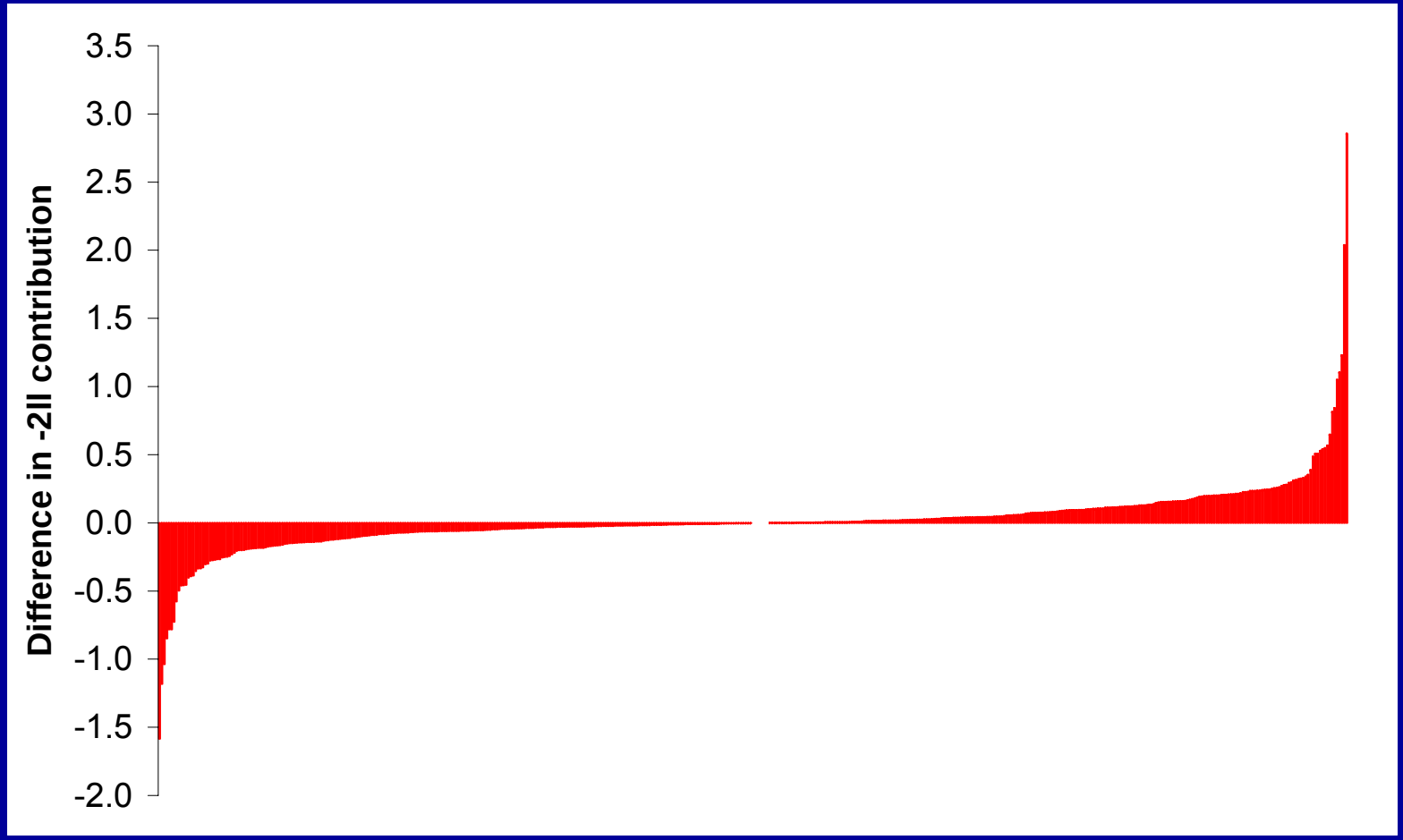
Which pairs contribute to the linkage?

- Mx allows to output the contribution to the -2ll per family:
- option %p=diagnostics.dat

1	2	3	4	5	6	7	8	9
9.000000000000000	7.336151039930395	1.540683866365682	8.343722785165869E-02	1	2	0	000	1
10.00000000000000	9.851302691037933	4.055835517473221	1.130602365719245	2	2	0	000	1
12.00000000000000	7.143777518583584	1.348310345018871	-3.614906755842623E-02	3	2	0	000	1

1. The first definition variable (wise to use a case identifier)
2. -2lnL the likelihood function for that vector of observations
3. the square root of the Mahalanobis distance
4. an estimated z-score
5. the number of the observation in the active (i.e. post selection) dataset. Note that with selection this may not correspond to the position of the vector in the data file
6. the number of data points in the vector (i.e. the family size if it is a pedigree with one variable per family member)
7. the number of times the log-likelihood was found to be incalculable during optimization
8. 000 if the likelihood was able to be evaluated at the solution, or 999 if it was incalculable
9. the model number if there are multiple models requested with the NModel argument to the Data line

- At marker 79 the lod score was highest
- The sum of all the individual -2ll's equals the -2ll given by mx. If you output the individual likelihoods for the FEQ model and the FE model, and subtract the two -2ll per family, you know how much each family contributes to the difference in -2ll and therefore to the lod-score


Difference in -2ll contribution
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0

Practical

- Take pihat1.mx and adjust this script to run at the position with the highest lod score (marker 79)
- Select variable fam and define fam as the first definition variable
- Run the AEQ model and add: options mx%p=diagnosticsAEQ.dat
- Run AE model with options mx%p=diagnosticsAE.dat
- Import the two dat files in excell, (contribution to LL.xls) select the first two columns of each dat file.
- Subtract the -2ll per family
- Sort the file on the difference in -2ll
- Produce a graph

PRACTICAL: F:\danielle\UnivariateLinkage\pract_linkage

If there is time -A little side step

Detecting outliers

- General must-do when analysing data
- Mx has a convenient mx%p output feature
- We'll illustrate the use of this feature in the context of univariate linkage, however, detecting outliers should normally be done in an earlier stage. In a linkage analysis, 'outlier' detection may help identifying most informative families

%p Viewer

- Java applet from QIMR to view the %p output in a convenient way
- Open stats_plot.jar,open diagnosticsAEQ.dat

Practical on outlier detection

- Select outliers
- Write to output (exclude.txt)
- Go back to Mx script, add: #include exclude.txt
- Run mx again, compare parameter estimates, and look at q-q plot