Copy the folder…

Faculty/Sarah/Tues_merlin

to

the **C Drive**

C:/Tues_merlin

MERLIN (and other Abecasis products)

Sarah Medland & Kate Morley

Boulder 2009

MERLIN software

Programs:

- GRR
- MERLIN
- MinX
- MERLIN-regress
- Pedstats
- Pedwipe
- Pedmerge

We will be using Cygwin…

- Unix emulator for windows
- Open by double clicking
- Migrate to this sessions working directory
  - cd C:/tues_merlin
- Check to see the files in the directory
  - ls

Data Input Files

Getting your data into Merlin

Input File Types

- 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>

Family

Individual

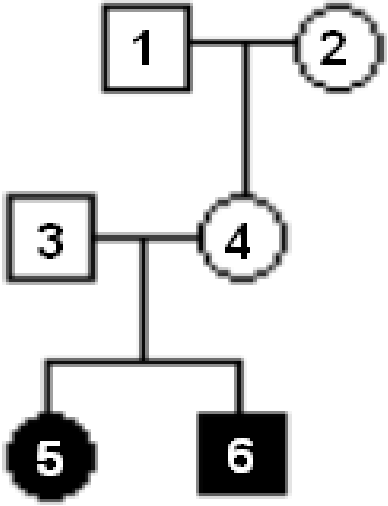
Parents

Sex

Phenotypes

Marker information

N.B. x = missing value



Data File Field Codes

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

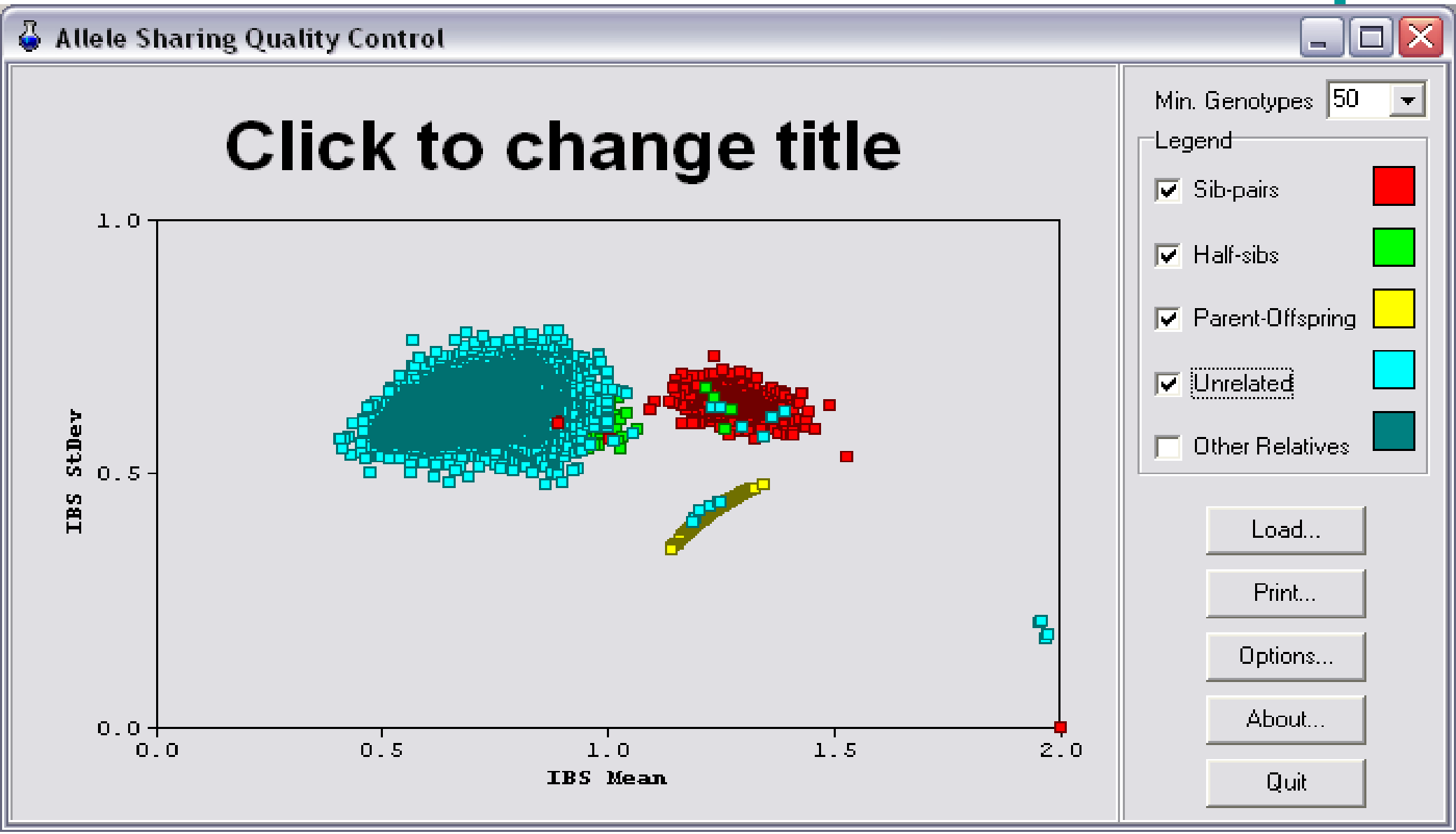
First step check relationships

GRR

Grr.exe

GRR - www.sph.umich.edu/csg/abecasis/GRR

- Graphs **mean** IBS against **sd** IBS
  - Either within families or across everyone in the sample
  - Ideally 200+ markers genotyped in common for each pair
- If you want to try this later…Sample.ped
  - 1300 individuals from 200 families
  - Genotyped on 320 markers across the genome

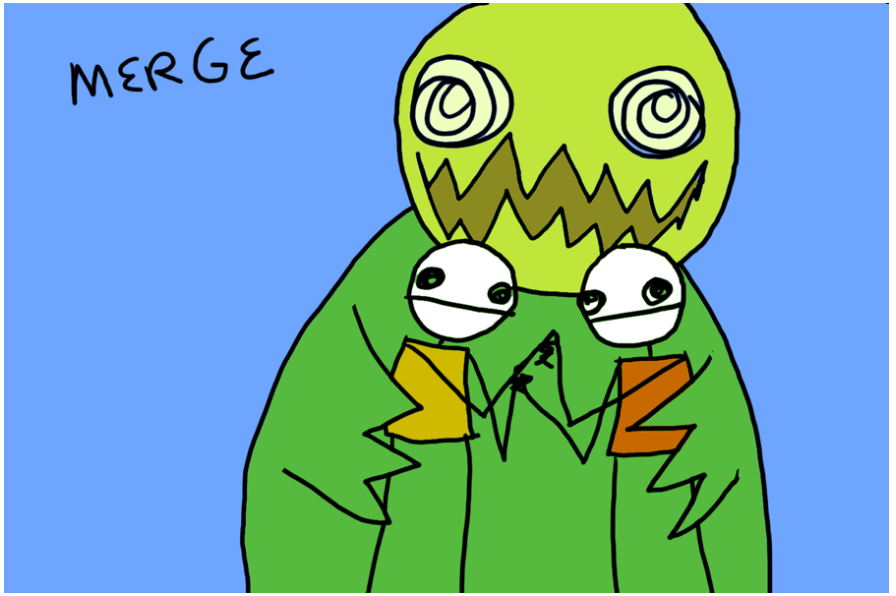

Allele Sharing Quality Control
Click to change title
1.0
0.5
0.0
IBS StDev
0.0
0.5
1.0
1.5
2.0
IBS Mean
Min. Genotypes
50
Legend
Sib-pairs
Half-sibs
Parent-Offspring
Unrelated
Other Relatives
Load...
Print...
Options...
About...
Quit

GRR is good for finding…

- MZ pairs labeled as sib-pairs
- Duplicates
- Dads that aren’t dads
- Full sibs who are half-sibs

Manipulating Data Files

Pedmerge



Manipulating Data Files

- Pedmerge
  - Combine multiple data files
  - Remove columns from a ped file
    - Recode the dat file so unwanted columns are skipped
  - Assumes ped and dat files have the same prefix – example.ped example.dat

Type ‘pedmerge’

```
$ ./pedmerge
PedMerge - Pedigree Merge (c) 1999 Goncalo Abecasis

Usage: pedmerge input1 input2 ... output

This program will try to merge a set of paired pedigree (.ped)
and data (.dat) files into a single composite pedigree.

For example:

    > pedmerge a b c

Will create the files c.dat and c.ped including all the phenotype
data and individuals in a.dat, a.ped, b.dat and b.ped.

WARNING: pedmerge will overwrite output files without checking
```

Checking for genotype error

Pedstats

Usage

- pedstats.exe –p pedstats.ped –d pedstats.dat

```
sarahm@medland-office /cygdrive/c/working/tim/monday
$ ./pedstats.exe -p pedstats.ped -d pedstats.dat
Pedigree Statistics - 0.6.10
(c) 1999-2006 Goncalo Abecasis, 2002-2006 Jan Wigginton

The following parameters are in effect:
                  Pedigree File :    pedstats.ped (-pname)
                      Data File :    pedstats.dat (-dname)
                       IBD File :    pedstats.ibd (-iname)
                  Adobe PDF File :    pedstats.pdf (-aname)
              Missing Value Code :         -99.999 (-xname)

Additional Options
     Pedigree File : --ignoreMendelianErrors, --chromosomeX, --trim
    Hardy-Weinberg : --hardyWeinberg, --showAll, --cutoff [0.05]
         HW Sample : --checkFounders, --checkAll, --checkUnrelated
            Output : --pairs, --rewritePedigree, --markerTables, --verbose
          Grouping : --bySex, --byFamily
       Age Checking : --age [], --birth []
        Generations : --minGap [13.00], --maxGap [70.00], --sibGap [30.00]
        PDF Options : --pdf, --familyPDF, --traitPDF, --affPDF, --markerPDF
             Filter : --minGenos, --minPhenos, --minCovariates, --affectedFor []
```

Summarizes pedigree

```
PEDIGREE STRUCTURE
==================
    Individuals: 1500
       Founders: 600 founders, 900 nonfounders
         Gender: 719 females, 781 males
       Families: 300

  Family Sizes
        Average: 5.00 (5 to 5)
  Distribution: 5 (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.
```

Trait summary

QUANTITATIVE TRAIT STATISTICS

[All Phenotypes]			Min	Max	Mean	Var	SibCorr
TRAIT	1500	100.0%	-2.877	3.496	-0.009	1.033	0.318
Total	1500	100.0%					

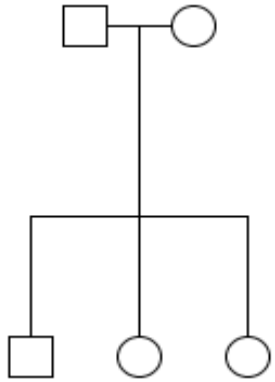
[Founders Only]			Min	Max	Mean	Var	SibCorr
TRAIT	600	100.0%	-2.536	3.496	0.016	1.066	-
Total	600	100.0%					

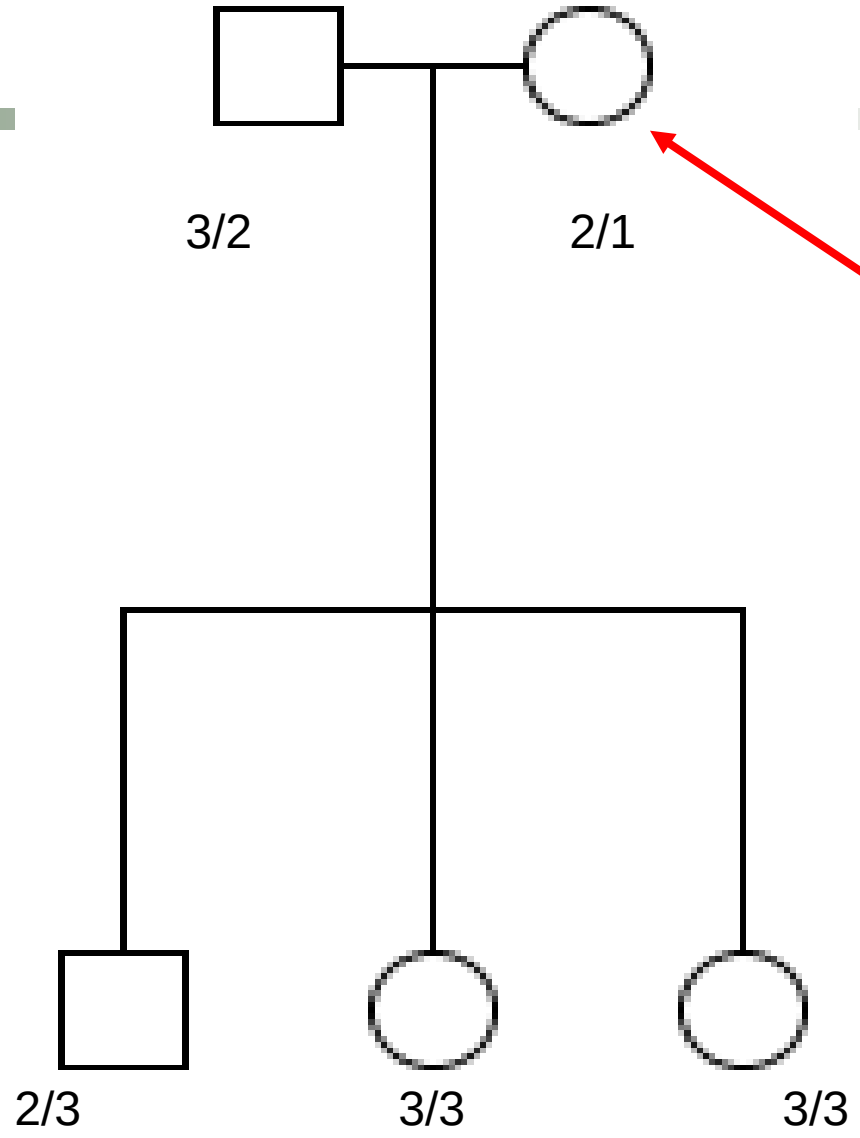
Pedstats will crash if there are Medelian errors

```
M2 - Fam 1: Child 4 [3/3] has parents [2/3]*[1/2]
M2 - Fam 1: Child 5 [3/3] has parents [2/3]*[1/2]

Mendelian inheritance errors detected

FATAL ERROR -
Mendelian inheritance errors detected
```

- Draw a diagram for this family

fam	id	dad	mum	sex	A1	A2
1	1	0	0	m	3	2
1	2	0	0	f	2	1
1	3	1	2	m	2	3
1	4	1	2	f	3	3
1	5	1	2	f	3	3




3/2
2/1
2/3
3/3
3/3

Mendelian errors

- Try to localize the error
- Short term solution – delete the bad genotypes
- Long term solution – retype the family at this marker

After fixing the problems

MARKER GENOTYPE STATISTICS

	[Genotypes]	[Founders]	Hetero
M2	1500 100.0%	600 100.0%	69.5%
M4	1500 100.0%	600 100.0%	66.6%
M6	1500 100.0%	600 100.0%	68.9%
M8	1500 100.0%	600 100.0%	68.7%
M10	1500 100.0%	600 100.0%	70.3%
M12	1500 100.0%	600 100.0%	70.3%
M14	1500 100.0%	600 100.0%	67.7%
M16	1500 100.0%	600 100.0%	69.6%
M18	1500 100.0%	600 100.0%	68.0%
M20	1500 100.0%	600 100.0%	68.1%
M22	1500 100.0%	600 100.0%	67.8%
M24	1500 100.0%	600 100.0%	64.5%
M26	1500 100.0%	600 100.0%	69.4%
M28	1500 100.0%	600 100.0%	66.0%
M30	1500 100.0%	600 100.0%	68.9%
M32	1500 100.0%	600 100.0%	66.5%
M34	1500 100.0%	600 100.0%	67.3%
M36	1500 100.0%	600 100.0%	70.5%
M38	1500 100.0%	600 100.0%	68.4%
M40	1500 100.0%	600 100.0%	69.2%
M42	1500 100.0%	600 100.0%	67.6%
M44	1500 100.0%	600 100.0%	70.1%
M46	1500 100.0%	600 100.0%	69.0%
M48	1500 100.0%	600 100.0%	69.7%
M50	1500 100.0%	600 100.0%	67.3%
M52	1500 100.0%	600 100.0%	68.9%
M54	1500 100.0%	600 100.0%	68.5%

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)

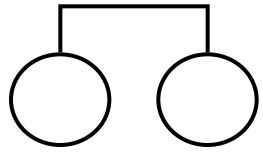
  –useCovariates (user-specified covariates)
- Means model
  - Can incorporate user-specified covariates
- Variance components model…

Merlin's Standard Variance Components Model - AQE

- Environmental component
  - Non shared, uses identity matrix
- Additive Polygenic component
  - Shared among relatives, according to kinship matrix
- QTL 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 $\Phi = .5$
For Full sibs $\Phi = .25$

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

General covariance model

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

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

Practical overview

- Using the LDL data from chromosome 19 (yesterday afternoon's practical)
- Data cleaning
  - Merging phenotype and genotype data
  - Checking you data with pedstats
- VC analysis in MERLIN
- MERLIN-regress analysis
- Comparison of MERLIN vs Mx

Step #1: combining phenotypes and genotypes

- Start with four files:
  - `pheno.ped` + `pheno.dat` (phenotype data)
  - `geno.ped` + `geno.dat` (genotype data)

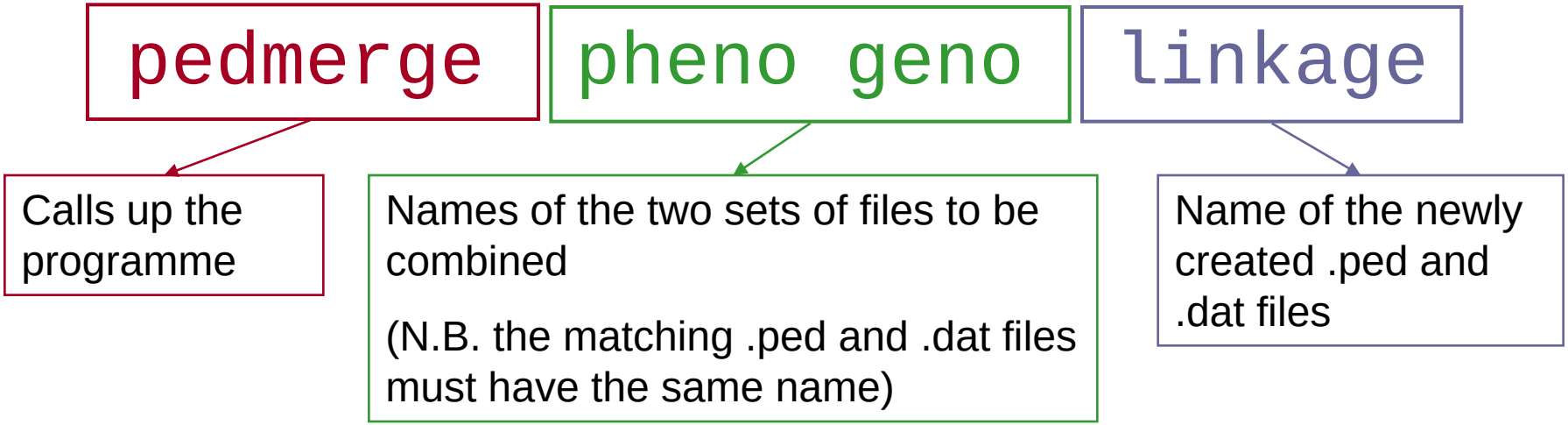
- Combine .ped files and combine .dat files using pedmerge to create 1 pedigree file and 1 .dat file

Practical #1: commands

- Have a look at your files

  `head <filename>`

- Combine your pedigree files and dat files



Check your file using the `head` command

linkage.ped

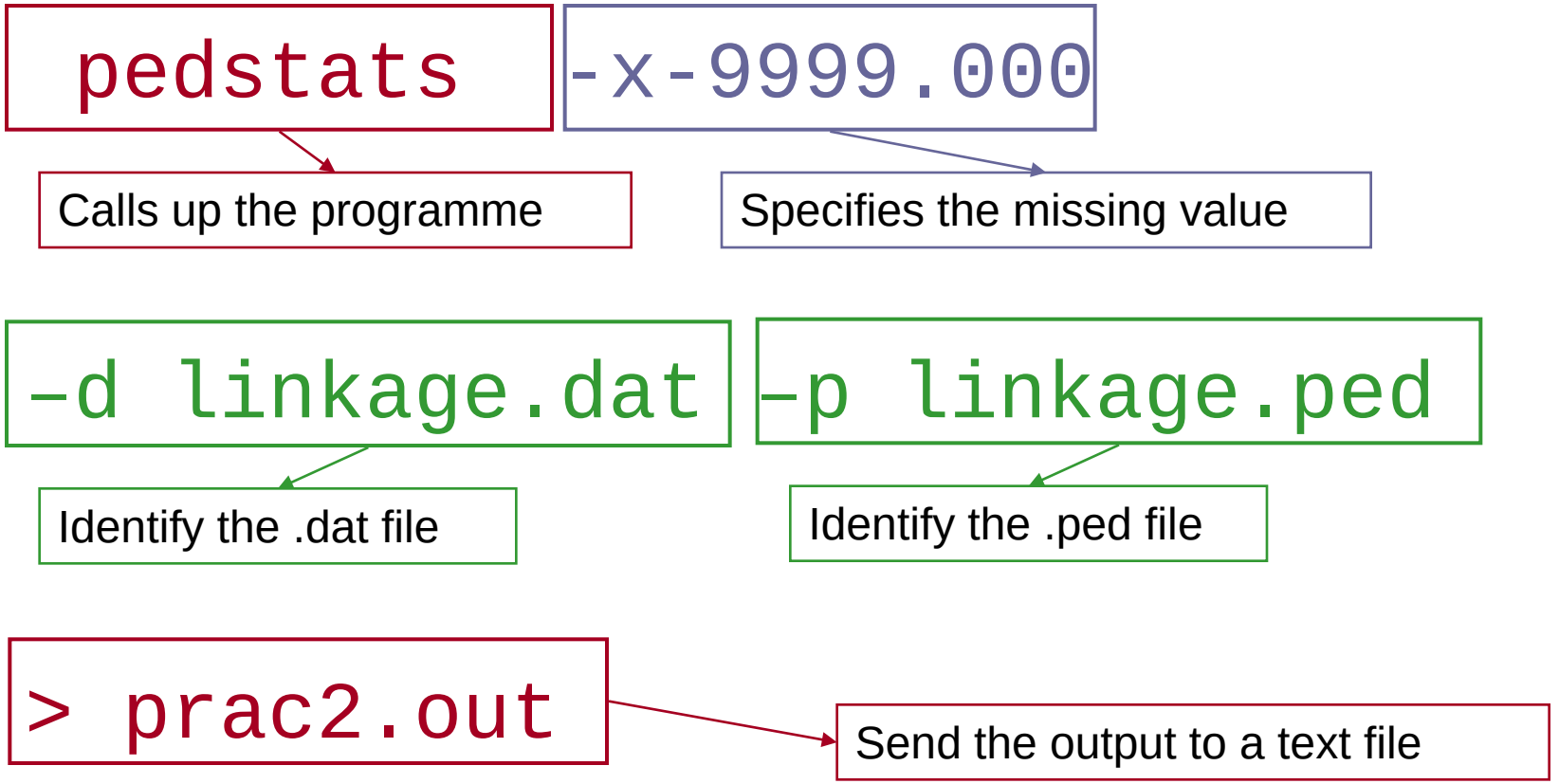
701	1	0	0	1	0/	0	0/	0	6/	15	7/	15	0/	0
701	2	0	0	2	0/	0	0/	0	5/	9	7/	14	0/	0
701	3	1	2	2	2/	8	6/	6	5/	15	14/	15	13/	23
701	4	1	2	2	2/	7	6/	6	6/	9	7/	7	7/	17
706	1	0	0	1	0/	0	0/	0	0/	0	0/	0	0/	0
706	2	0	0	2	0/	0	0/	0	0/	0	0/	0	0/	0
706	3	1	2	2	2/	2	4/	5	8/	8	2/	8	20/	20
706	4	1	2	2	2/	2	5/	5	8/	8	2/	7	18/	20
713	1	0	0	1	0/	0	0/	0	0/	0	0/	0	0/	0
713	2	0	0	2	0/	0	0/	0	0/	0	0/	0	0/	0

Step #2: checking your data with pedstats

- Pedstats provides preliminary data checks
  - Initial check of input files
  - Pedigree consistency
  - Information on genetic marker data
    - Marker heterozygosity
    - Proportion of individuals genotyped
    - Tests of Hardy Weinberg equilibrium

Prac #2: commands

```
./pedstats -x-9999.000 -d linkage.dat -p linkage.ped > prac2.out
```



```
Pedigree Statistics - 0.6.10
(c) 1999-2006 Goncalo Abecasis, 2002-2006 Jan Wigginton

The following parameters are in effect:
                  Pedigree File :     linkage.ped (-pname)
                      Data File :     linkage.dat (-dname)
                       IBD File :    pedstats.ibd (-iname)
                  Adobe PDF File :    pedstats.pdf (-aname)
              Missing Value Code :       -9999.000 (-xname)

Additional Options
    Pedigree File : --ignoreMendelianErrors, --chromosomeX, --trim
   Hardy-Weinberg : --hardyWeinberg, --showAll, --cutoff [0.05]
        HW Sample : --checkFounders, --checkAll, --checkUnrelated
           Output : --pairs, --rewritePedigree, --markerTables, --verbose
         Grouping : --bySex, --byFamily
      Age Checking : --age [], --birth []
       Generations : --minGap [13.00], --maxGap [70.00], --sibGap [30.00]
       PDF Options : --pdf, --familyPDF, --traitPDF, --affPDF, --markerPDF
            Filter : --minGenos, --minPhenos, --minCovariates, --affectedFor []
```

```
PEDIGREE STRUCTURE
==================
   Individuals: 452
      Founders: 226 founders, 226 nonfounders
        Gender: 231 females, 221 males
      Families: 113

  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
ldl	225	49.8%	0.580	6.390	3.616	0.997	0.443
Total	225	49.8%					

COVARIATE STATISTICS
==========

[All Phenotypes]			Min	Max	Mean	Var	SibCorr
sex	226	50.0%	0.000	1.000	0.522	0.251	0.326
age	226	50.0%	34.500	59.500	44.199	44.905	1.000
Total	452	50.0%					

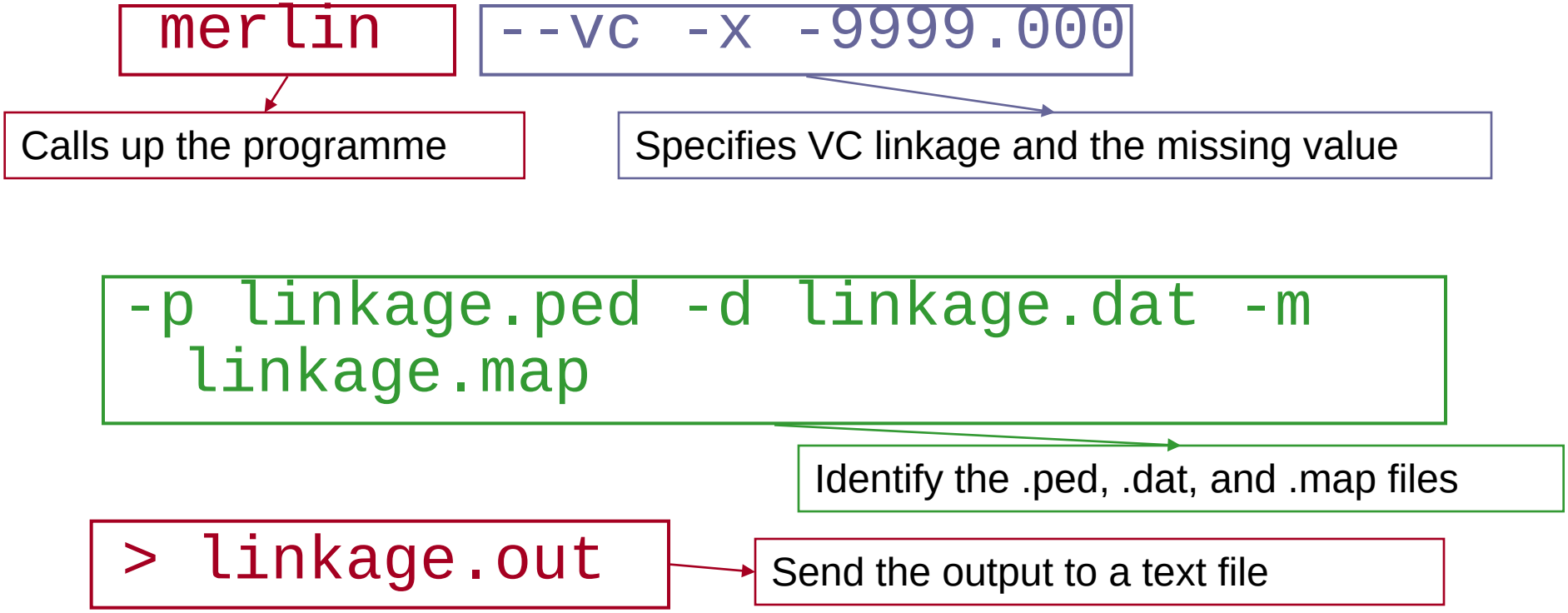
MARKER GENOTYPE STATISTICS
==========

	[Genotypes]		[Founders]		Hetero
d19M47	220	48.7%	0	0.0%	78.2%
d19s1034	219	48.5%	0	0.0%	71.7%
d19s391	251	55.5%	41	18.1%	83.7%
d19s865	248	54.9%	42	18.6%	86.3%
d19s394	219	48.5%	0	0.0%	87.7%
D19S588	221	48.9%	0	0.0%	80.1%
d19s49	224	49.6%	0	0.0%	80.4%
d19s433	225	49.8%	0	0.0%	80.0%
d19s47	210	46.5%	0	0.0%	67.1%
d19s420	242	53.5%	55	24.3%	80.6%
d19s178	249	55.1%	57	25.2%	75.1%
apoe	269	59.5%	44	19.5%	46.8%
apoc2	205	45.4%	0	0.0%	88.8%
d19M46	216	47.8%	0	0.0%	88.9%
d19s180	211	46.7%	0	0.0%	73.9%
d19M10	224	49.6%	0	0.0%	78.1%
d19M54	223	49.3%	0	0.0%	83.9%
Total	3876	50.4%	239	6.2%	78.0%

Total markers: 17

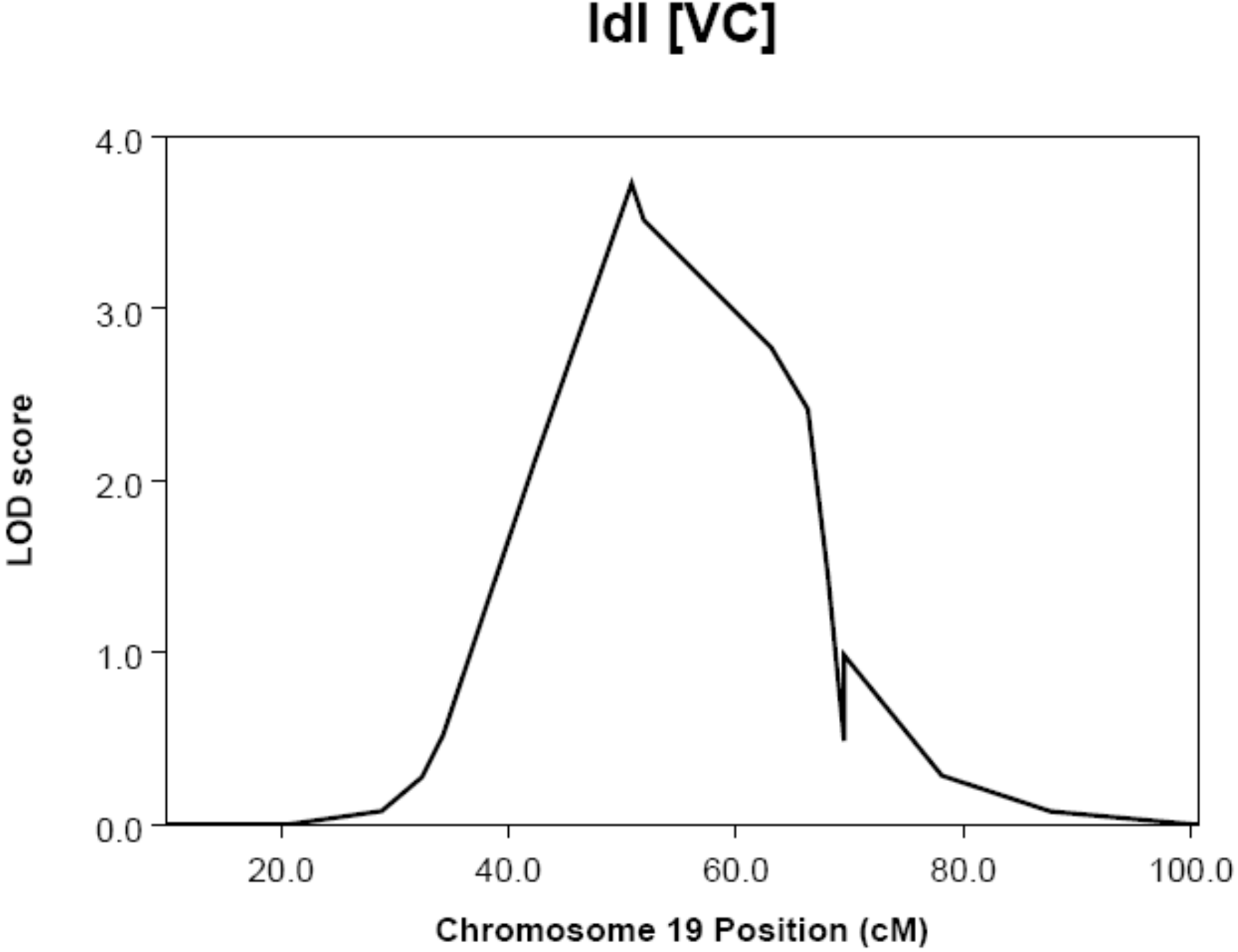
Step #3: running VC linkage

```
./merlin --vc -x -9999.000 -p linkage.ped -d linkage.dat -m
   linkage.map > linkage.out
```



Phenotype: ldl [VC] (113 families, h2 = 88.38%)

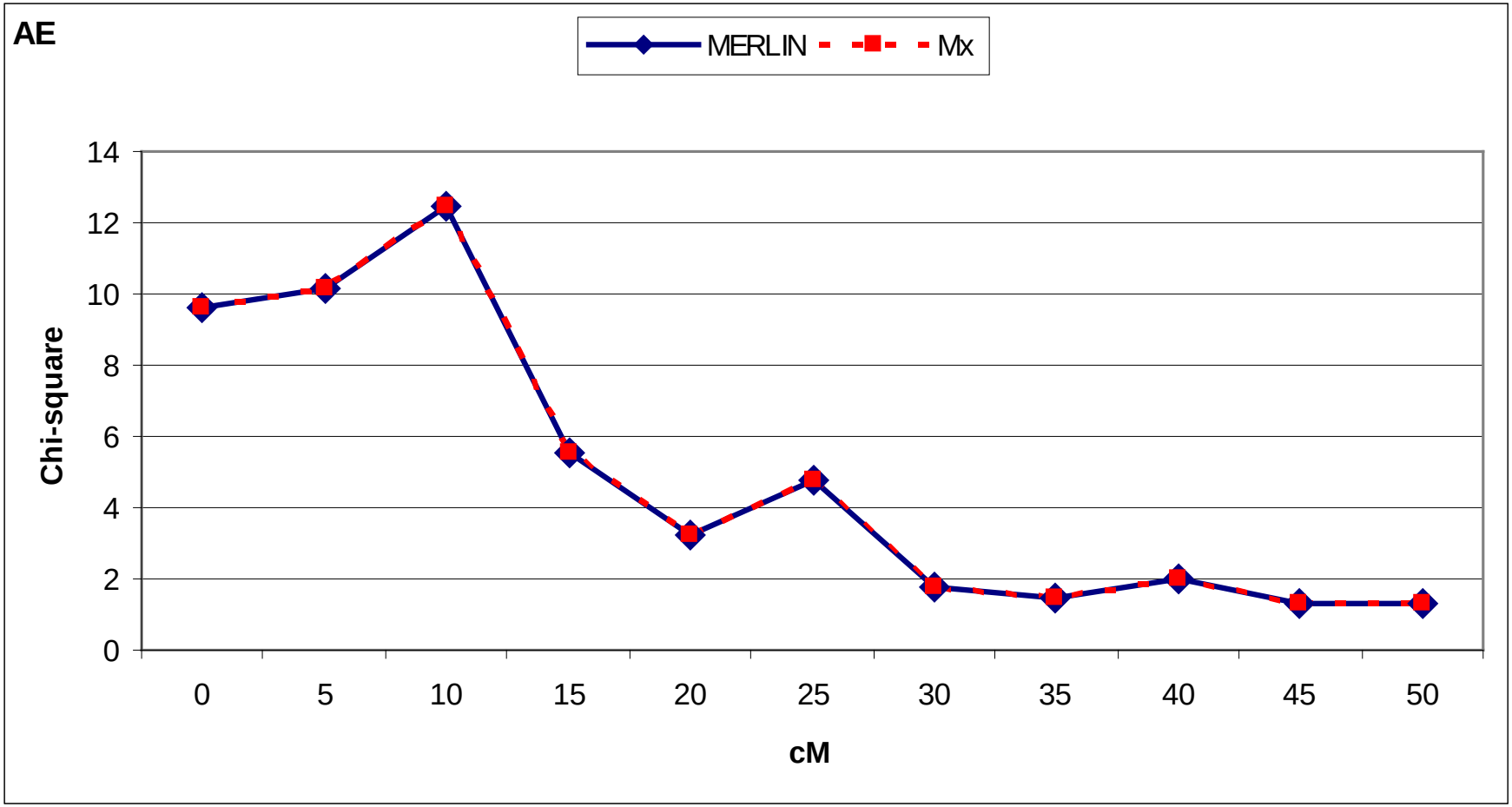
Position	H2	ChiSq	LOD	pvalue
9.840	0.00%	0.00	0.00	0.5
20.750	0.00%	0.00	0.00	0.5
28.830	14.31%	0.35	0.08	0.3
32.390	25.05%	1.26	0.27	0.13
34.250	34.56%	2.39	0.52	0.06
42.280	63.45%	9.71	2.11	0.0009
50.810	77.33%	17.17	3.73	0.00002
51.880	76.57%	16.20	3.52	0.00003
63.100	80.53%	12.77	2.77	0.0002
66.300	74.89%	11.12	2.42	0.0004
68.080	58.58%	6.66	1.45	0.005
69.490	37.96%	2.24	0.49	0.07
69.500	45.02%	4.53	0.98	0.02
78.080	25.38%	1.30	0.28	0.13
87.660	12.81%	0.34	0.07	0.3
100.010	0.00%	0.00	0.00	0.5
100.610	0.00%	0.00	0.00	0.5


ldl [VC]
LOD score
4.0
3.0
2.0
1.0
0.0
20.0
40.0
60.0
80.0
100.0
Chromosome 19 Position (cM)

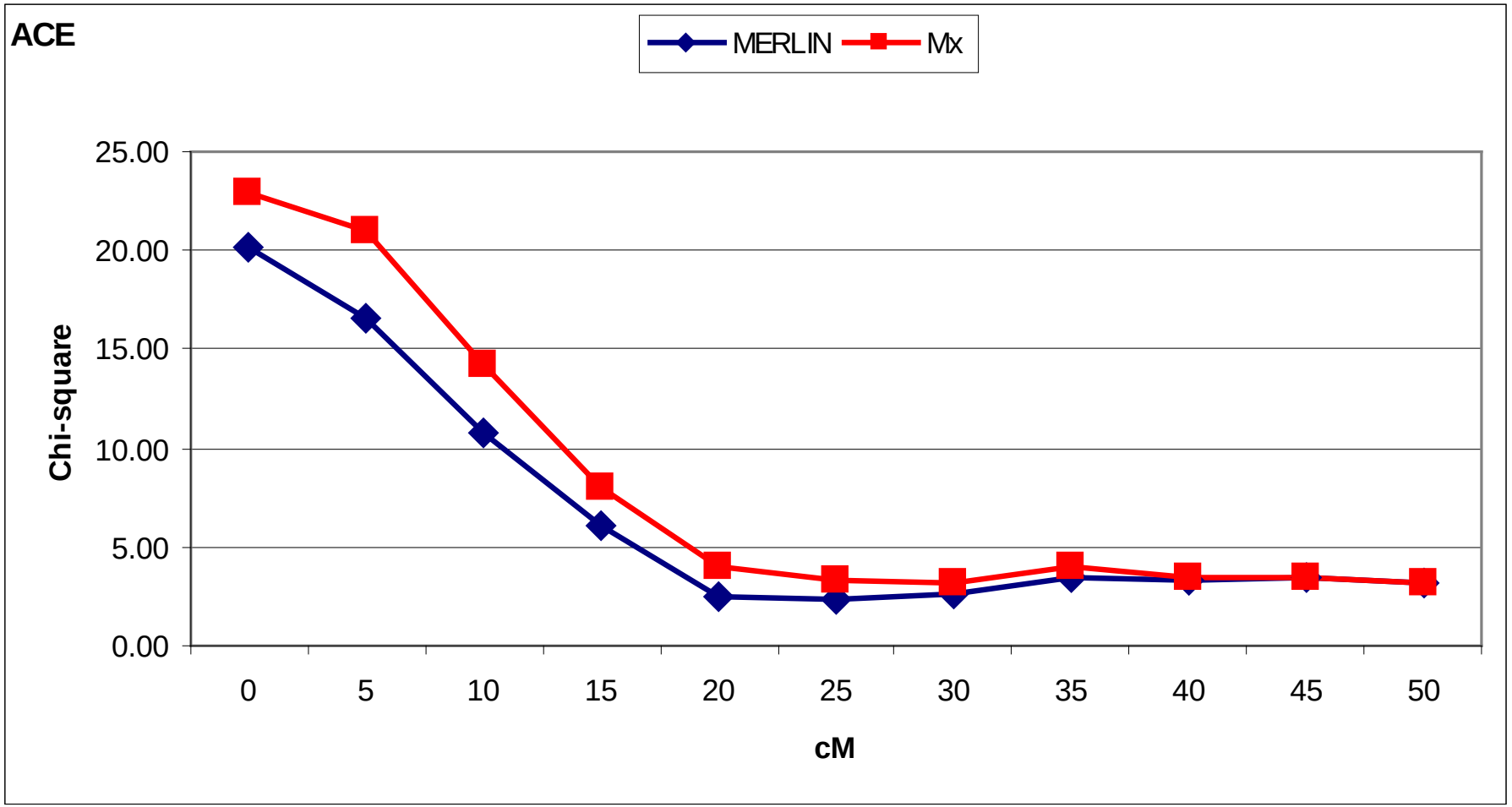
So why would we run Mx

- Merlin can not analyse ordinal data
- Limited correction for ascertainment
- Limited multivariate linkage
  - repeated measures using the mean and TRT correlation
- Only runs an AE model – no C or D

A 86% E 14%


AE
MERLIN
Mx
Chi-square
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
40
45
50
cM

A 60% C 30% E 10%


ACE
MERLIN
Mx
25.00
20.00
15.00
10.00
5.00
0.00
Chi-square
0
5
10
15
20
25
30
35
40
45
50
cM

Merlin Regress

Aim

- To develop a regression-based method that
  - Has same power as maximum likelihood variance components, for sib pair data
  - Will generalise to general pedigrees
  - Is computationally efficient

Am. J. Hum. Genet. 71:238–253, 2002

Powerful Regression-Based Quantitative-Trait Linkage Analysis of General Pedigrees

Pak C. Sham,[1] Shaun Purcell,[1] Stacey S. Cherny,[1,2] and Gonçalo R. Abecasis[3]

[1]Institute of Psychiatry, King's College, London; [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford; and [3]Center for Statistical Genetics, University of Michigan, Ann Arbor

- Multivariate Regression Model
- Weighted Least Squares Estimation
- Weight matrix based on IBD information
  - Dependent variables = IBD
  - Independent variables = Trait

General approach

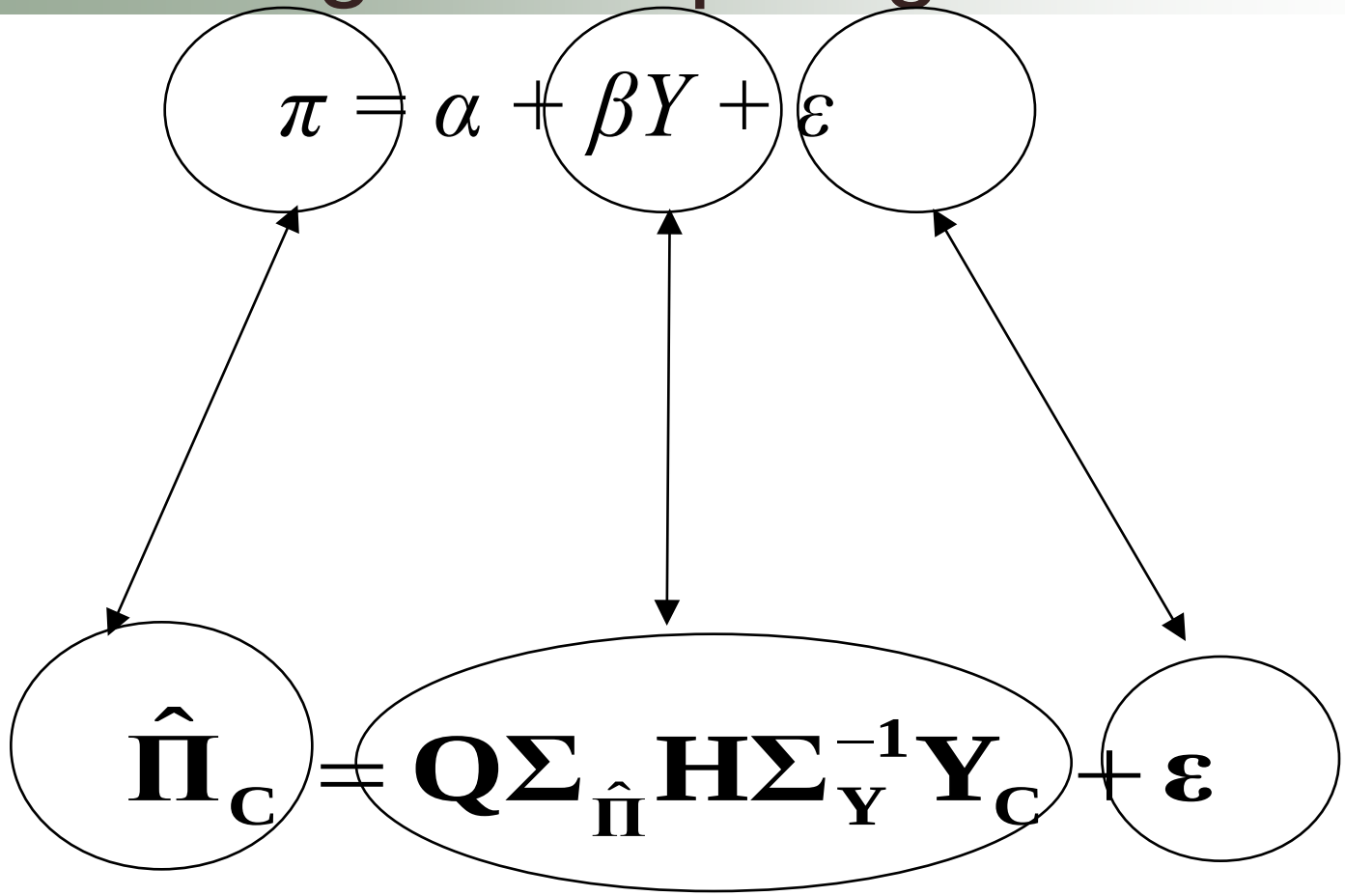
- Standard regression based methods model trait (D^2, S^2) in terms of estimated IBD status

$$Y = \alpha + \beta\pi + \varepsilon$$

- Instead IBD estimate is regressed on trait value

$$\pi = \alpha + \beta Y + \varepsilon$$

Extend to general pedigrees



Dependent Variables

- Estimated IBD sharing of all pairs of relatives
- Example:

$$\hat{\boldsymbol{\Pi}} = \begin{bmatrix} \hat{\pi}_{12} \\ \hat{\pi}_{13} \\ \hat{\pi}_{14} \\ \hat{\pi}_{23} \\ \hat{\pi}_{24} \\ \hat{\pi}_{34} \end{bmatrix}$$

Independent Variables

- Squares and cross-products
  - (equivalent to non-redundant squared sums and differences)
- Example

$$\mathbf{Y} = \begin{bmatrix} x_1x_2 \\ x_1x_3 \\ x_1x_4 \\ x_2x_3 \\ x_2x_4 \\ x_3x_4 \\ x_1x_1 \\ x_2x_2 \\ x_3x_3 \\ x_4x_4 \end{bmatrix}$$

Estimation

For a family, regression model is

$$\hat{\mathbf{\Pi}}_{\mathbf{C}} = \mathbf{Q}\boldsymbol{\Sigma}_{\hat{\mathbf{\Pi}}}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_{\mathbf{C}} + \boldsymbol{\varepsilon}$$

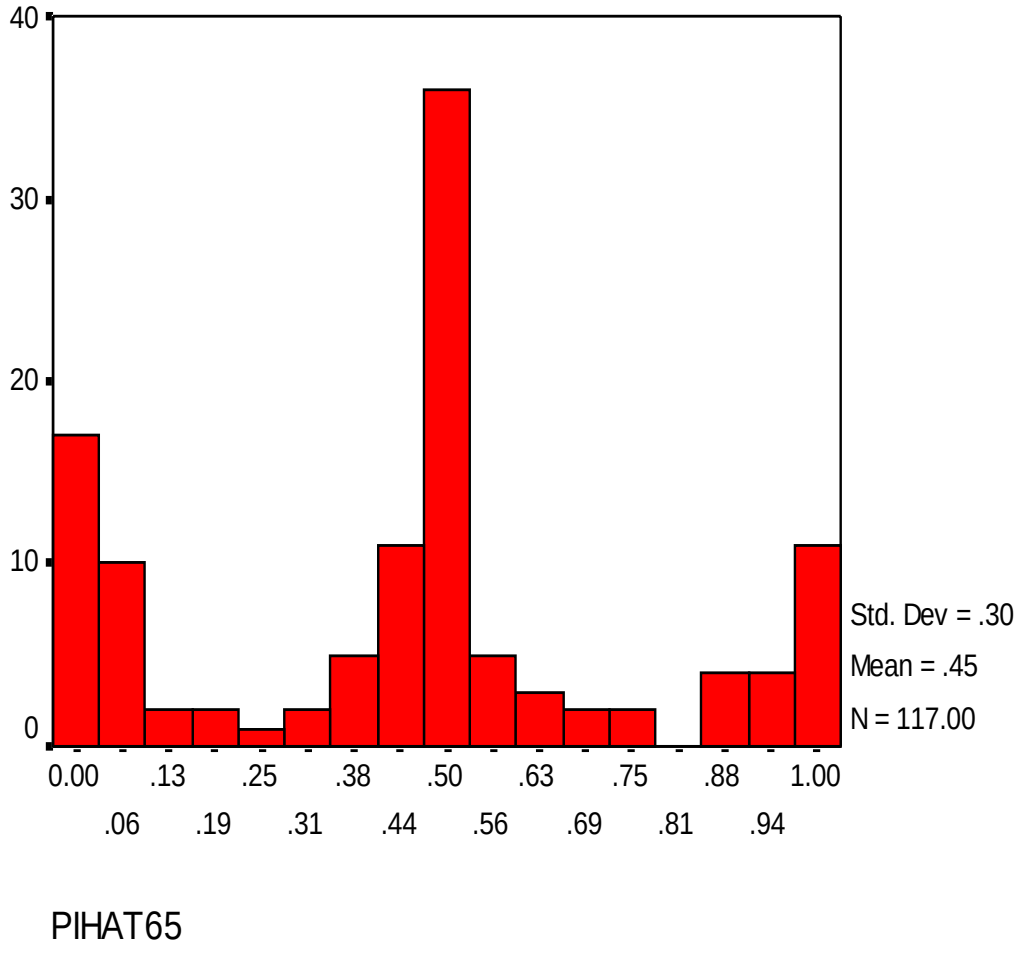
Estimate Q by weighted least squares, and obtain sampling variance, family by family

Combine estimates across families, inversely weighted by their variance, to give overall estimate, and its sampling variance

Why is that better?

- Regression methods assume that the dependant variable (left hand side) is normally distributed

Distribution of pi-hat



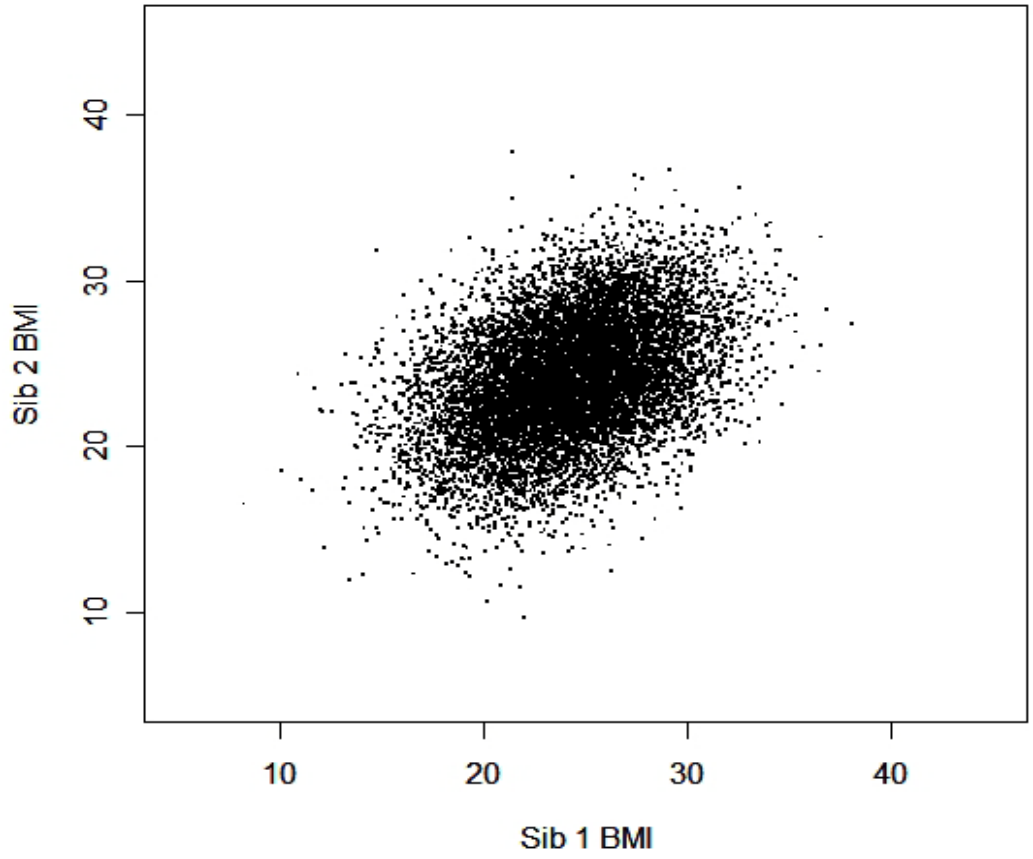
Why is that better?

- But “central limit theorem” works well when data a symmetric with mode in the centre
- In a general pedigree, sib-pairs provide the most information on linkage
- IBD under null hypothesis (with complete inheritance information)
  - 0 – 25%
  - 0.5 – 50%
  - 1 – 25%

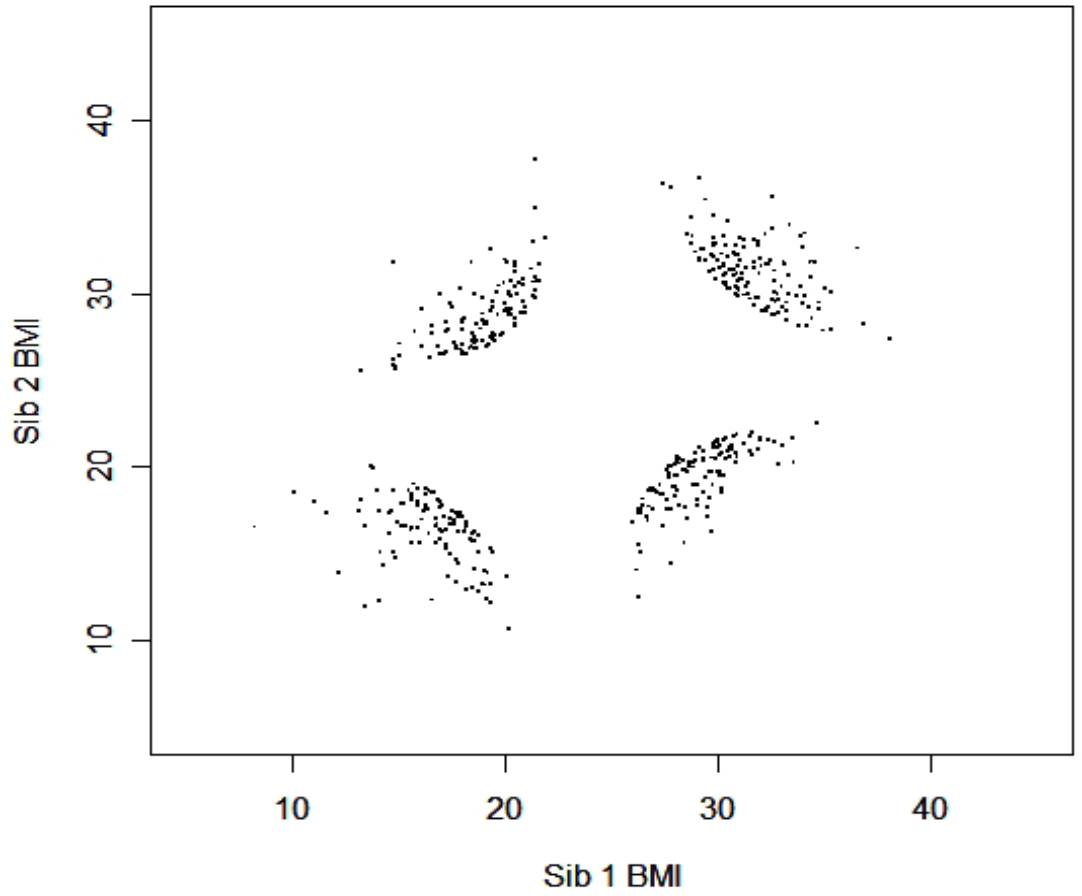
Selected Samples

- Merlin-regress is particularly suited to the analysis of selected samples
  - Ordinary variance component analysis (e.g. using Merlin) gives biased QTL estimates
  - Merlin-regress is designed to be robust to data selection

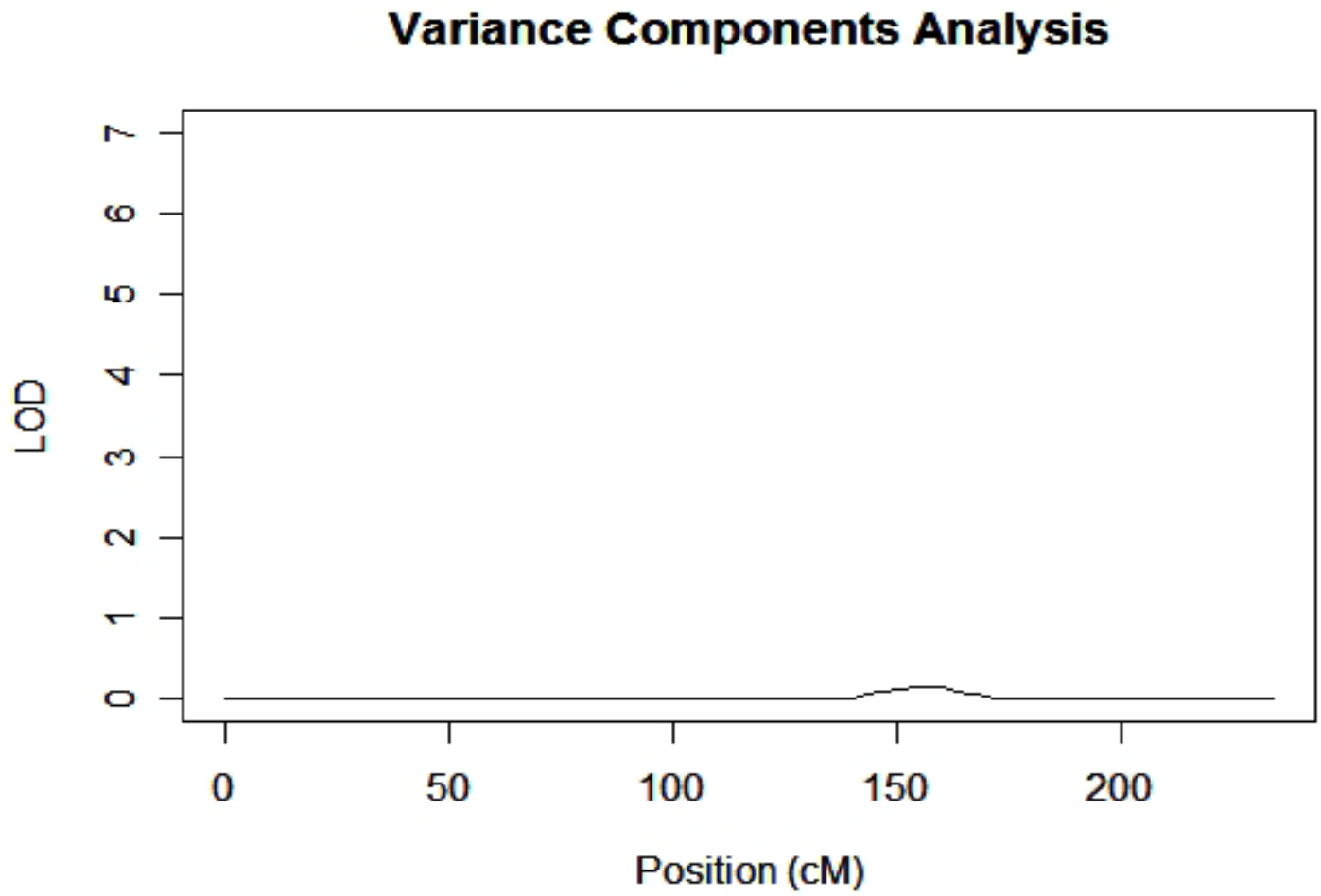
Example Data – BMI 10000 pairs



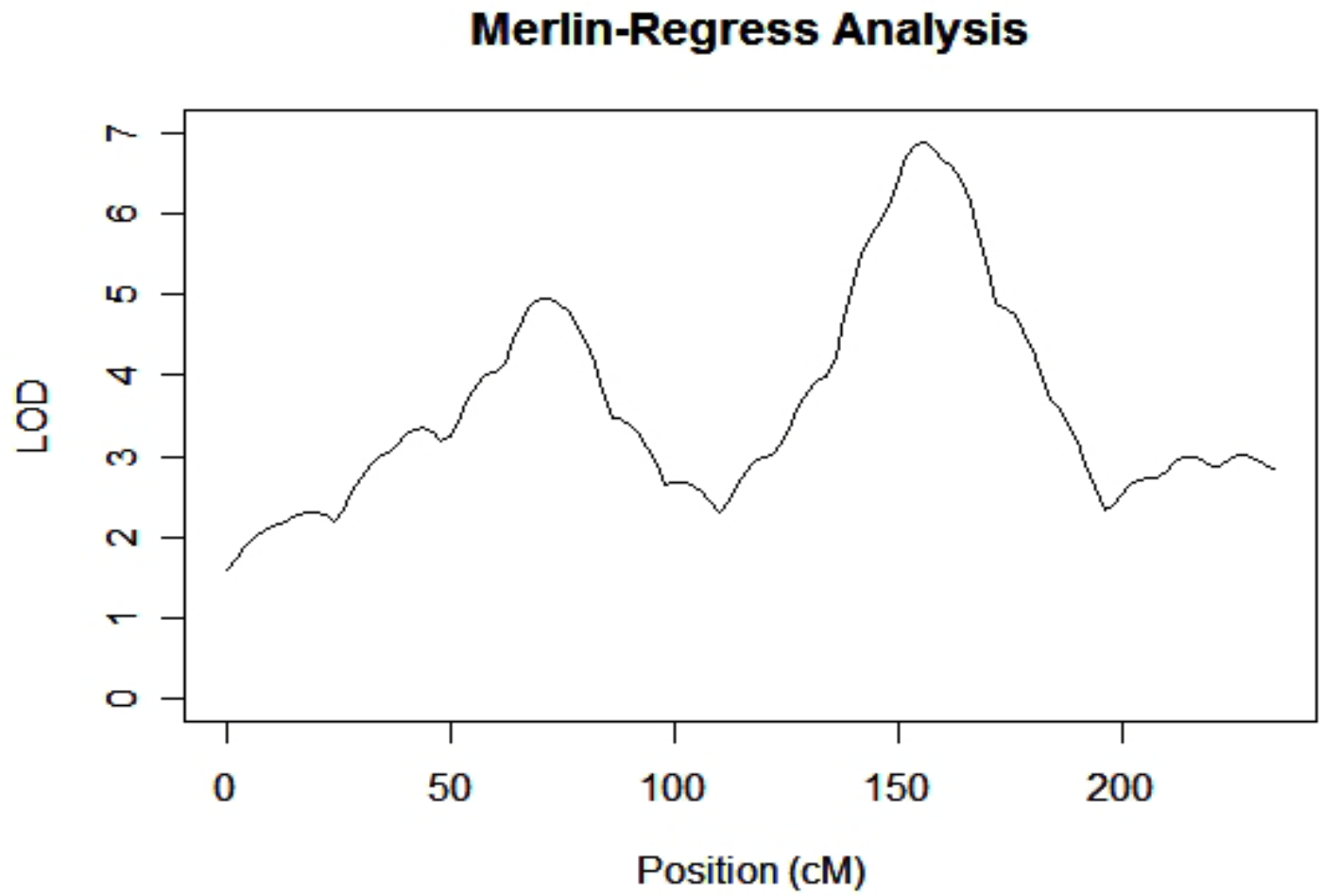
Selected Sample – 500 pairs



Results – VC


Variance Components Analysis
LOD
0
1
2
3
4
5
6
7
0
50
100
150
200
Position (cM)

Results – Merlin-Regress



Practical #4: running regress

```
./merlin-regress -x -9999.000 -p linkage.ped -d
   linkage.dat -m linkage.map --mean ? --variance ?
    --heritability ? > linkage2.out
```

`merlin-regress`

Calls up the programme

`--vc -x -9999.000`

Specifies VC linkage and the missing value

`-p linkage.ped -d linkage.dat -m linkage.map`

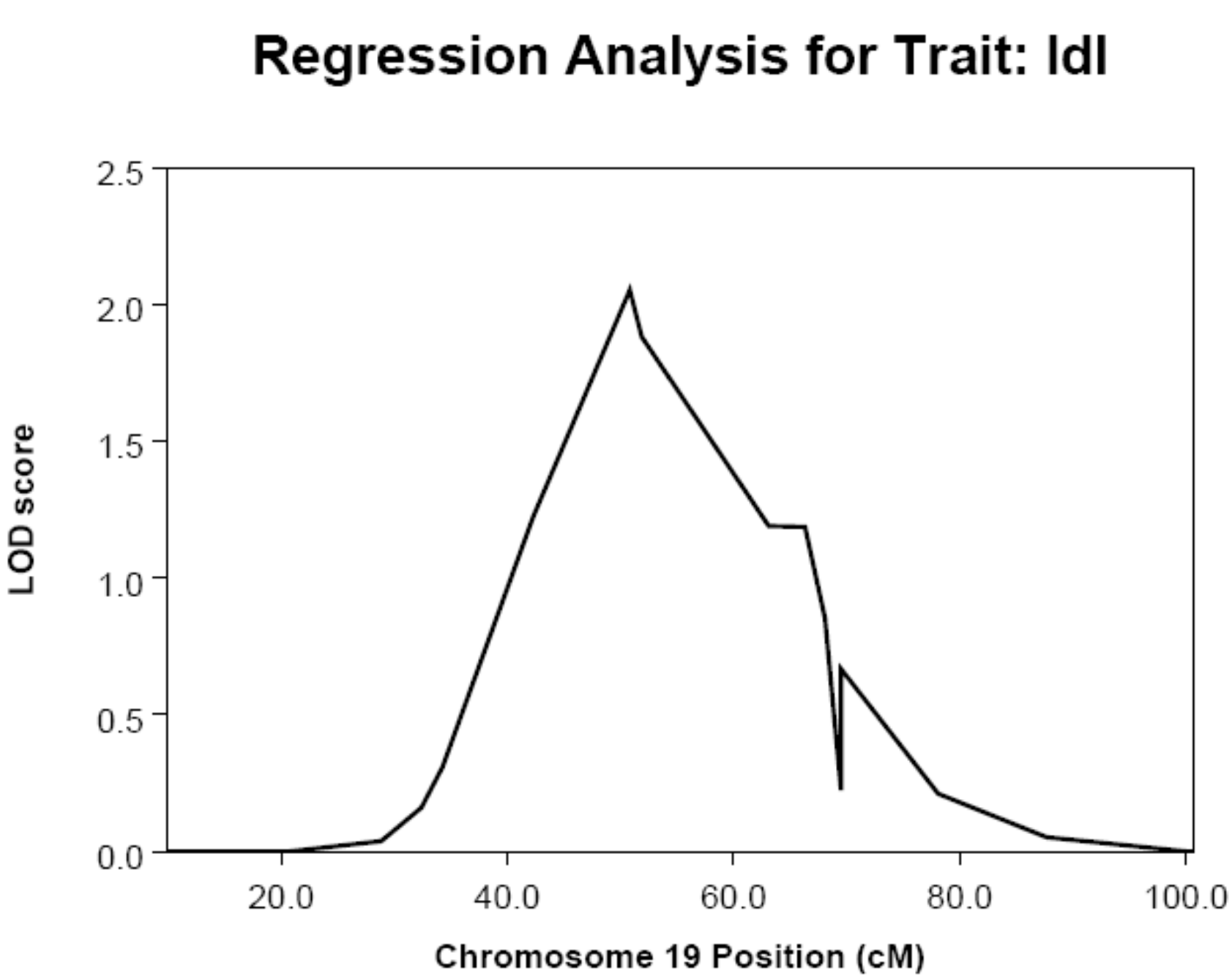
Identify the .ped, .dat, and .map files

`--mean ? --variance ? --heritability ?`

Specify the mean, variance, and heritability from the *whole* population (Pedstats)

`> linkage.out`

Send the output to a text file


Regression Analysis for Trait: ldl
LOD score
2.5
2.0
1.5
1.0
0.5
0.0
20.0
40.0
60.0
80.0
100.0
Chromosome 19 Position (cM)