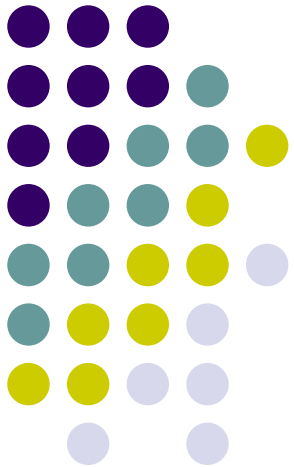
Linkage Analysis in Merlin

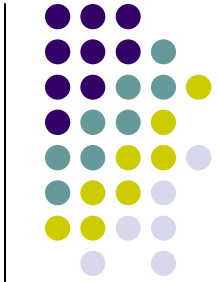
Meike Bartels

Kate Morley

Danielle Posthuma



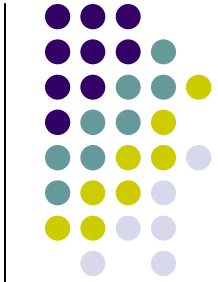
Software for linkage analyses



- Genehunter
- Mendel
- Vitesse
- Allegro
- Simwalk
- Loki
- Merlin
- ….

- Mx
- R
- Lisrel
- …

MERLIN software

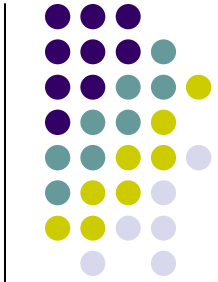


Programs:

- **MERLIN**
- MinX
- MERLIN-regress
- **Pedstats**
- Pedwipe
- Pedmerge

http://www.sph.umich.edu/csg/abecasis/Merlin/

MERLIN

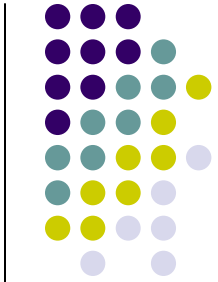


- Automates simple linkage tests (“black box”)
- Uses fast multipoint calculations to generate IBD and kinship matrices
- Key options are

  –vc (variance components analysis)

  –useCovariates (user-specified covariates)
- Means model
  - Can incorporate user-specified covariates
- Variance components model…

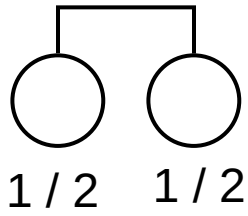
Merlin's Standard Variance Components Model



- Environmental component
  - Non shared, uses identity matrix
- Polygenic component
  - Shared among relatives, according to kinship matrix
- Major gene component
  - Shared when individuals are IBD, kinship matrix at marker

What is a Kinship Coefficient?

- Kinship coefficient (Φ): probability that two alleles sampled at random, one from each individual, are identical by descent



For MZ twins…
¼ 1 & 1
¼ 1 & 2
¼ 2 & 1
¼ 2 & 2

- 2 x Φ_{ij} = expected proportion of alleles IBD across genome for individuals i and j (π)
- But will vary at each locus → $\hat{\pi}$

General covariance model

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

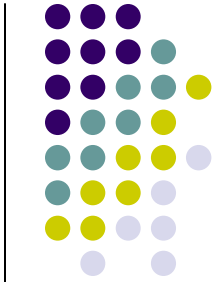
Where,

φ is the theoretical kinship coefficient for the two individuals

$\hat{\pi}$ depends on the number of alleles shared IBD for individuals j, k

j and k index different individuals in the family

Input Files (again)



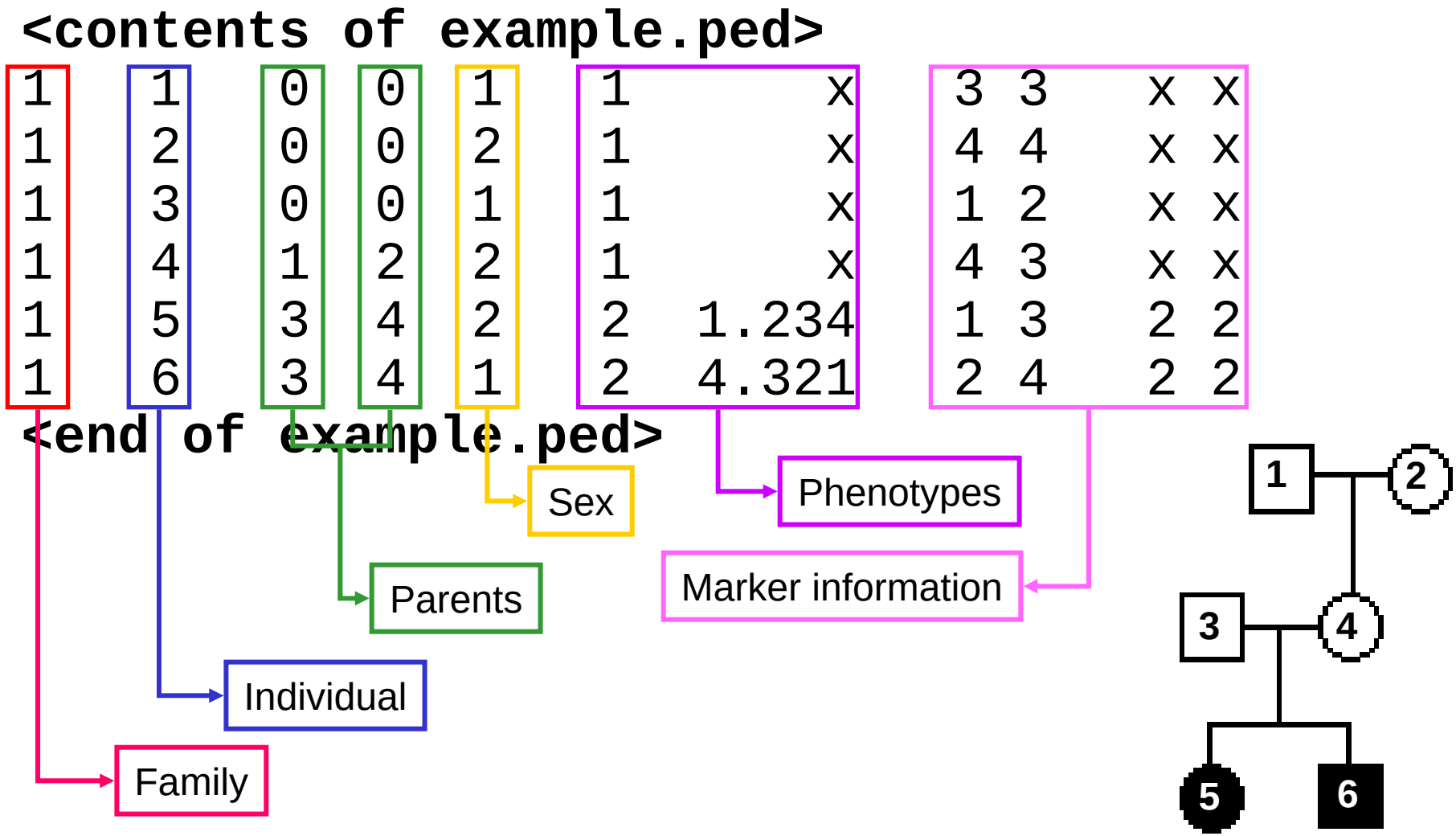
- Pedigree File
  - Family relationships
  - Phenotype data
  - Genotype data
- Data File
  - Describes contents of pedigree file
- Map File
  - Records location of genetic markers

Example Pedigree File

```
<contents of example.ped>
1   1   0  0  1   1       x    3 3    x x
1   2   0  0  2   1       x    4 4    x x
1   3   0  0  1   1       x    1 2    x x
1   4   1  2  2   1       x    4 3    x x
1   5   3  4  2   2   1.234    1 3    2 2
1   6   3  4  1   2   4.321    2 4    2 2
<end of example.ped>
```

Encodes family relationships, marker and phenotype information

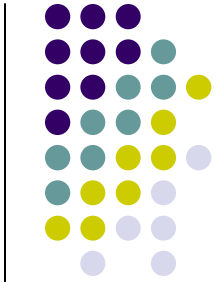
Example Pedigree File



Data File Field Codes

Code	Description
M	Marker Genotype
A	Affection Status.
T	Quantitative Trait.
C	Covariate.
Z	Zygosity.
S[n]	Skip n columns.

Example Data File

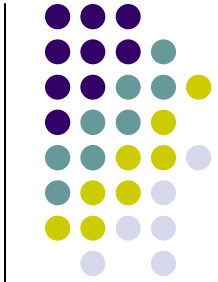


```
<contents of example.dat>
T    some_trait_of_interest
M    some_marker
M    another_marker
<end of example.dat>
```

Provides information necessary to decode pedigree file.

First five columns assumed to follow standard format: family, individual, father, mother, sex

Example Map File



```
<contents of example.map>
CHROMOSOME      MARKER      POSITION
2               D2S160      160.0
2               D2S308      165.0
…
<end of example.map>
```

Indicates location of individual markers, necessary to derive recombination fractions between them

Brisbane Twin and Family Study of Cognition

Memory
Attention
Problem
Solving

Queensland Institute of Medical Research and University of Queensland, Australia

Example Dataset

- Performance IQ Data
  - 710 sib-pairs
  - 59 micro-satellite markers on chromosome 2

Genome-wide Scan of IQ Finds Significant Linkage to a Quantitative Trait Locus on 2q

Behavior Genetics, Vol. 36, No. 1, January 2006

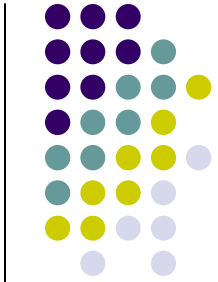
M. Luciano,[1,2] M. J. Wright,[1] D. L. Duffy,[1] M. A. Wainwright,[1] G. Zhu,[1] D. M. Evans,[1] G. M. Geffen,[1] G. W. Montgomery,[1] and N. G. Martin[1]

A Genomewide Scan for Intelligence Identifies Quantitative Trait Loci on 2q and 6p

Am. J. Hum. Genet. 77:318–326, 2005

Danielle Posthuma,[1] Michelle Luciano,[2] Eco J. C. de Geus,[1] Margie J. Wright,[2] P. Eline Slagboom,[3] Grant W. Montgomery,[2] Dorret I. Boomsma,[1] and Nicholas G. Martin[2]

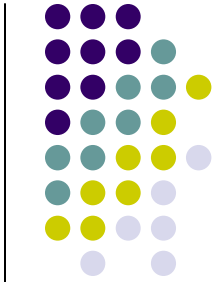
PIQ Dataset



- Analyses using chromosome 2 data
  1. Quick check and summary of data using PEDSTATS
  2. Variance components linkage analysis using Merlin
- Merlin input files
  - **`piq.ped`**
  - **`piq.dat`**
  - **`piq.map`**
- Copy this folder to your directory:

`F:\kate\merlin_prac`

Practical 1 - PEDSTATS

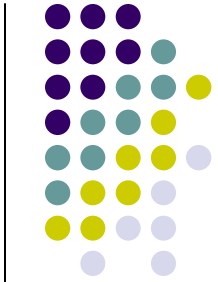


- An easy way to summarise your data…
  - Initial check of input files, pedigree consistency, genetic marker data, phenotypic data

- Open ms-dos prompt
- Navigate to your folder
  - **`dir`** to view files in a **dir**ectory
  - **`cd`** to **c**hange **d**irectory

http://www.sph.umich.edu/csg/abecasis/PedStats/index.html

Commands



- Run PEDSTATS

  `pedstats –d piq.dat –p piq.ped`

- Output as PDF document

  `--pdf`

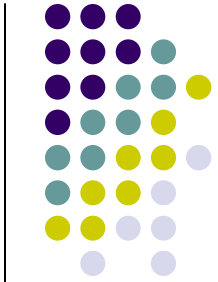
- Test Hardy Weinberg equilibrium of markers

  `--HardyWeinberg`

- Save the output to a file

  `> pedstats.out`

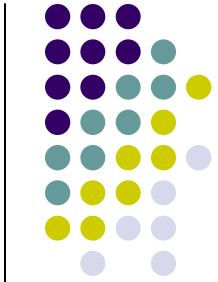
Pedigree & Trait Statistics



pedstats.out

```
PEDIGREE STRUCTURE
==================
   Individuals: 2840
      Founders: 1420 founders, 1420 nonfounders
        Gender: 1426 females, 1414 males
      Families: 710

  Family Sizes
       Average: 4.00 (4 to 4)
  Distribution: 4 (100.0%), 0 (0.0%) and 1 (0.0%)

  Generations
       Average: 2.00 (2 to 2)
  Distribution: 2 (100.0%), 0 (0.0%) and 1 (0.0%)

Checking family connectedness ...
   All individuals in each family are connected.

QUANTITATIVE TRAIT STATISTICS
=============================

             [All Phenotypes]      Min      Max     Mean      Var  SibCorr
          PIQ      856  30.1%   68.000  150.000  111.792  275.639    0.328
        Total      856  30.1%

             [Founders Only]      Min      Max     Mean      Var  SibCorr
          PIQ        0   0.0%        -        -        -        -        -
        Total        0   0.0%
```

Genotypic Data Quality



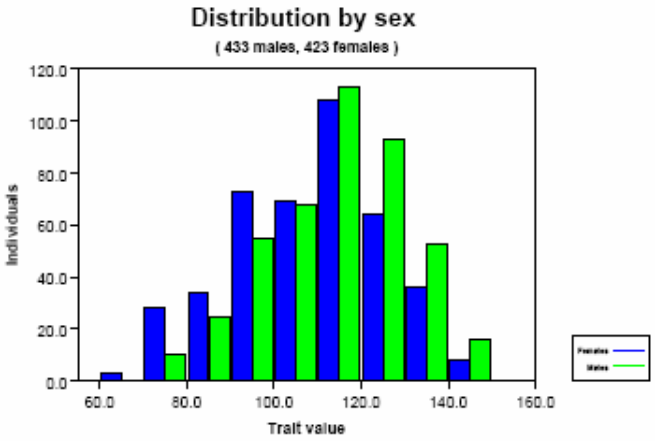
pedstats.out

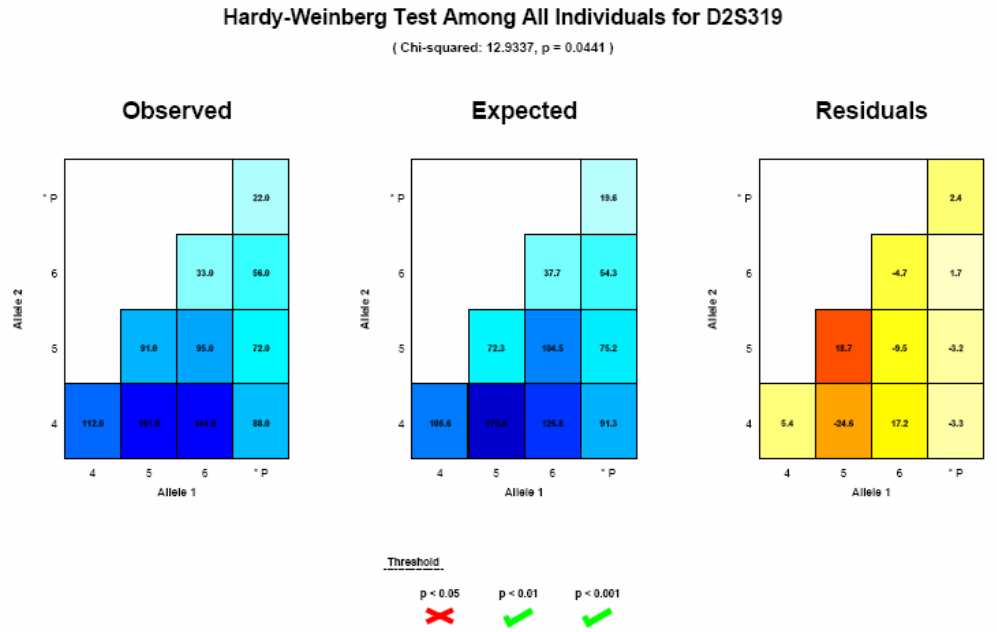
```
DATA QUALITY
============

HIGHEST AND LOWEST GENOTYPING RATES BY MARKER

         MARKER    RANK     PROP    N_GENO |         MARKER    RANK     PROP    N_GENO
-------------------------------------------------------------------------------------
        D2S2952       1    44.6%      1266 |         D2S151      59     5.4%       152
        D2S1777       2    44.3%      1258 |          XRCC5      58    23.4%       665
        D2S1394       3    44.1%      1253 |        D2S2313      57    24.9%       706
        D2S1776       4    44.1%      1253 |        D2S2259      56    29.6%       841
         D2S405       5    43.9%      1247 |         D2S367      55    30.0%       851
        D2S1352       6    43.8%      1243 |        D2S2330      54    30.0%       851
        D2S2944       7    43.6%      1238 |        D2S2382      53    30.0%       853
        D2S1353       8    43.5%      1235 |         D2S337      52    30.0%       853
        D2S1360       9    43.5%      1235 |         D2S396      51    30.1%       855
         D2S427      10    43.4%      1233 |         D2S305      50    30.2%       857

         Totals     59    35.5%     59414

HIGHEST AND LOWEST HETEROZYGOSITIES BY MARKER

         MARKER    RANK      HET    N_GENO |         MARKER    RANK      HET    N_GENO
-------------------------------------------------------------------------------------
        D2S1788       1    89.6%      1220 |        D2S1777      59    57.0%      1258
         D2S367       2    88.8%       851 |        D2S2968      58    61.2%      1189
         D2S337       3    87.5%       853 |         D2S286      57    62.1%       857
        D2S2330       4    86.7%       851 |        D2S1352      56    65.4%      1243
        D2S1360       5    86.7%      1235 |         D2S405      55    65.6%      1247
        D2S2976       6    86.7%      1217 |        D2S1400      54    66.2%      1216
         D2S168       7    86.2%       865 |         D2S442      53    66.5%      1203
         D2S165       8    85.3%       865 |        D2S1394      52    69.1%      1253
        D2S2313       9    85.0%       706 |          XRCC5      51    69.6%       665
         D2S162      10    84.4%       868 |        D2S2972      50    70.0%      1231

         Totals     59    77.4%     59414

Graphical output written to pedstats.pdf
```

Graphical Output

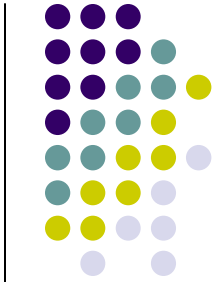
- Graphical output for
  - Pedigree and trait statistics
  - HWE tests

`pedstats.pdf`





Practical 2 – Merlin VC



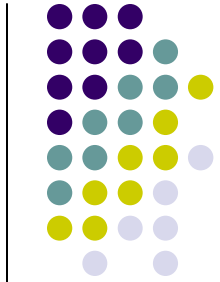
- In the same directory, type

```
merlin –d piq.dat –p piq.ped –m piq.map --vc
```

- `--pdf` → PDF file output
- `--grid 2` → Analysis at every 2 cM
- `--start 0` → Start grid at position 0 cM
- `--perFamily` → Per family contributions to log-likelihood and LOD score

Don’t forget to send text output to a file: `> merlin.out`

Output

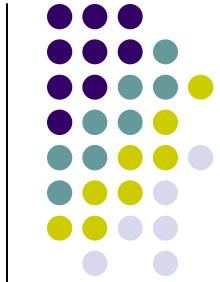


Phenotype: PIQ [VC] (550 families, h2 = 66.56%)

Position	H2	ChiSq	LOD	pvalue
0.000	0.00%	0.00	0.00	0.5
2.000	0.00%	0.00	0.00	0.5
4.000	0.00%	0.00	0.00	0.5
6.000	0.00%	0.00	0.00	0.5
8.000	0.00%	0.00	0.00	0.5
10.000	0.00%	0.00	0.00	0.5
12.000	0.00%	0.00	0.00	0.5
14.000	0.00%	0.00	0.00	0.5
16.000	0.00%	0.00	0.00	0.5
18.000	0.00%	0.00	0.00	0.5
20.000	0.00%	0.00	0.00	0.5
22.000	0.00%	0.00	0.00	0.5
24.000	0.00%	0.00	0.00	0.5
26.000	0.00%	0.00	0.00	0.5
28.000	0.00%	0.00	0.00	0.5
30.000	0.00%	0.00	0.00	0.5
32.000	0.00%	0.00	0.00	0.5
34.000	0.00%	0.00	0.00	0.5
36.000	1.81%	0.01	0.00	0.5
38.000	3.95%	0.05	0.01	0.4
40.000	7.63%	0.17	0.04	0.3
42.000	6.80%	0.13	0.03	0.4
44.000	0.33%	0.00	0.00	0.5

merlin.out

Output



```
Phenotype: PIQ [VC] (550 families, h2 = 66.56%)
=================================================================
          Position      H2     ChiSq      LOD   pvalue
             0.000   0.00%      0.00     0.00      0.5
             2.000   0.00%      0.00     0.00      0.5
             4.000   0.00%      0.00     0.00      0.5
             6.000   0.00%      0.00     0.00      0.5
             8.000   0.00%      0.00     0.00      0.5
            10.000   0.00%      0.00     0.00      0.5
            12.000   0.00%      0.00     0.00      0.5
            14.000   0.00%      0.00     0.00      0.5
            16.000   0.00%      0.00     0.00      0.5
            18.000   0.00%      0.00     0.00      0.5
            20.000   0.00%      0.00     0.00      0.5
            22.000   0.00%      0.00     0.00      0.5
            24.000   0.00%      0.00     0.00      0.5
            26.000   0.00%      0.00     0.00      0.5
            28.000   0.00%      0.00     0.00      0.5
            30.000   0.00%      0.00     0.00      0.5
            32.000   0.00%      0.00     0.00      0.5
            34.000   0.00%      0.00     0.00      0.5
            36.000   1.81%      0.01     0.00      0.5
            38.000   3.95%      0.05     0.01      0.4
            40.000   7.63%      0.17     0.04      0.3
            42.000   6.80%      0.13     0.03      0.4
            44.000   0.33%      0.00     0.00      0.5
```

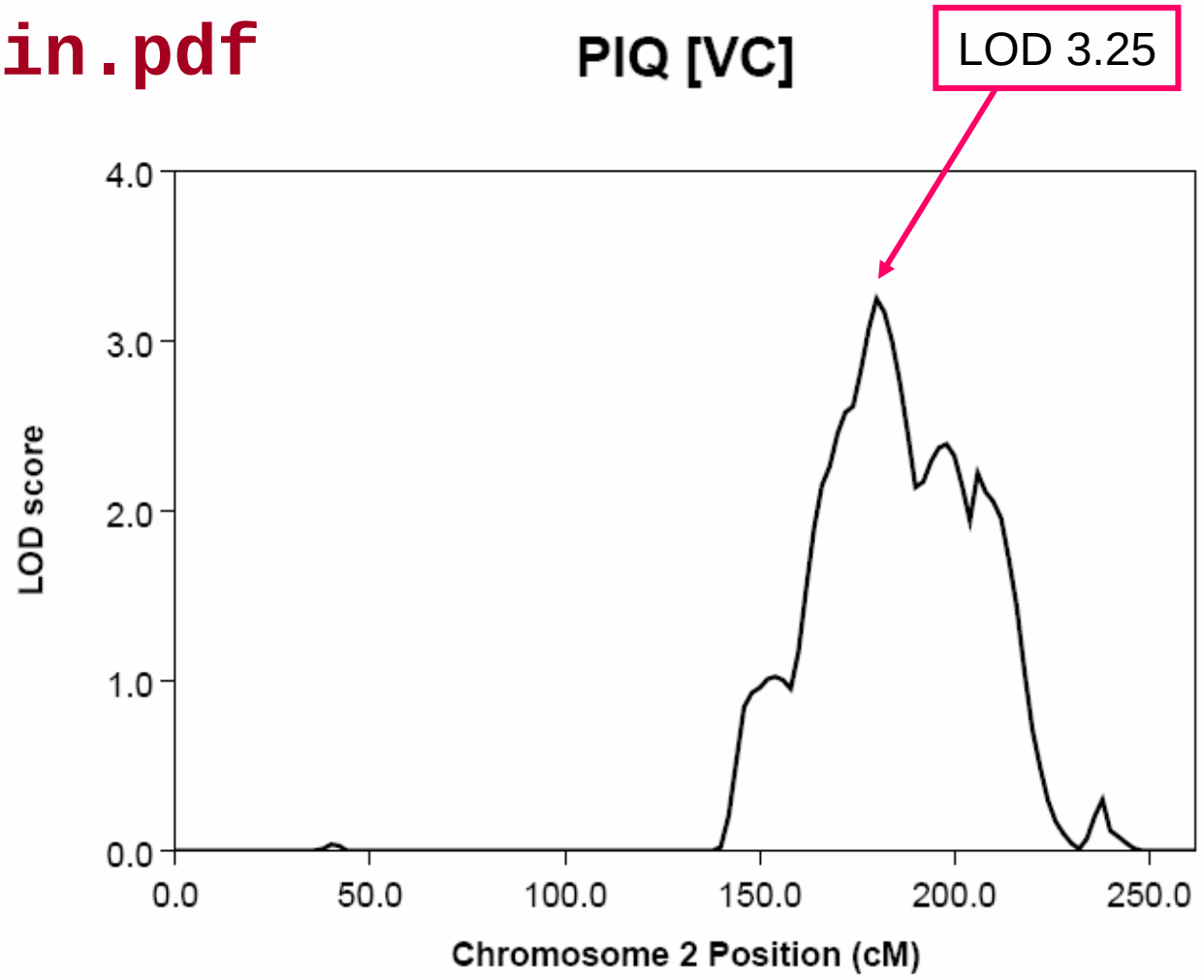
sample heritability

evidence for linkage?

merlin.out

Results

`merlin.pdf`



Family Contributions

merlin.vc

FAMILY	POSITION	LLK_NULL	LLK_ALT	LOD
80020	0.000	-5.63768	-5.63762	0.00003
80030	0.000	-6.01550	-6.01566	-0.00007
80033	0.000	-5.80161	-5.80163	-0.00001
80040	0.000	-6.88094	-6.88093	0.00000
80090	0.000	-7.80469	-7.80433	0.00015
80092	0.000	-5.99204	-5.99181	0.00010
80110	0.000	-2.81080	-2.81072	0.00003
80140	0.000	-6.15116	-6.15085	0.00013
80182	0.000	-2.99884	-2.99893	-0.00004
80200	0.000	-7.10584	-7.10547	0.00016
80201	0.000	-5.70219	-5.70195	0.00010
80210	0.000	-3.17603	-3.17619	-0.00007
80331	0.000	-8.19427	-8.19508	-0.00035
80332	0.000	-3.06162	-3.06144	0.00008
80400	0.000	-7.75628	-7.75637	-0.00004
80402	0.000	-5.74423	-5.74427	-0.00002
80404	0.000	-2.96361	-2.96368	-0.00003
80410	0.000	-5.69553	-5.69554	-0.00001
80413	0.000	-6.21017	-6.21030	-0.00006

Family Contributions

merlin.vc

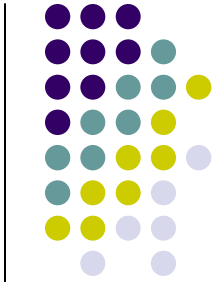
FAMILY	POSITION	LLK_NULL	LLK_ALT	LOD
80020	0.000	-5.63768	-5.63762	0.00003
80030	0.000	-6.01550	-6.01566	-0.00007
80033	0.000	-5.80161	-5.80163	-0.00001
80040	0.000	-6.88094	-6.88093	0.00000
80090	0.000	-7.80469	-7.80433	0.00015
80092	0.000	-5.99204	-5.99181	0.00010
80110	0.000	-2.81080	-2.81072	0.00003
80140	0.000	-6.15116	-6.15085	0.00013
80182	0.000	-2.99884	-2.99893	-0.00004
80200	0.000	-7.10584	-7.10547	0.00016
80201	0.000	-5.70219	-5.70195	0.00010
80210	0.000	-3.17603	-3.17619	-0.00007
80331	0.000	-8.19427	-8.19508	-0.00035
80332	0.000	-3.06162	-3.06144	0.00008
80400	0.000	-7.75628	-7.75637	-0.00004
80402	0.000	-5.74423	-5.74427	-0.00002
80404	0.000	-2.96361	-2.96368	-0.00003
80410	0.000	-5.69553	-5.69554	-0.00001
80413	0.000	-6.21017	-6.21030	-0.00006

Null hypothesis log-likelihood

Alternative hypothesis

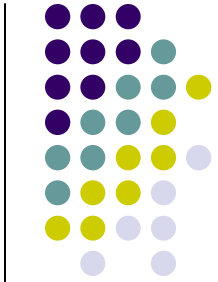
LOD score

Creating Input Files



- Create your own Merlin input files
- Small example data set: 10 families, 2 offspring each (no parents!), one trait, one marker
- Initial data in **`Input.Exercise.xls`**
- Create **`ex.ped ex.dat ex.map`**
  - Use a text editor e.g. PFE (included in prac folder)
  - Use 3 and 4 to denote father and mother extensions (remember – need parental information to link siblings, even if parents not genotyped)
  - Use x for missing data
  - Save files to your directory

Analysing Your Data…



- Check your files using PEDSTATS

  `pedstats -d ex.dat -p ex.ped`

- Run VC linkage analysis in Merlin:

  `merlin -d ex.dat -p ex.ped -m ex.map --vc`

- Your LOD score should be 0.41