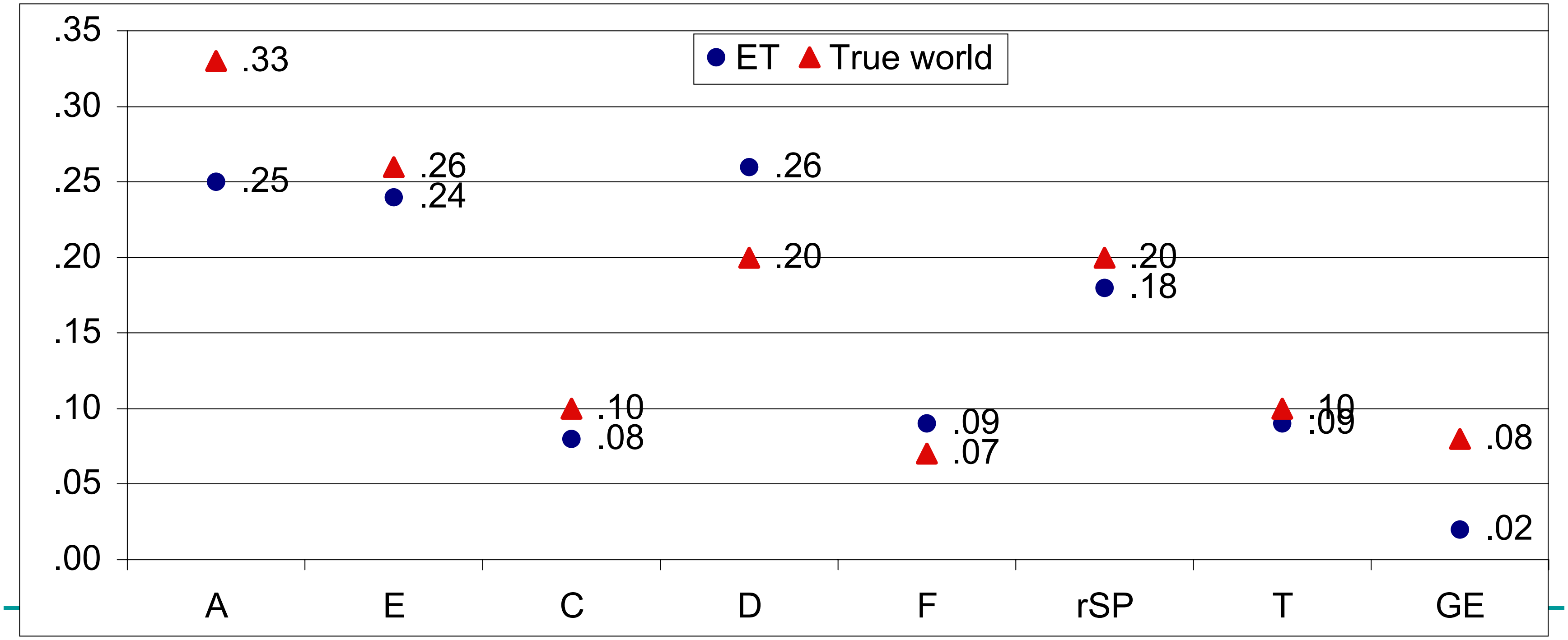
Corrected version of the model building ppt in hmaes/a21/maes/Extended_Pedigrees/Model_building.ppt



Introduction to Linkage

Sarah Medland - Boulder 2008

Aim of QTL mapping…

LOCALIZE and then IDENTIFY a locus that regulates a trait (QTL)

  - *Locus: Nucleotide or sequence of nucleotides with variation in the population, with different variants associated with different trait levels.*

- Linkage
  - localize region of the genome where a QTL that regulates the trait is likely to be harboured
  - Family-specific phenomenon: Affected individuals in a family share the same ancestral predisposing DNA segment at a given QTL
- Association
  - identify a QTL that regulates the trait
  - Population-specific phenomenon: Affected individuals in a population share the same ancestral predisposing DNA segment at a given QTL

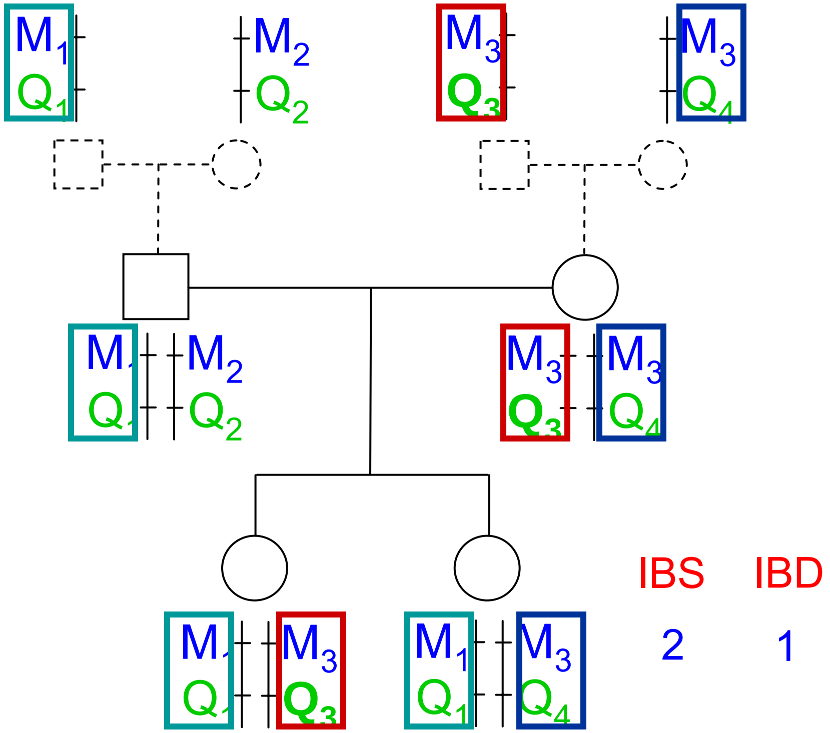
Genotypic similarity – basic principals

- Loci that are close together are more likely to be inherited together than loci that are further apart
- Loci are likely to be inherited in context – ie with their surrounding loci
- Because of this, knowing that a loci is transmitted from a common ancestor is more informative than simply observing that it is the same allele

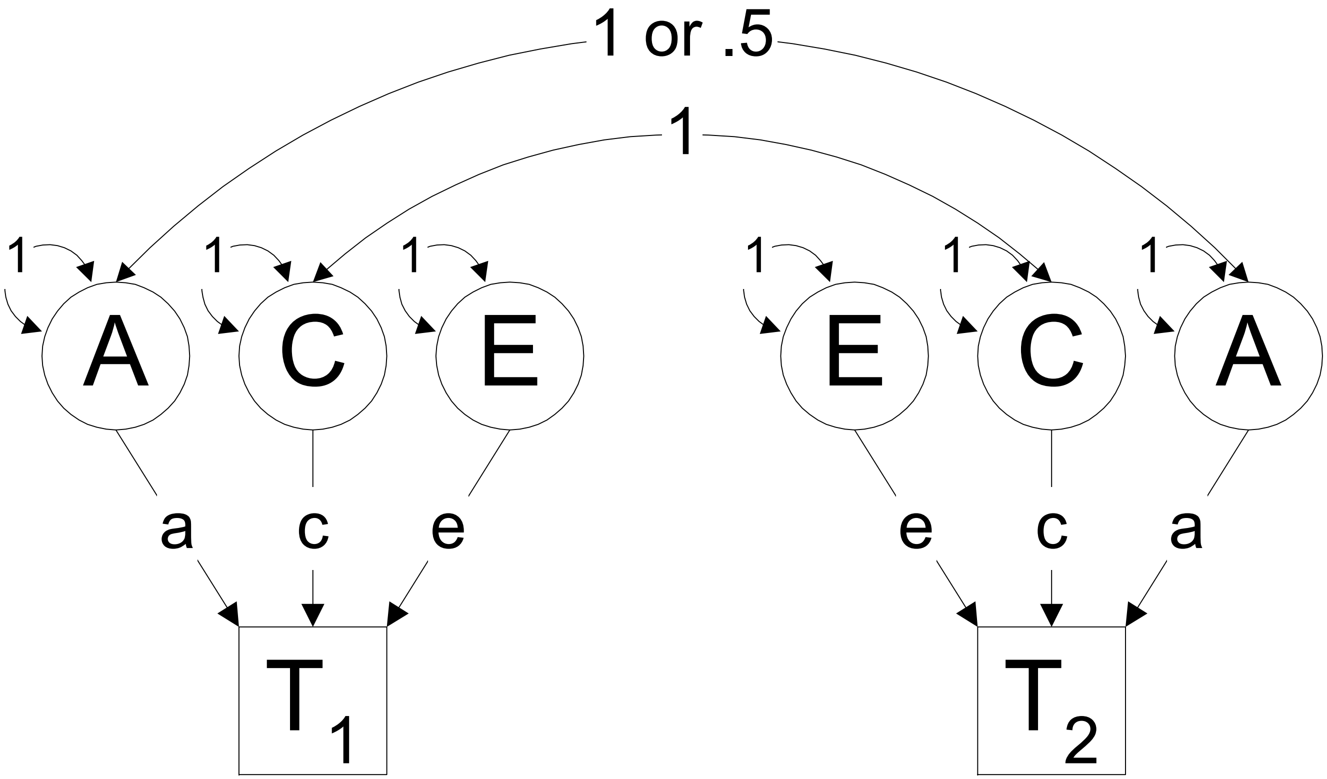
Genotypic similarity between relatives

- IBS Alleles shared Identical By State "look the same", may have the same DNA sequence but they are not necessarily derived from a known common ancestor - focus for association

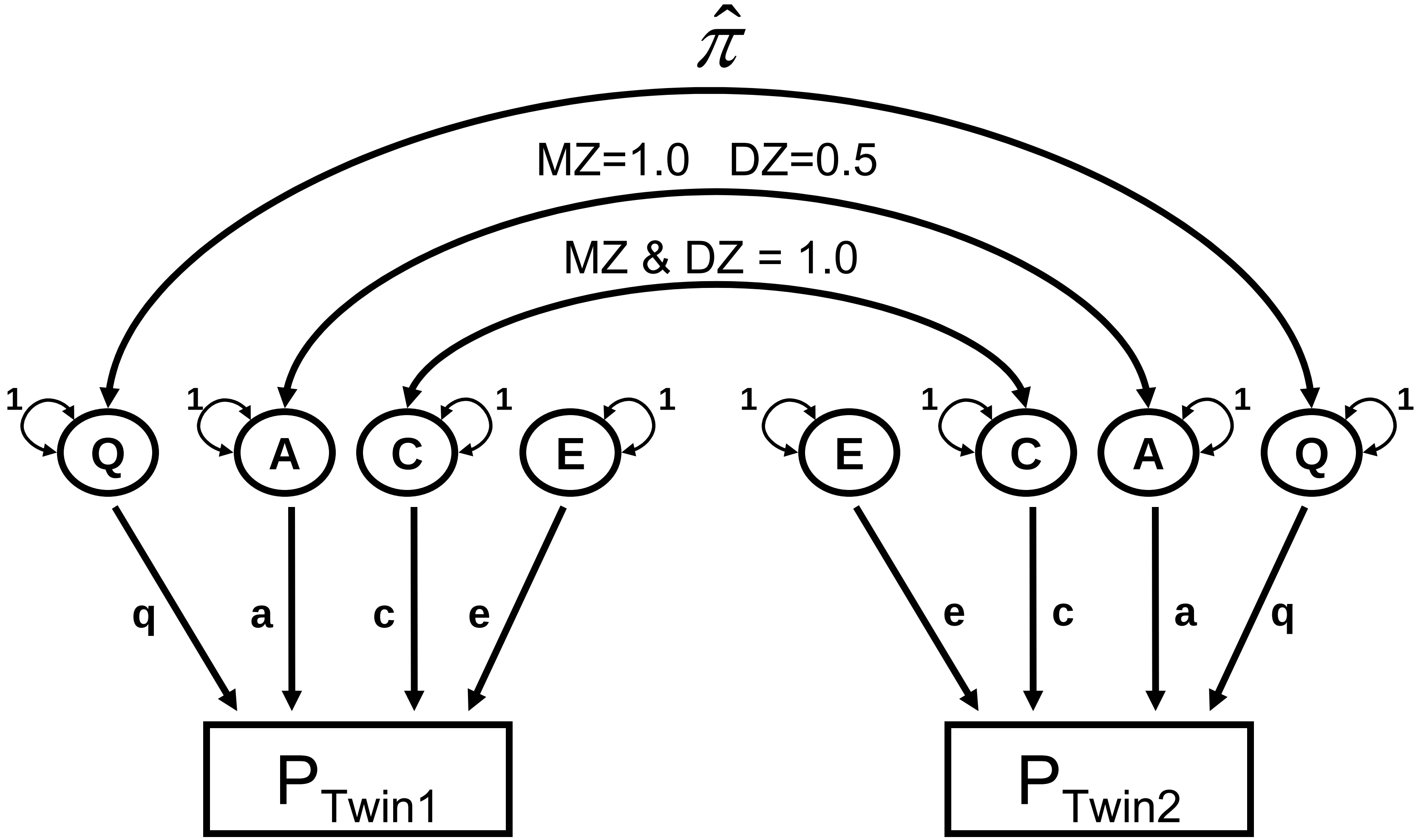
- IBD Alleles shared Identical By Descent are a copy of the same ancestor allele - focus for linkage



- In biometrical modeling A is correlated at 1 for MZ twins and .5 for DZ twins
  - .5 is the average genome-wide sharing of genes between full siblings (DZ twin relationship)



- In linkage analysis we will be estimating an additional variance component Q
  - For each locus under analysis the coefficient of sharing for this parameter will vary for each pair of siblings
    - The coefficient will be the probability that the pair of siblings have both inherited the same alleles at a given locus from a common ancestor $\hat{\pi}$


$\hat{\pi}$
MZ=1.0   DZ=0.5
MZ & DZ = 1.0
1
Q
1
A
C
1
E
1
1
E
1
C
A
1
Q
1
q
a
c
e
e
c
a
q
P_{Twin1}
P_{Twin2}

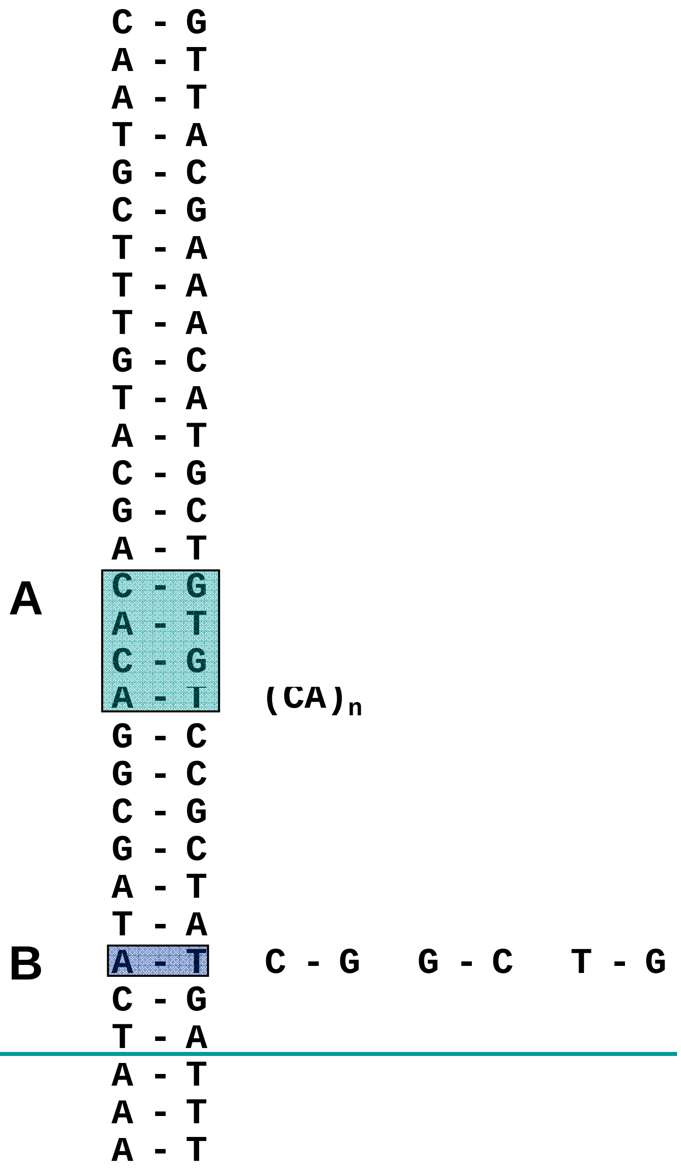
DNA polymorphisms

▷ Microsatellites

- >100,000
- Many alleles, $(CA)_n$
- Very Informative
- Not intended to be functional variants
- Used in linkage

▷ SNPs

- 10,054,521 (25 Jan '05)
- 10,430,753 (11 Mar '06)
- Most with 2 alleles (up to 4)
- Not very informative
- Intended to by functional variants
- Used in association or linkage



```
     C - G
     A - T
     A - T
     T - A
     G - C
     C - G
     T - A
     T - A
     T - A
     G - C
     T - A
     A - T
     C - G
     G - C
     A - T
A    C - G
     A - T
     C - G
     A - T   (CA)n
     G - C
     G - C
     C - G
     G - C
     A - T
     T - A
B    A - T   C - G   G - C   T - G
     C - G
     T - A
     A - T
     A - T
     A - T
```

Microsatellite data

- Ideally positioned at equal genetic distances across chromosome
- Mostly di/tri nucleotide repeats
- Raw data consists of allele lengths/calls (bp)

Binning

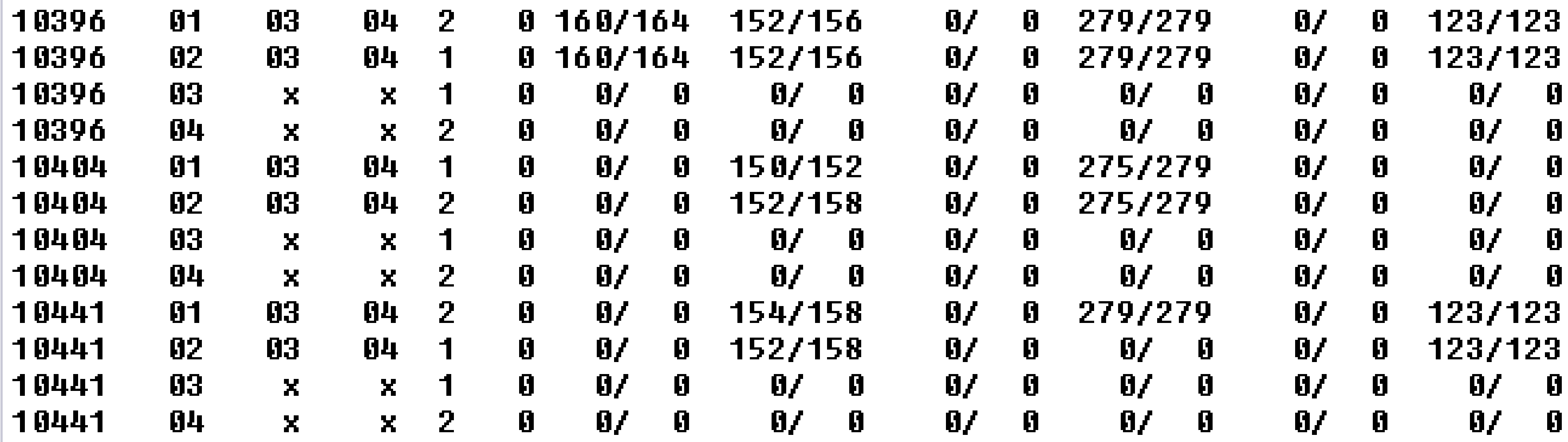
- Raw allele lengths are converted to allele numbers or lengths
  - Example:D1S1646 tri-nucleotide repeat size range130-150
    - Logically: Work with binned lengths
    - Commonly: Assign allele 1 to 130 allele, 2 to 133 allele …

Error checking

- After binning check for errors
  - Family relationships (GRR, Rel-pair)
  - Mendelian Errors (Sib-pair)
  - Double Recombinants (MENDEL, ASPEX, ALEGRO)
- An iterative process

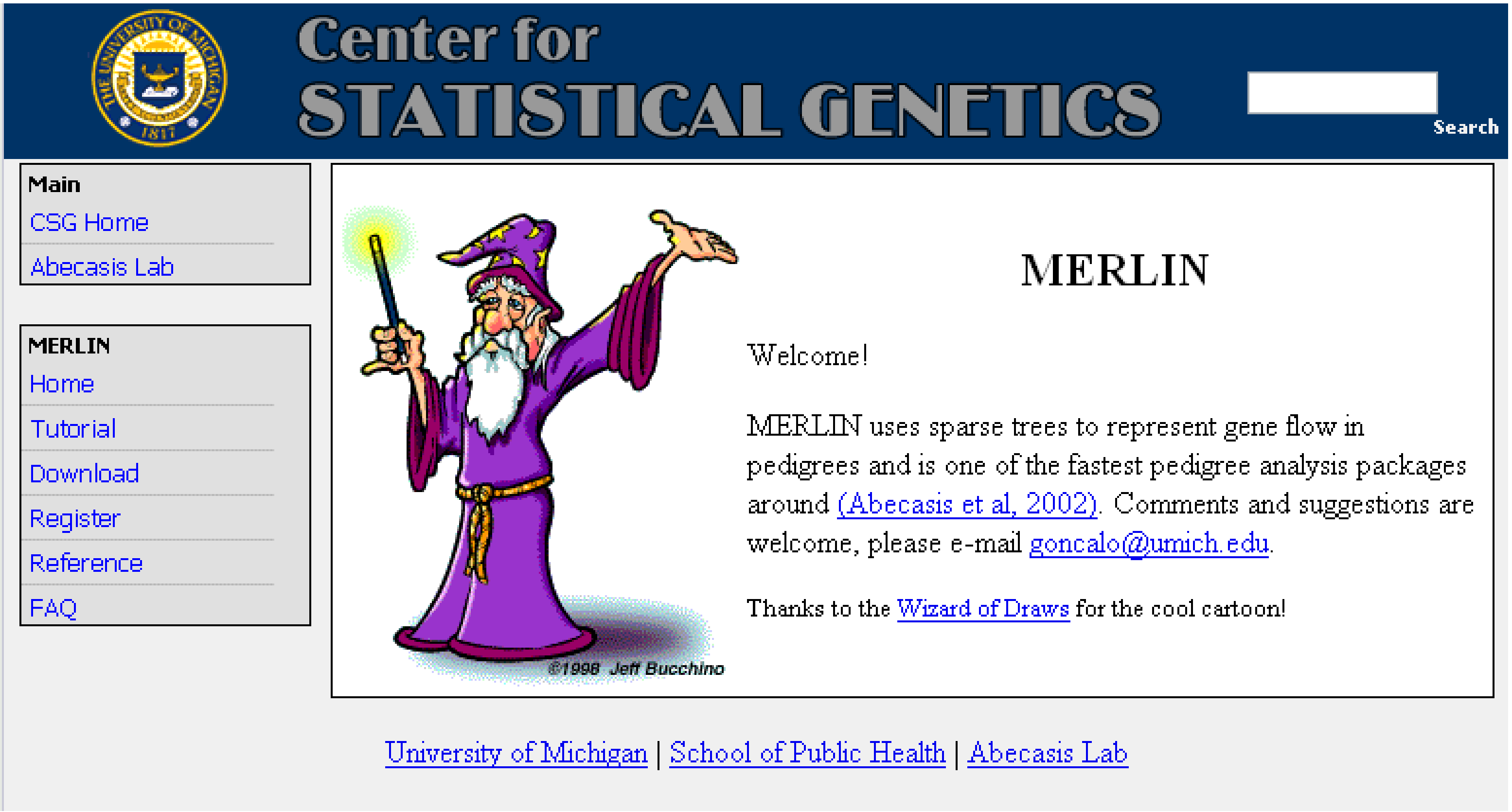
‘Clean’ data

- ped file
  - Family, individual, father, mother, sex, dummy, genotypes
  - The ped file is used with ‘map’ files to obtain estimates of genotypic sharing between relatives at each of the locations under analysis - MERLIN



```
10396   01   03   04   2    0 160/164  152/156     0/   0  279/279     0/   0  123/123
10396   02   03   04   1    0 160/164  152/156     0/   0  279/279     0/   0  123/123
10396   03    x    x   1    0    0/   0    0/   0  0/   0     0/   0  0/   0     0/   0
10396   04    x    x   2    0    0/   0    0/   0  0/   0     0/   0  0/   0     0/   0
10404   01   03   04   1    0    0/   0  150/152  0/   0  275/279  0/   0     0/   0
10404   02   03   04   2    0    0/   0  152/158  0/   0  275/279  0/   0     0/   0
10404   03    x    x   1    0    0/   0    0/   0  0/   0     0/   0  0/   0     0/   0
10404   04    x    x   2    0    0/   0    0/   0  0/   0     0/   0  0/   0     0/   0
10441   01   03   04   2    0    0/   0  154/158  0/   0  279/279  0/   0  123/123
10441   02   03   04   1    0    0/   0  152/158  0/   0     0/   0  0/   0  123/123
10441   03    x    x   1    0    0/   0    0/   0  0/   0     0/   0  0/   0     0/   0
10441   04    x    x   2    0    0/   0    0/   0  0/   0     0/   0  0/   0     0/   0
```

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



MERLIN

Welcome!

MERLIN uses sparse trees to represent gene flow in pedigrees and is one of the fastest pedigree analysis packages around (Abecasis et al, 2002). Comments and suggestions are welcome, please e-mail goncalo@umich.edu.

Thanks to the Wizard of Draws for the cool cartoon!

University of Michigan | School of Public Health | Abecasis Lab

More on IBD

- Chapter 8 - Abecasis
  - Neale, Ferreira, Medland, Posthuma (2007) Statistical Genetics: Gene mapping through linkage and Association
- Advanced workshop

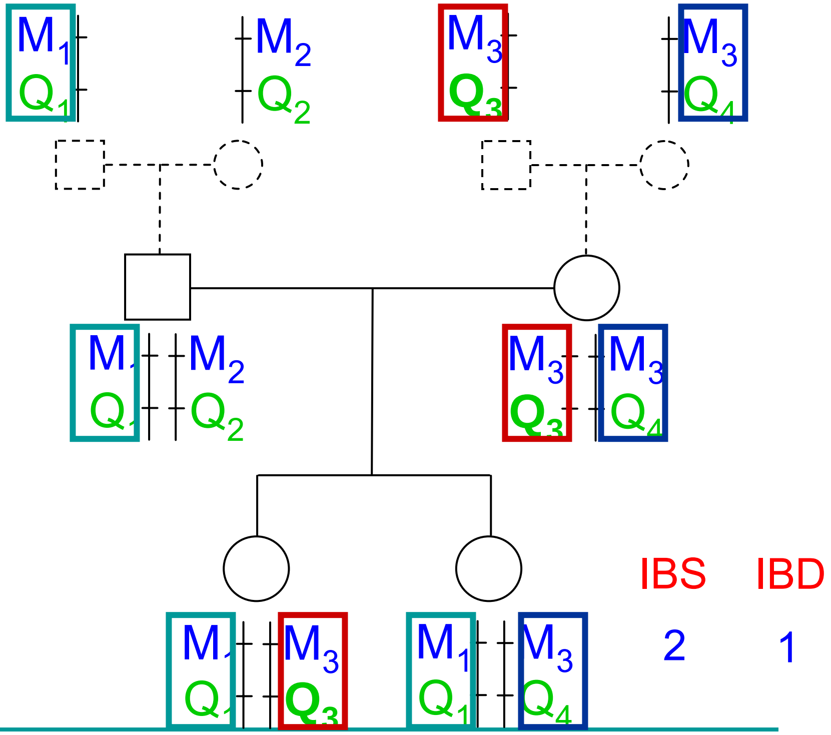
Genotypic similarity between relatives

- IBD Alleles shared Identical By Descent are a copy of the same ancestor allele

Pairs of siblings may share 0, 1 or 2 alleles IBD

The probability of a pair of relatives being IBD is called pi-hat

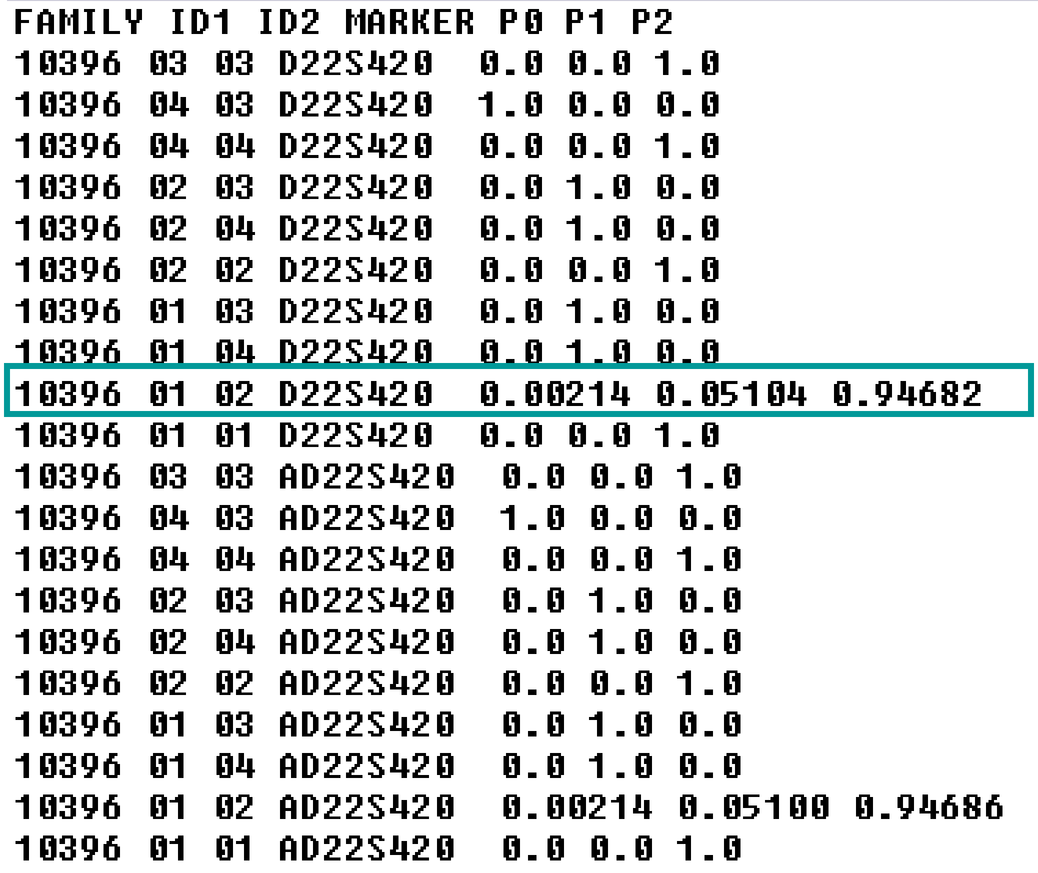
$$\hat{\pi} = p(IBD2) + .5 * p(IBD1)$$



Estimating genotypic sharing…

- Output

$$\hat{\pi} = p(IBD2) + .5 * p(IBD1)$$



```
FAMILY ID1 ID2 MARKER P0 P1 P2
10396 03 03 D22S420  0.0 0.0 1.0
10396 04 03 D22S420  1.0 0.0 0.0
10396 04 04 D22S420  0.0 0.0 1.0
10396 02 03 D22S420  0.0 1.0 0.0
10396 02 04 D22S420  0.0 1.0 0.0
10396 02 02 D22S420  0.0 0.0 1.0
10396 01 03 D22S420  0.0 1.0 0.0
10396 01 04 D22S420  0.0 1.0 0.0
10396 01 02 D22S420  0.00214 0.05104 0.94682
10396 01 01 D22S420  0.0 0.0 1.0
10396 03 03 AD22S420  0.0 0.0 1.0
10396 04 03 AD22S420  1.0 0.0 0.0
10396 04 04 AD22S420  0.0 0.0 1.0
10396 02 03 AD22S420  0.0 1.0 0.0
10396 02 04 AD22S420  0.0 1.0 0.0
10396 02 02 AD22S420  0.0 0.0 1.0
10396 01 03 AD22S420  0.0 1.0 0.0
10396 01 04 AD22S420  0.0 1.0 0.0
10396 01 02 AD22S420  0.00214 0.05100 0.94686
10396 01 01 AD22S420  0.0 0.0 1.0
```

$$\hat{\pi} = ?$$

=.94682 + .5*.05104

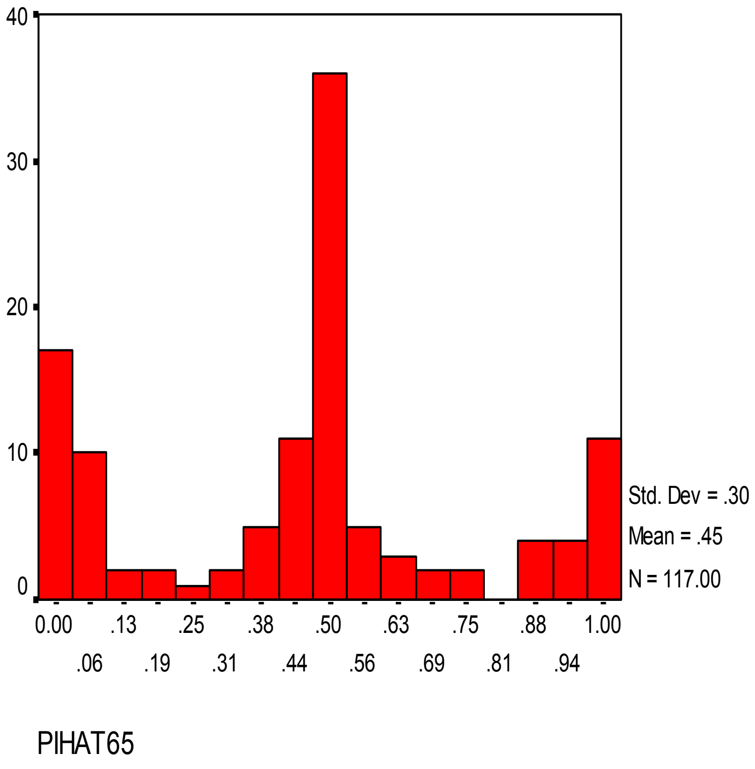
=.972

Identity by Descent (IBD) in sibs

		Sib1			
		AC	AD	BC	BD
Sib 2	AC	2	1	1	0
	AD	1	2	0	1
	BC	1	0	2	1
	BD	0	1	1	2

- Four parental marker alleles: A-B and C-D
- Two siblings can inherit 0, 1 or 2 alleles IBD
- IBD 0:1:2 = 25%:50%:25%
- Derivation of IBD probabilities at one marker (Haseman & Elston 1972

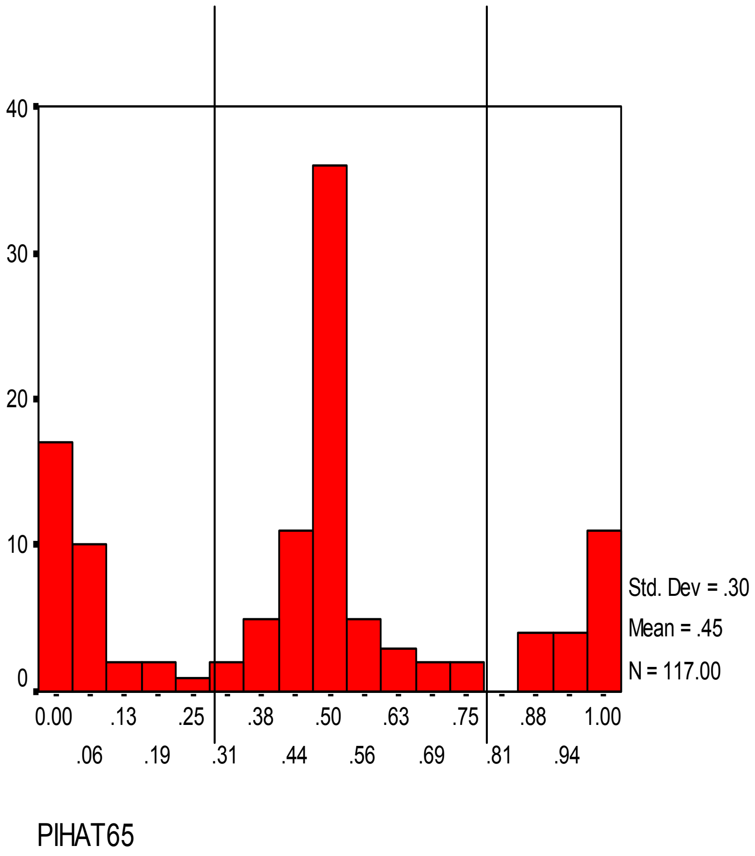
Distribution of pi-hat



- Adult Dutch DZ pairs: distribution of pi-hat $\hat{\pi}$ at 65 cM on chromosome 19
- Model resemblance (e.g. correlations, covariances) between sib pairs, or DZ twins, as a function of DNA marker sharing at a particular chromosomal location

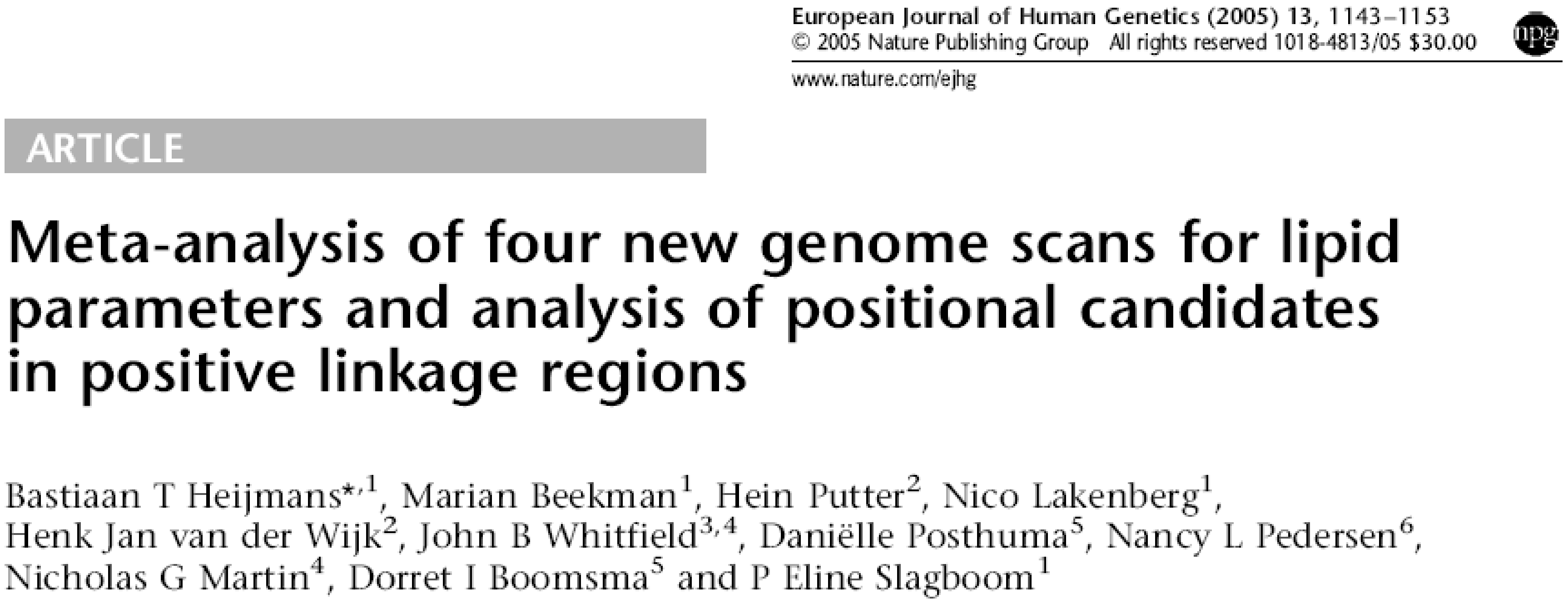
Linkage with full siblings (DZ twins)

Partitioned twin analysis



- Adult Dutch DZ pairs: distribution of pi-hat $\hat{\pi}$ at 65 cM on chromosome 19
  - ❑ $\hat{\pi} < 0.25$: IBD=0 group
  - ❑ $\hat{\pi} > 0.75$: IBD=2 group
  - ❑ others: IBD=1 group
  - ❑ pi65cat= (0,1,2)

http://www.nature.com/ejhg/journal/v13/n10/pdf/5201466a.pdf



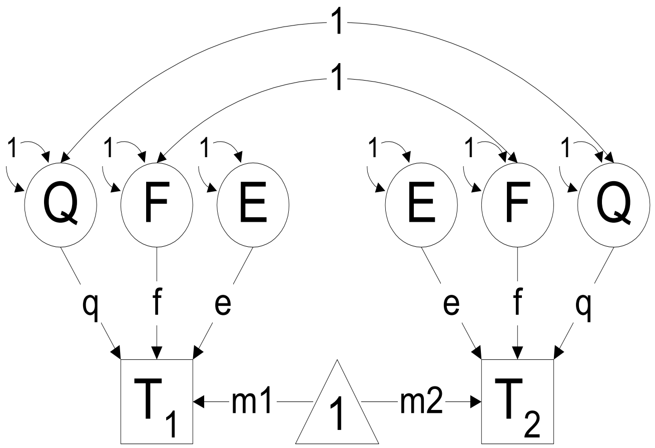
European Journal of Human Genetics (2005) 13, 1143–1153
www.nature.com/ejhg

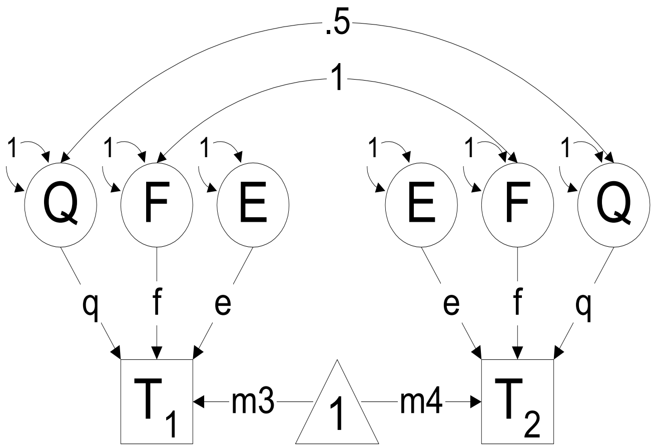
ARTICLE

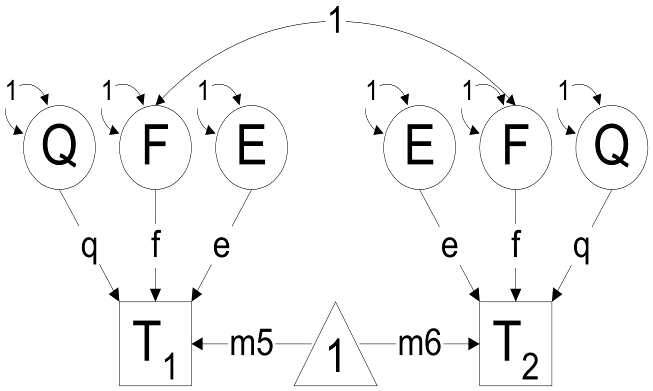
Meta-analysis of four new genome scans for lipid parameters and analysis of positional candidates in positive linkage regions

Bastiaan T Heijmans*[,1], Marian Beekman[1], Hein Putter[2], Nico Lakenberg[1], Henk Jan van der Wijk[2], John B Whitfield[3,4], Daniëlle Posthuma[5], Nancy L Pedersen[6], Nicholas G Martin[4], Dorret I Boomsma[5] and P Eline Slagboom[1]

DZ by IBD status







- Variance = Q + F + E
- Covariance = πQ + F

partitioned.mx

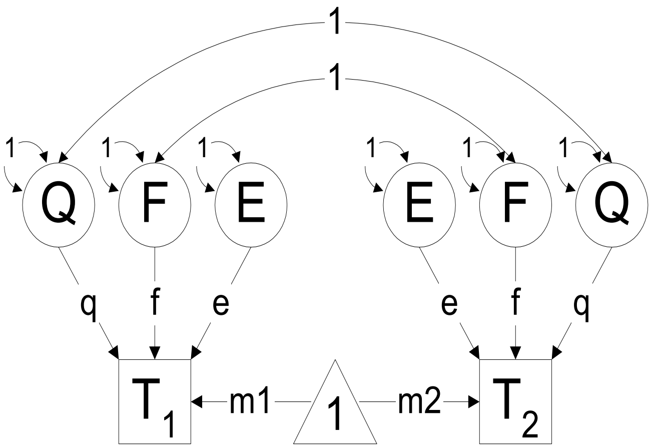
```
! Estimate Genetic (QTL) and Environmental Components - FEQ model
! Dutch Adult Twins: Lipid levels (position 65 cM chromosome 19)
#define $var ldl
!3 variables in the file ldl apob apoe
#define nvar 1
#define nvarx2 2
#NGroups 5

G1: Model Parameters
Calculation
 Begin Matrices;
  X Lower nvar nvar Free    ! residual familial path coefficients
  Z Lower nvar nvar Free    ! nonshared environment path coefficients
  T Lower nvar nvar Free    ! QTL path coefficients
  H Full 1 1
 End Matrices;
  Matrix H .5
  Start .3 All
 Begin Algebra;
  F=X*X';                   ! residual familial variance components
  E=Z*Z';                   ! nonshared environment variance components
  Q=T*T';                   ! QTL variance components
 End Algebra;
 Option Rsiduals
End
```

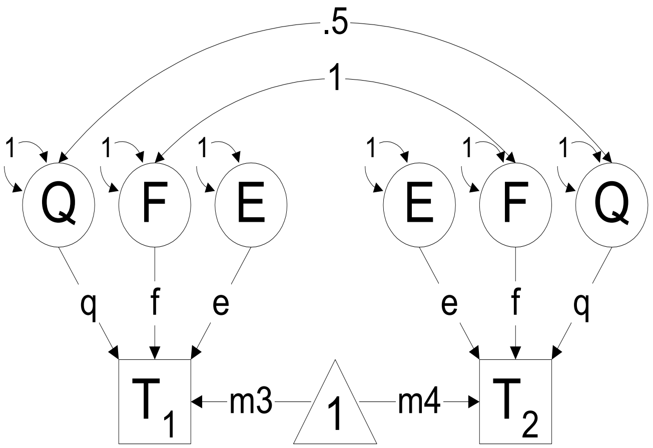
partitioned.mx

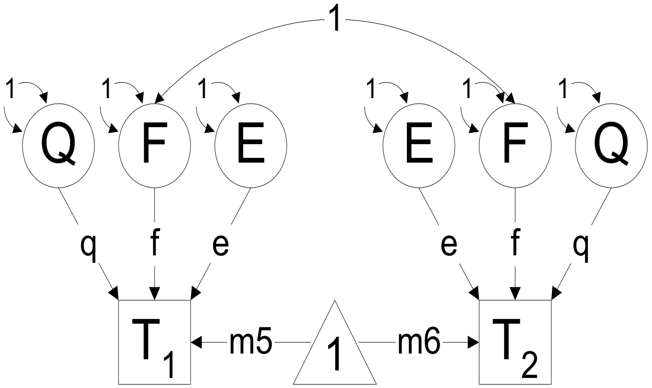
3 Data groups

```
G2: DZ IBD2 twins
 Data NInput=18
  Rectangular File=DutchDZ.rec
  Labels zyg  sex1 age1 med1 t1ldl t1apob t1lnapoe  sex2 age2 med2 t2ldl t2apob t2lnapoe
   ibd0_65 ibd1_65 ibd2_65 pihat65 pi65cat
  Select if pi65cat =2;
  Select
   t1$var
   t2$var ;
 Begin Matrices = Group 1;
  M Full nvar nvarx2 Free
  K Full 1 1                ! correlation of QTL effects
 End Matrices;
  Matrix M 4 4
  Matrix K 1
 Means M;
 Covariance
    F+Q+E | F+K@Q _
    F+K@Q | F+Q+E;
End
```

DZ by IBD status







- Variance = Q + F + E
- Covariance = πQ + F

Covariance Statements

```
G2: DZ IBD2 twins
  Matrix K 1
 Covariance
    F+Q+E | F+K@Q _
    F+K@Q | F+Q+E;

G3: DZ IBD1 twins
  Matrix K .5
 Covariance
    F+Q+E | F+K@Q _
    F+K@Q | F+Q+E;

G4: DZ IBD0 twins
 Covariance
    F+Q+E | F_
    F     | F+Q+E;
```

partitioned.mx

```
G5: Standardization
 Calculation
 Begin Matrices = Group 1;
 Begin Algebra;
  V=F+E+Q;                  ! total variance
  P=F|E|Q;                  ! concatenate parameter estimates
  S=P@V~;                   ! standardized parameter estimates
 End Algebra;
  Label Col P f^2 e^2 q^2
  Label Col S f^2 e^2 q^2
!FEQ model
 Interval S 1 1 - S 1 3
 Option Rsiduals Iterations=5000 NDecimals=4
 Option Multiple Issat
End

! Test for QTL
Drop T 1 1 1
Exit
```

Variance Components FEQ

		f^2	e^2	q^2
LDL		0	.2263	.7737
ApoB				
lnApoE				

Chi-square Tests for QTL

	DZ pairs (df=1)	
	Chi-square	Mx P-value
LDL	12.25	0.00004
ApoB		
lnApoE		

Your task…

- The data file has 3 traits
  - Labels zyg sex1 age1 med1 t1ldl t1apob t1lnapoe sex2 age2 med2 t2ldl t2apob t2lnapoe
- Change the variable being analyses
  - Left side of the room (your left) - apob
  - Right side of the room (your right) - lnapoe

Variance Components FEQ

		f^2	e^2	q^2
LDL		0	.2263	.7737
ApoB		.2712	.4136	.3152
lnApoE		.1885	.1607	.6508

Chi-square Tests for QTL

	DZ pairs (df=1)	
	Chi-square	Mx p-value
LDL	12.25	0.00047
ApoB	1.95	0.163
lnApoE	12.45	0.00042

Converting chi-squares to p values

- Complicated
  - Distribution of genotypes and phenotypes
  - Boundary problems

- For univariate linkage analysis
  (where you have 1 QTL estimate)
  p(linkage)= $\chi_1^2 / 2$

Chi-square Tests for QTL

	DZ pairs (df=1)	
	Chi-square	Asymptotic p-value
LDL	12.25	0.00024
ApoB	1.95	0.08150
lnApoE	12.45	0.00021

Converting chi-squares to LOD scores

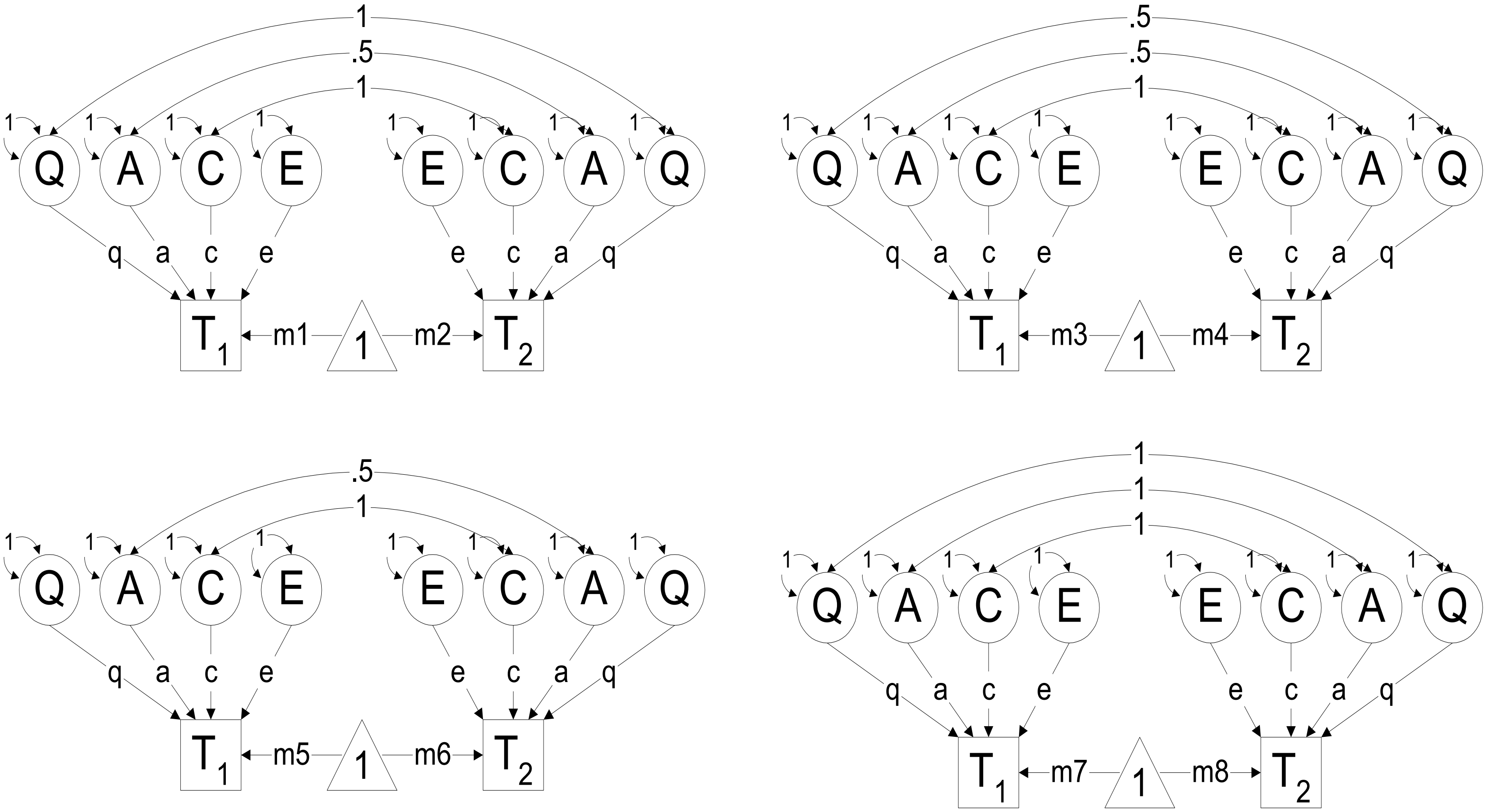
- For univariate linkage analysis
  (where you have 1 QTL estimate)
  $X^2/4.6 = \text{LOD}$

Adding MZ twins

Partitioned+MZ.mx

- Adding MZ pairs allows you to partitioned F into A and C
- Do MZ contribute to linkage?
- In what ways do MZs help in a linkage analysis?

DZ by IBD status + MZ



Covariance Statements +MZ

```
G2: DZ IBD2 twins
  Matrix K 1
 Covariance
    A+C+Q+E | H@A+C+K@Q _
    H@A+C+K@Q | A+C+Q+E;

G3: DZ IBD1 twins
  Matrix K .5
 Covariance
    A+C+Q+E | H@A+C+K@Q _
    H@A+C+K@Q | A+C+Q+E;

G4: DZ IBD0 twins
 Covariance
    A+C+Q+E | H@A+C_
    H@A+C   | A+C+Q+E;

G5: MZ twins
 Covariance
    A+C+Q+E | A+C+Q _
    A+C+Q   | A+C+Q+E;
```

Variance Components ACEQ

	a^2	c^2	e^2	q^2
LDL	0.04 (0 – 0.39)	0 (0 – 0.27)	0.21 (0.15 – 0.29)	0.75 (0.37 – 0.84)
ApoB				
lnApoE				

Chi-square Tests for QTL

	DZ+MZ pairs (df=1)	
	Chi-square	Asymptotic p-value
LDL	12.561	0.0002
ApoB		
lnApoE		

Your task…

- The data file has 3 traits
  - Labels zyg sex1 age1 med1 t1ldl t1apob t1lnapoe sex2 age2 med2 t2ldl t2apob t2lnapoe
- Change the variable being analyses
  - Left side of the room (your left) - apob
  - Right side of the room (your right) - lnapoe

Variance Components ACEQ

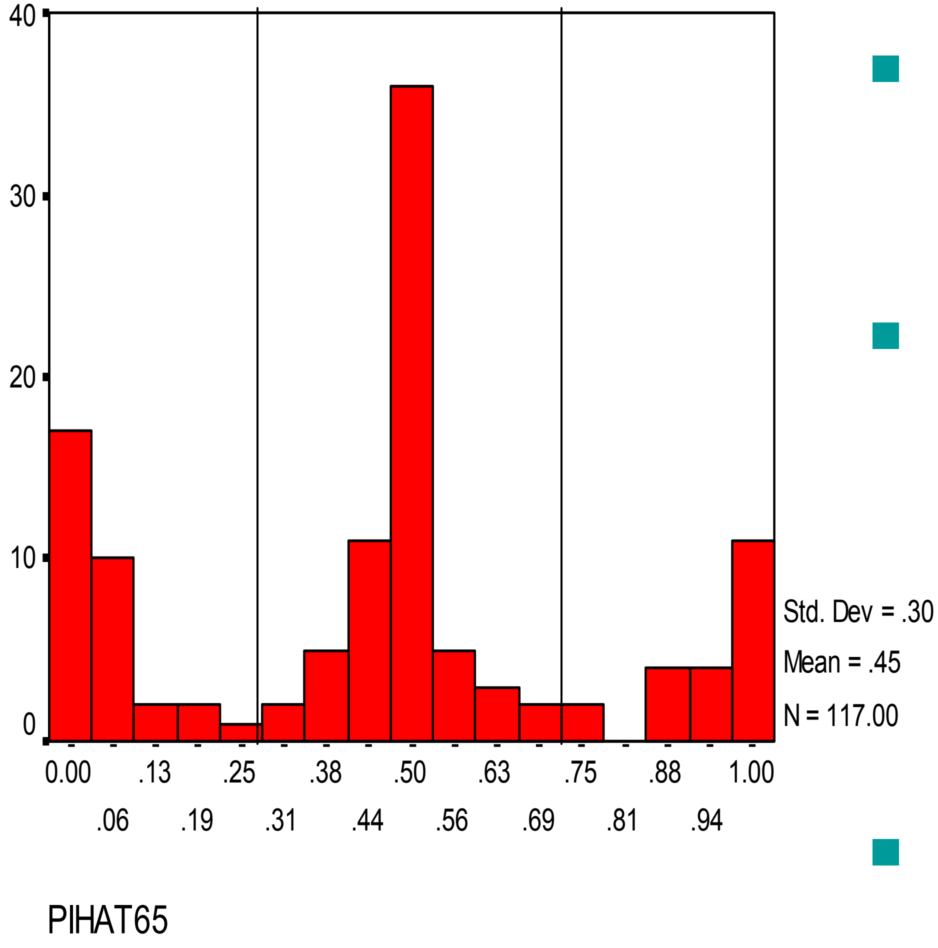
	a^2	c^2	e^2	q^2
LDL	0.04 (0 – 0.39)	0 (0 – 0.27)	0.21 (0.15 – 0.29)	0.75 (0.37 – 0.84)
ApoB	0.46 (0.11 – 0.84)	0.02 (0 – 0.29)	0.19 (0.14 – 0.27)	0.33 (0 – 0.67)
lnApoE	0.03 (0 – 0.33)	0.22 (0 – 0.45)	0.13 (0.10 – 0.18)	0.63 (0.32 – 0.87)

Chi-square Tests for QTL

	DZ+MZ pairs (df=1)	
	Chi-square	Asymptotic p-value
LDL	12.561	0.00020
ApoB	2.128	0.07231
lnApoE	12.292	0.00023

Using the full distribution of pi-hat

Using the full distribution



- More power if we use all the available information
- So instead of dividing the sample we will use $\hat{\pi}$ as a continuous coefficient that will vary between sib-pair across loci
- No MZs in this analysis

Pihat.mx

```
!script for univariate linkage - pihat approach
!DZ/SIB
#loop $i 1 4 1

#define nvar 1
#NGroups 1

DZ / sib TWINS genotyped
 Data NInput=324
 Missing =-1.0000
 Rectangular File=lipidall.dat
  Labels sample fam ldl1 apob1 ldl2  apob2 …

  Select  apob1 apob2
   ibd0m$i
   ibd1m$i
   ibd2m$i
;
  Definition_variables
   ibd0m$i
   ibd1m$i
   ibd2m$i
 ;
```

This use of the loop command allows you to run the same script over and over moving along the chromosome

The format of the command is:
#loop variable start end increment
So…#loop $i 1 4 1
Starts at marker 1 goes to marker 4 and runs each locus in turn
Each occurrence of $i within the script will be replaced by the current number ie on the second run $i will become 2

With the loop command the last end statement becomes an exit statement and the script ends with #end loop

Pihat.mx

```
!script for univariate linkage - pihat approach
!DZ/SIB
#loop $i 1 4 1

#define nvar 1
#NGroups 1

DZ / sib TWINS genotyped
 Data NInput=324
 Missing =-1.0000
 Rectangular File=lipidall.dat
  Labels sample fam ldl1 apob1 ldl2  apob2 …

  Select  apob1 apob2
   ibd0m$i
   ibd1m$i
   ibd2m$i
;
  Definition_variables
   ibd0m$i
   ibd1m$i
   ibd2m$i
 ;
```

This use of the ‘definition variables’ command allows you to specify which of the selected variables will be used as covariates

The value of the covariate displayed in the mxo will be the values for the last case read

Pihat.mx

```
!script for univariate linkage - pihat approach
!DZ/SIB
#loop $i 1 2 1

#define nvar 1
#NGroups 1

DZ / sib TWINS genotyped
 Data NInput=324
 Missing =-1.0000
 Rectangular File=lipidall.dat
  Labels sample fam ldl1 apob1 ldl2  apob2 …

  Select  apob1 apob2
   ibd0m$i
   ibd1m$i
   ibd2m$i
;
  Definition_variables
   ibd0m$i
   ibd1m$i
   ibd2m$i
 ;
```

```
Begin Matrices;
  X Lower nvar nvar free      ! residual familial F
  Z Lower nvar nvar free      ! unshared environment E
  L Full nvar 1 free          ! qtl effect Q
  G Full 1 nvar free          ! grand means
  H Full 1 1                  ! scalar, .5
  K Full 3 1                  ! IBD probabilities (from Merlin)
  J Full 1 3                  ! coefficients 0.5,1 for pihat
 End Matrices;
Specify K
   ibd0m$i
   ibd1m$i
   ibd2m$i

  Matrix H .5
  Matrix J 0 .5 1
  Start .1 X 1 1 1
  Start .1 L 1 1 1
  Start .1 Z 1 1 1
  Start .5 G 1 1 1
```

Pihat.mx

```
 Begin Algebra;
  F= X*X';
 ! residual familial variance
  E= Z*Z';
 ! unique environmental variance
  Q= L*L';
! variance due to QTL
  V= F+Q+E;
! total variance
  T= F|Q|E;
! parameters in one matrix
  S= F%V| Q%V| E%V;
! standardized variance component estimates
  P= ???? ;
! estimate of pihat
 End Algebra;

  Labels Row S standest
  Labels Col S f^2 q^2 e^2
  Labels Row T unstandest
  Labels Col T f^2 q^2 e^2
```

```
 Means
  G| G ;
 Covariance
   F+E+Q | F+P@Q_
   F+P@Q | F+E+Q ;

 Option NDecimals=4
 Option RSiduals
 Option Multiple Issat
!End

!test significance of QTL effect
! Drop L 1 1 1
Exit

#end loop
```

What should this be?

Pihat.mx

```
 Begin Algebra;
  F= X*X';
 ! residual familial variance
  E= Z*Z';
 ! unique environmental variance
  Q= L*L';
! variance due to QTL
  V= F+Q+E;
! total variance
  T= F|Q|E;
! parameters in one matrix
  S= F%V| Q%V| E%V;
! standardized variance component estimates
  P= ???? ;
! estimate of pihat
 End Algebra;

  Labels Row S standest
  Labels Col S f^2 q^2 e^2
  Labels Row T unstandest
  Labels Col T f^2 q^2 e^2
```

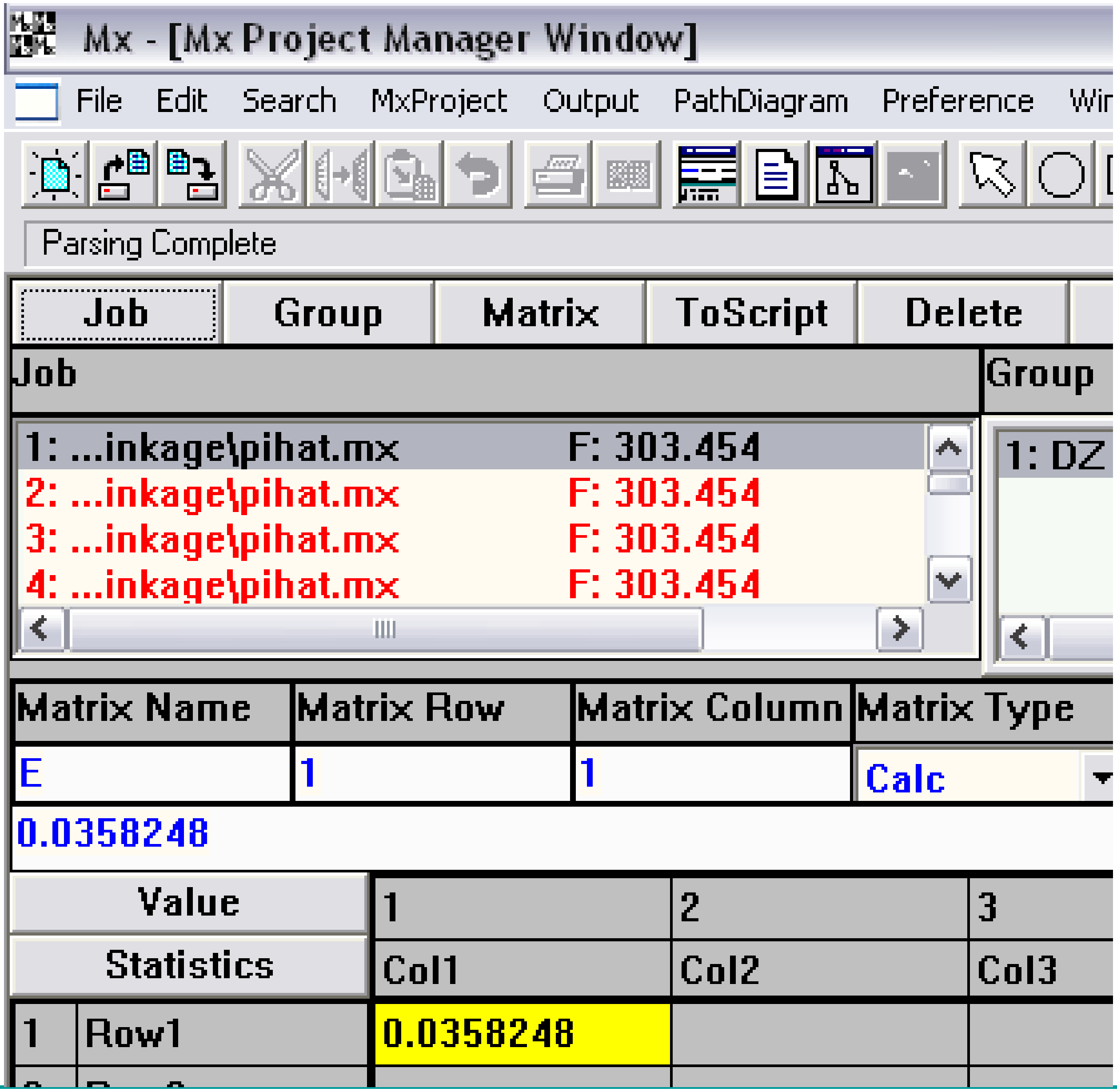
```
 Means
  G| G ;
 Covariance
   F+E+Q | F+P@Q_
   F+P@Q | F+E+Q ;

 Option NDecimals=4
 Option RSiduals
 Option Multiple Issat
!End

!test significance of QTL effect
! Drop L 1 1 1
Exit

#end loop
```

J*K

Your task…

- Change the loci being analysed
  - Left side of the room (your left) – 70 to 75
  - Right side of the room (your right) – 76 -80

Mx



Parsing Complete

Job	
1: ...inkage\pihat.mx	F: 303.454
2: ...inkage\pihat.mx	F: 303.454
3: ...inkage\pihat.mx	F: 303.454
4: ...inkage\pihat.mx	F: 303.454

Group
1: DZ

Matrix Name	Matrix Row	Matrix Column	Matrix Type
E	1	1	Calc

0.0358248

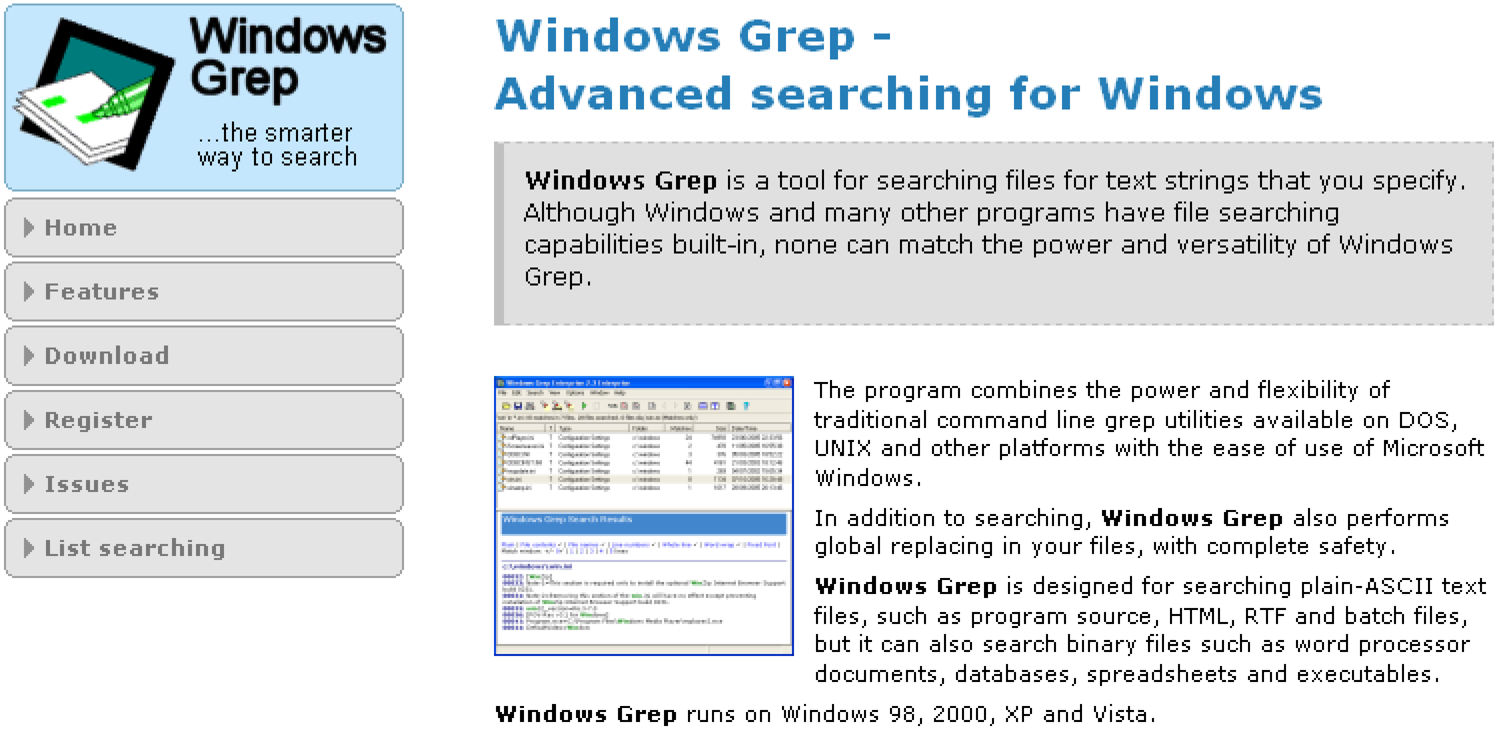
		1 Col1	2 Col2	3 Col3
1	Row1	0.0358248		

Grepping the results

- Under Unix/Linux/Cygwin

  -
    ```
    grep ‘of data’ pihat.mxo > output.txt
    ```

Grepping the results



Windows Grep - Advanced searching for Windows

Windows Grep is a tool for searching files for text strings that you specify. Although Windows and many other programs have file searching capabilities built-in, none can match the power and versatility of Windows Grep.

The program combines the power and flexibility of traditional command line grep utilities available on DOS, UNIX and other platforms with the ease of use of Microsoft Windows.

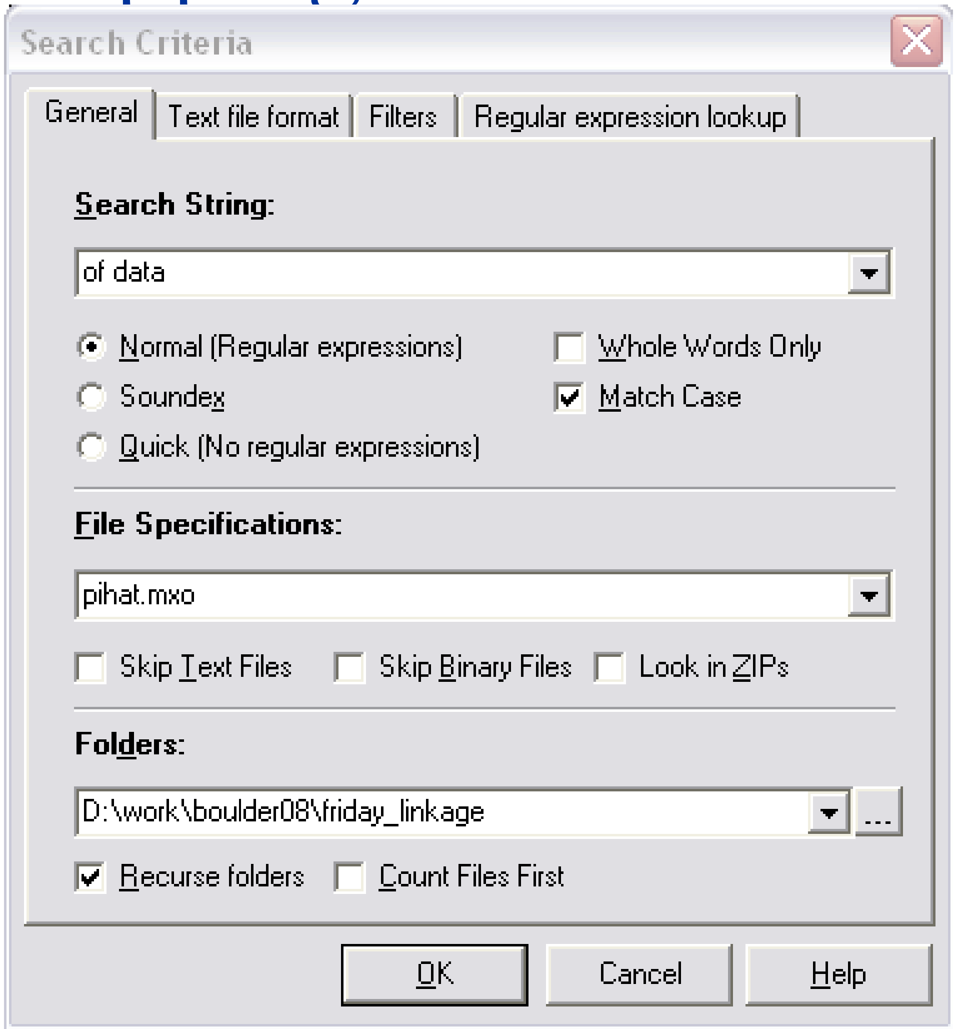
In addition to searching, **Windows Grep** also performs global replacing in your files, with complete safety.

Windows Grep is designed for searching plain-ASCII text files, such as program source, HTML, RTF and batch files, but it can also search binary files such as word processor documents, databases, spreadsheets and executables.

Windows Grep runs on Windows 98, 2000, XP and Vista.

http://www.wingrep.com/index.htm

Grepping the results



Search String:

of data

● Normal (Regular expressions)
○ Soundex
○ Quick (No regular expressions)

□ Whole Words Only
☑ Match Case

File Specifications:

pihat.mxo

□ Skip Text Files □ Skip Binary Files □ Look in ZIPs

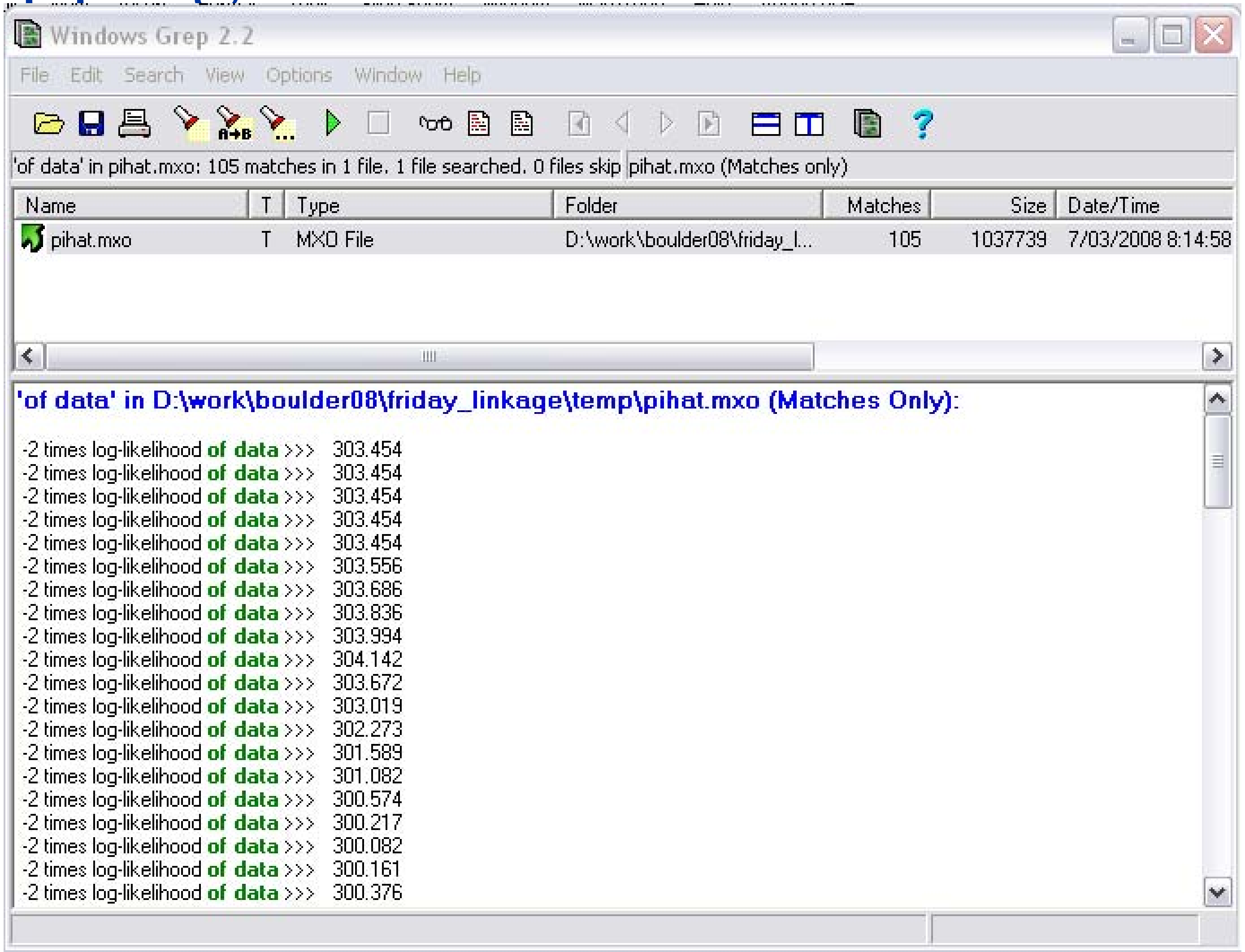
Folders:

D:\work\boulder08\friday_linkage

☑ Recurse folders □ Count Files First

http://www.wingrep.com/index.htm

Grepping the results



'of data' in pihat.mxo: 105 matches in 1 file. 1 file searched. 0 files skip | pihat.mxo (Matches only)

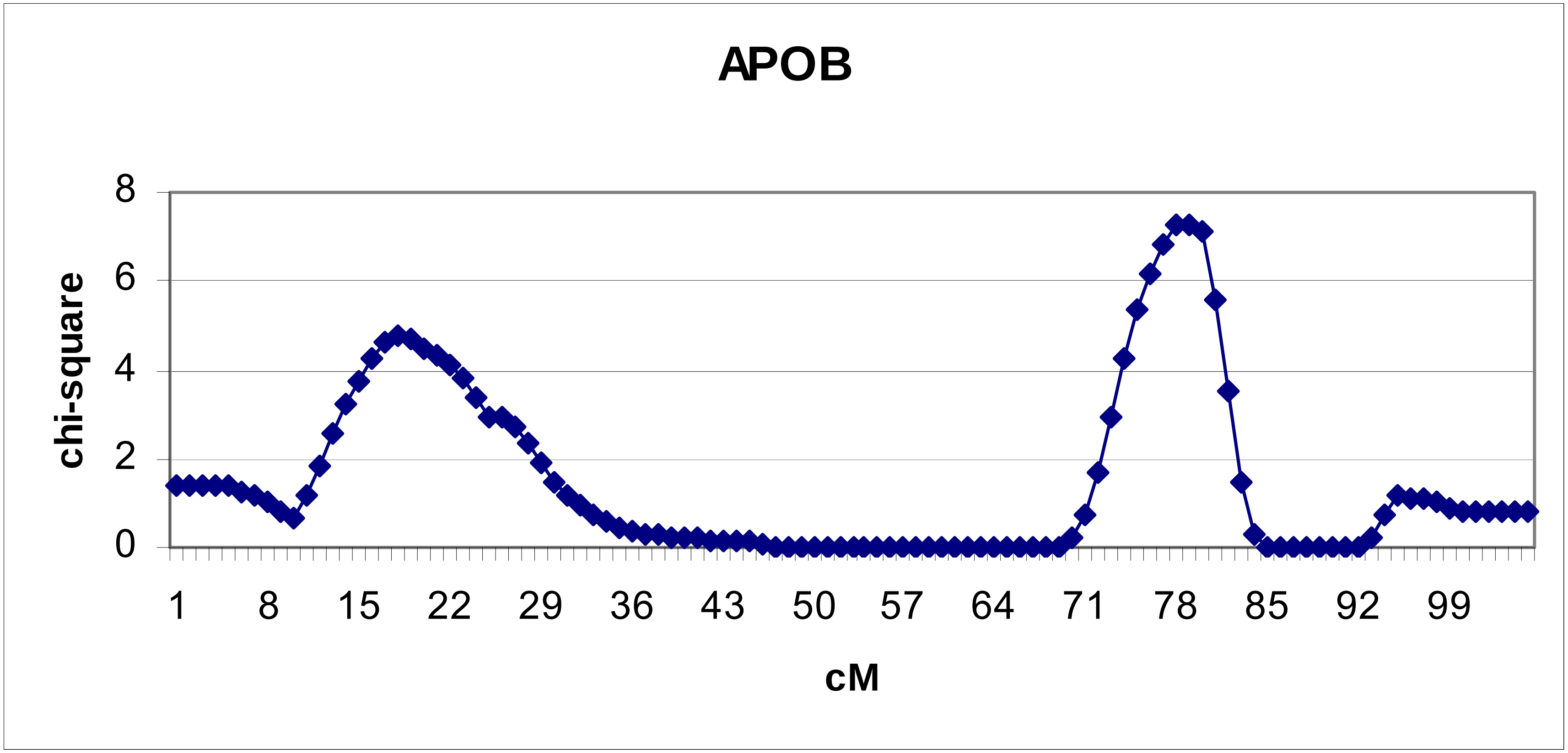
Name	T	Type	Folder	Matches	Size	Date/Time
pihat.mxo	T	MXO File	D:\work\boulder08\friday_l...	105	1037739	7/03/2008 8:14:58

'of data' in D:\work\boulder08\friday_linkage\temp\pihat.mxo (Matches Only):

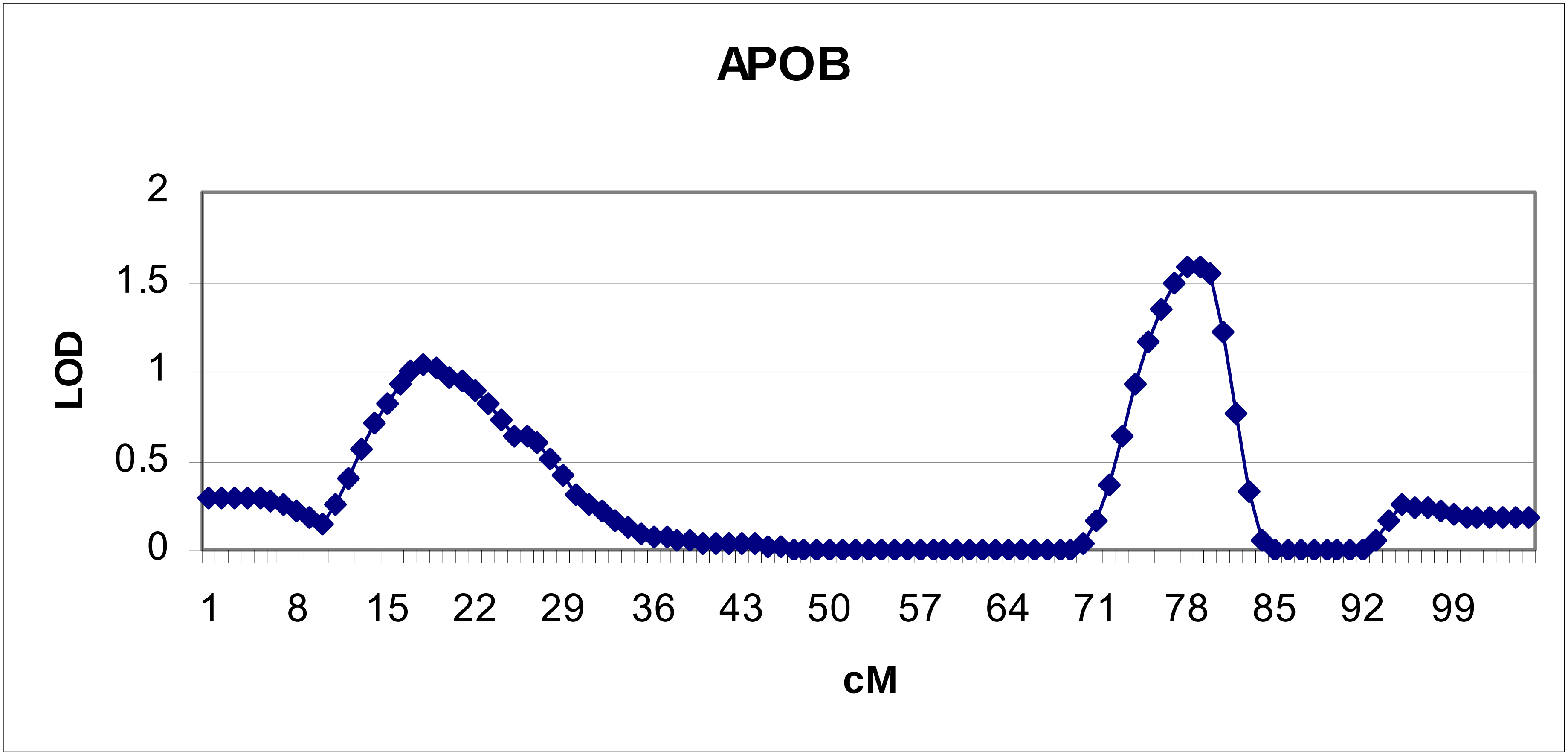
```
-2 times log-likelihood of data >>>   303.454
-2 times log-likelihood of data >>>   303.454
-2 times log-likelihood of data >>>   303.454
-2 times log-likelihood of data >>>   303.454
-2 times log-likelihood of data >>>   303.454
-2 times log-likelihood of data >>>   303.556
-2 times log-likelihood of data >>>   303.686
-2 times log-likelihood of data >>>   303.836
-2 times log-likelihood of data >>>   303.994
-2 times log-likelihood of data >>>   304.142
-2 times log-likelihood of data >>>   303.672
-2 times log-likelihood of data >>>   303.019
-2 times log-likelihood of data >>>   302.273
-2 times log-likelihood of data >>>   301.589
-2 times log-likelihood of data >>>   301.082
-2 times log-likelihood of data >>>   300.574
-2 times log-likelihood of data >>>   300.217
-2 times log-likelihood of data >>>   300.082
-2 times log-likelihood of data >>>   300.161
-2 times log-likelihood of data >>>   300.376
```

http://www.wingrep.com/index.htm

Difference in Chi-square



LOD score


APOB
LOD
cM