

Family Based Association

Danielle Posthuma

Stacey Cherny

TC18-Boulder 2005

Overview

- Simple association test
- Practical population stratification
- Family based association
- Practical family based association and linkage in Mx

Life after Linkage

- Fine mapping
- Searching for putative candidate genes
- Searching for the functional polymorphism
- Testing for **association**

Simple Association Model

- Model association in the means model
- Each copy of an allele changes trait by a fixed amount
 - Use covariate counting copies for allele of interest

$$E(y_i) = \mu + a * [\text{number of copies of allele}] \quad \text{Or;}$$

$$E(y_i) = \mu + \beta_x X_i$$

X is the number of copies of the allele of interest.

β_x is the estimated effect of each copy (the additive genetic value)

Results in estimate of additive genetic value. Evidence for association when $\beta_x \neq 0$

Simple association model is sensitive to population stratification

Occurs when

- differences in allele frequencies, AND
- differences in prevalence or means of a trait

Case-control study

- Often used
- High statistical power

BUT:

- Spurious association (false positives/negatives): population stratification



NEWS & VIEWS

Beware the chopsticks gene

D Hamer and L Sirota
Laboratory of Biochemistry
National Cancer Institute
National Institutes of Health
Bethesda, MD, USA

Once upon a time, an ethnogeneticist decided to figure out why some people eat with chopsticks and others do not. His experiment was simple. He rounded up several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and mapped them for a series of anonymous and candidate genes.

The results were astounding. One of the markers, located right in the middle of a region previously linked to several behavioral traits, showed a huge correlation to chopstick use, enough to account for nearly half of the observed variance. When the experiment was repeated with students from a different university, precisely the same marker lit up. Eureka! The delighted scientist popped a bottle of champagne and quickly submitted an article to *Molecular Psychiatry* heralding the discovery of the ‘successful-use-of-selected-handinstruments gene’ (SUSHI).

Where did the delighted scientist go wrong?

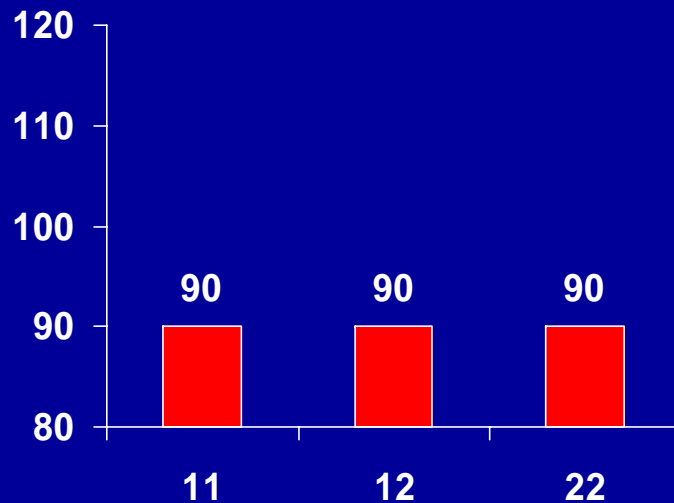
- All the 'cases' were from Asian descent, while the 'controls' were from European descent
- Due to historical differences allele frequencies for many genes differ between the Asians and Europeans
- Due to cultural differences many Asians eat with chopsticks while Europeans generally will not

Thus, every allele with a different frequency is now falsely identified as being associated with eating with chopsticks ...

Practical – Find a gene for sensation seeking:

- Two populations (A & B) of 100 individuals in which sensation seeking was measured
- In population A, gene X (alleles 1 & 2) does not influence sensation seeking
- In population B, gene X (alleles 1 & 2) does not influence sensation seeking
- Mean sensation seeking score of population A is 90
- Mean sensation seeking score of population B is 110
- Frequencies of allele 1 & 2 in population A are .1 & .9
- Frequencies of allele 1 & 2 in population B are .5 & .5

Population A



Genotypic freq.

11

.01

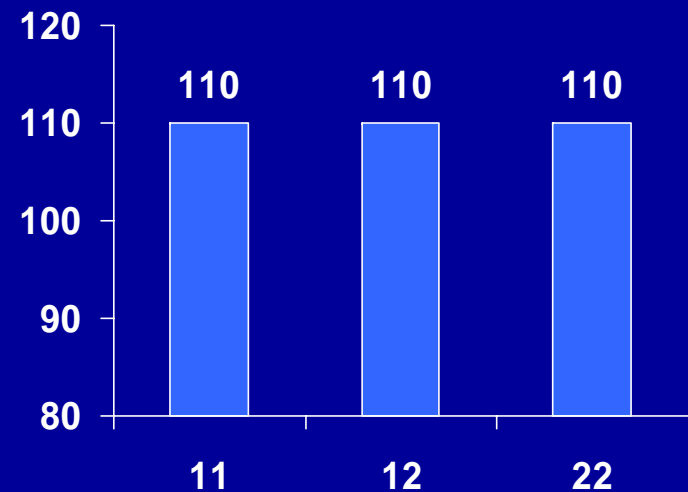
12

.18

22

.81

Population B



11

.25

12

.50

22

.25

Sensation seeking score is the same across genotypes, within each population.

Population B scores higher than population A

Differences in genotypic frequencies

Suppose we are unaware of these two populations and have measured 200 individuals and typed gene X

The mean sensation seeking score of this mixed population is 100

What are our observed genotypic frequencies and means?

Calculating genotypic frequencies in the mixed population

Genotype 11:

1 individual from population A, 25 individuals from population B on a total of 200 individuals: $(1+25)/200=.13$

Genotype 12: $(18+50)/200=.34$

Genotype 22: $(81+25)/200=.53$

Calculating genotypic means in the mixed population

Genotype 11:

1 individual from population A with a mean of 90, 25 individuals from population B with a mean of 110 = $((1*90) + (25*110))/26 = 109.2$

Genotype 12: $((18*90) + (50*110))/68 = 104.7$

Genotype 22: $((81*90) + (25*110))/106 = 94.7$

Gene X is the gene for sensation seeking!

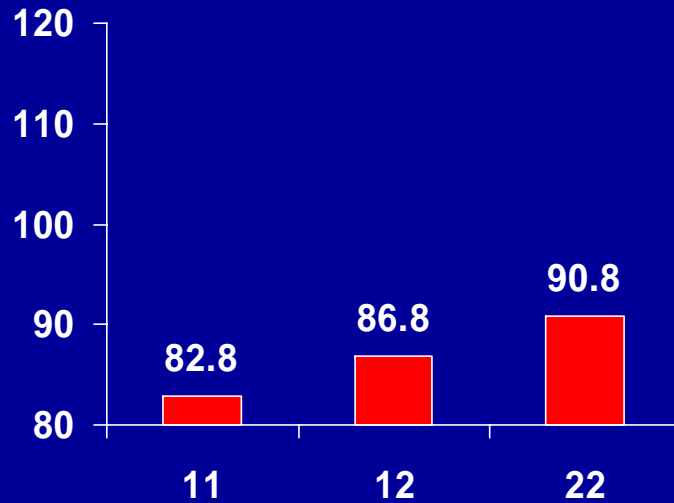


Now, allele 1 is associated with higher sensation seeking scores, while in both populations A and B, the gene was not associated with sensation seeking scores...

FALSE ASSOCIATION

What if there is true association?

Population A



Genotypic freq. **.01** **.18** **.81**

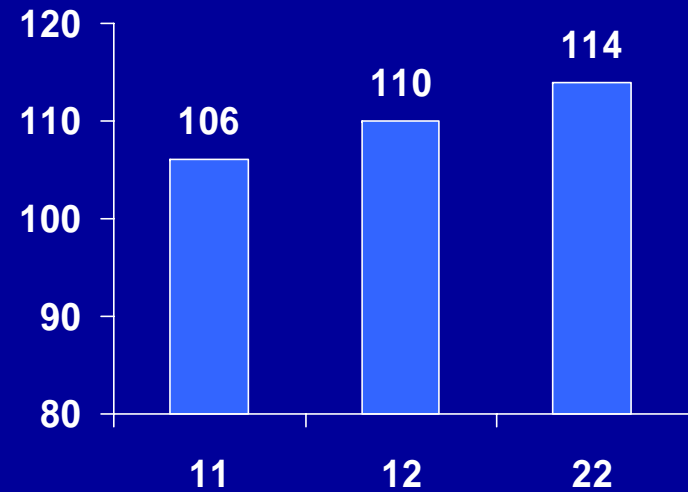
allele 1 frequency 0.1

allele 2 frequency 0.9

allele 1 = -2,
allele 2 = +2

Pop mean = 90

Population B



.25 **.50** **.25**

allele 1 frequency 0.5

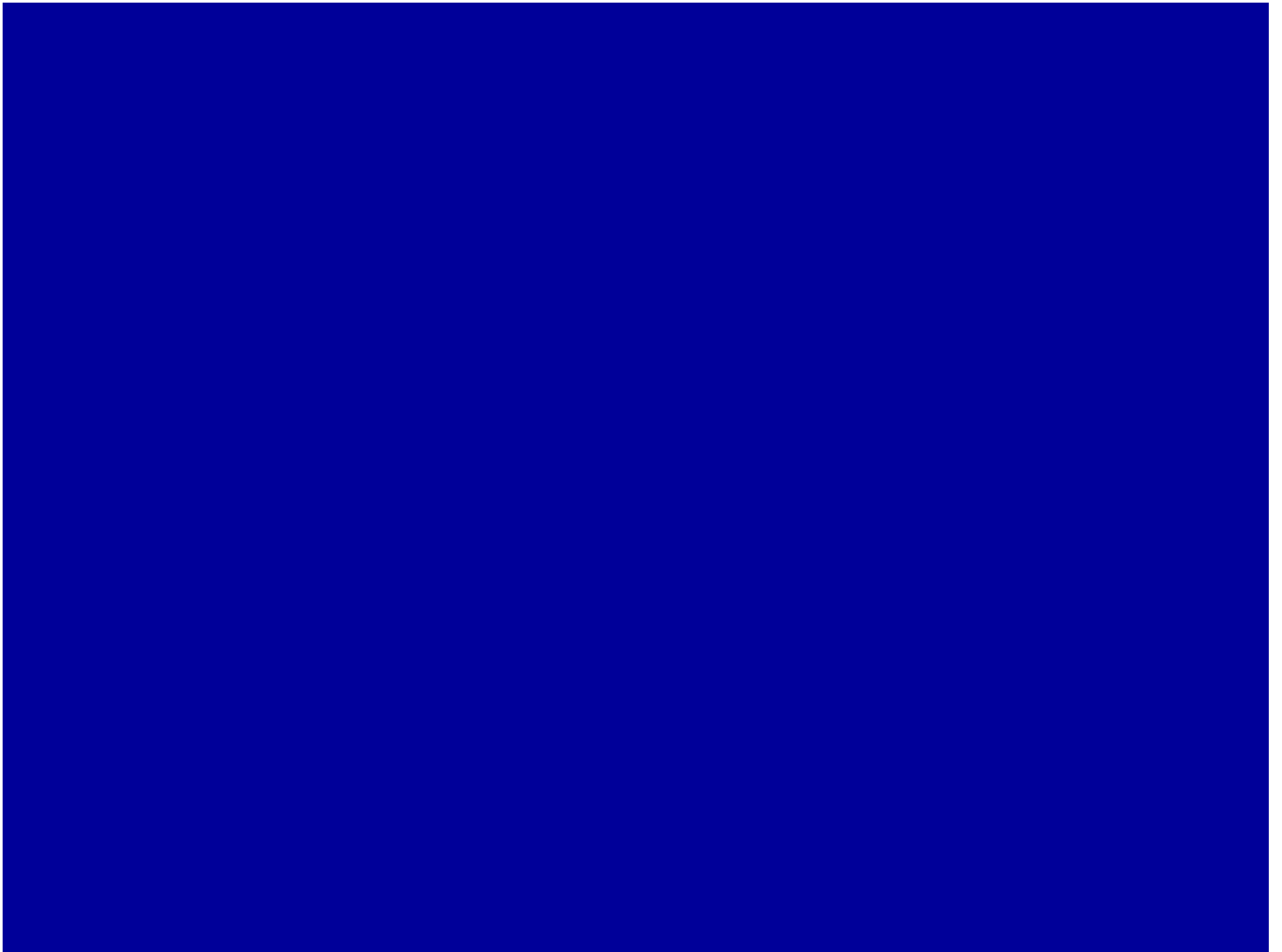
allele 2 frequency 0.5.

allele 1 = -2
allele 2 = +2

Pop mean = 110

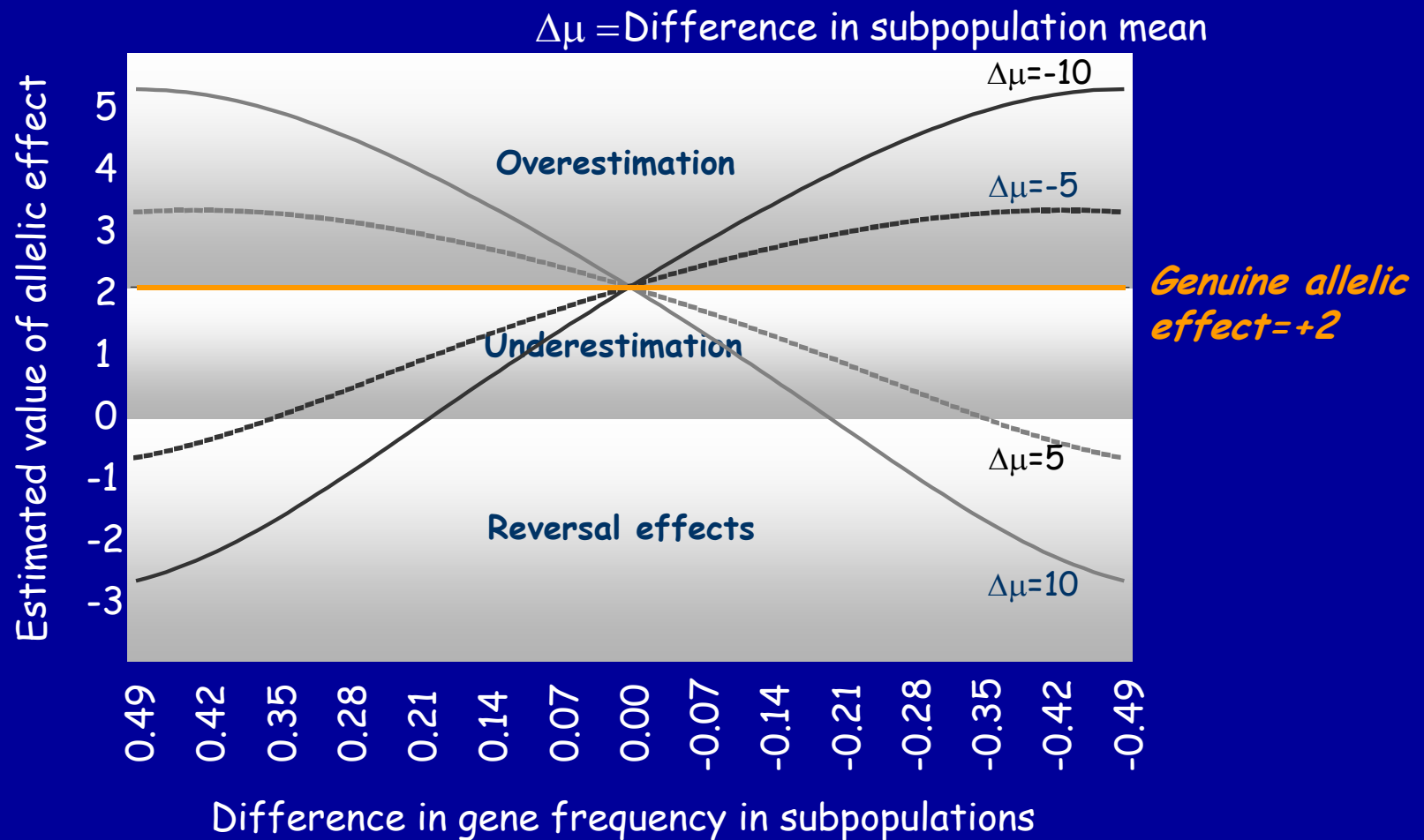
Calculate:

- Genotypic means in mixed population
- Genotypic frequencies in mixed population
- Is there an association between the gene and sensation seeking score? If yes which allele is the increaser allele?



- There is an excell sheet with which you can play around, and which calculates the extent of false association for you:
- Association.xls

False positives and false negatives



How to avoid spurious association?

True association is detected in people coming from the same genetic stratum

Controlling for Stratification

- Stratification produces differences between families NOT within families
- Partition g_{ij} (no. of copies of allele - 1) into a between families component (b_{ij}) and a within families component (w_{ij}) (Fulker et al., 1999)

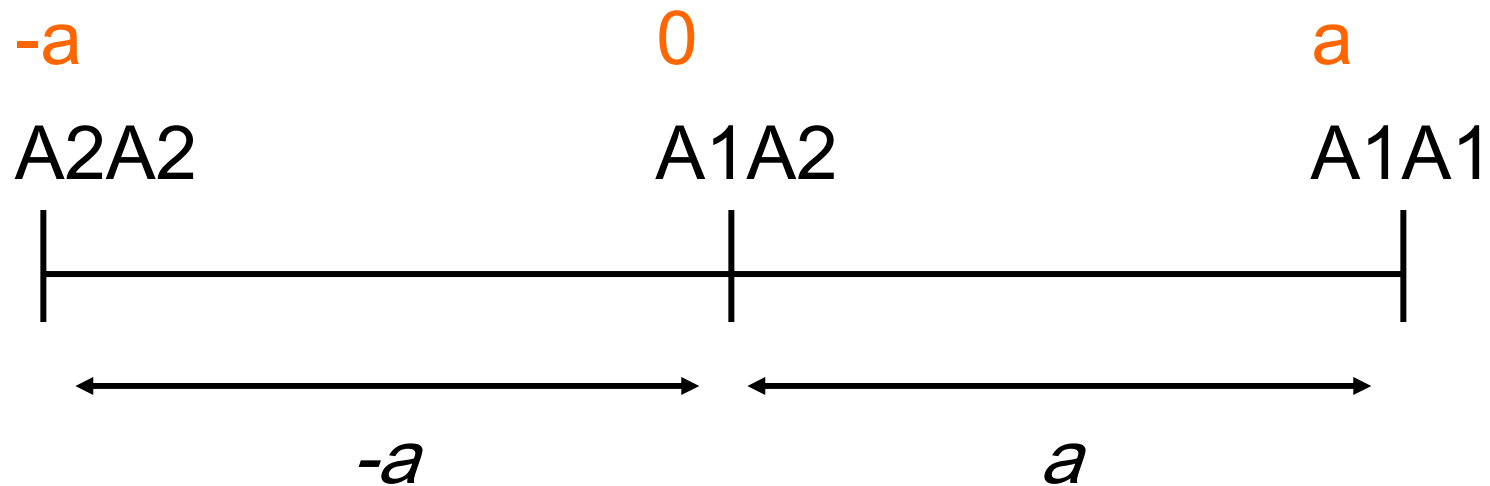
$$\hat{y}_{ij} = \mu + \hat{\beta}_b b_{ij} + \hat{\beta}_w w_{ij}$$

b_{ij} as Family Control

- b_{ij} is the expected genotype for each individual
 - Ancestors
 - Siblings
- w_{ij} is the deviation of each individual from this expectation
- Informative individuals
 - To be “informative” an individual’s genotype should differ from expected
 - Have heterozygous ancestor in pedigree
- $\beta_b \neq \beta_w$ is a test for population stratification
- $\beta_w > 0$ is a test for association free from stratification

Partitioning of Additive Effect into Between- and Within-Pairs Components

GENOTYPE		ADDITIVE EFFECT			
Sib 1	Sib 2	Sib 1	Sib 2	MEAN	DIFFERENCE/2
A_1A_1	A_1A_1	a_b	a_b	a_b	0
A_1A_1	A_1A_2	$(a_b/2) + (a_w/2)$	$(a_b/2) - (a_w/2)$	$a_b/2$	$a_w/2$



Fulker (1999) model extended to include dominance effects, conditional on parental genotypes, multiple alleles, multiple sibs

Table II. Partitioning of Additive and Dominance Genotypic Effects into Between and Within Components for a Diallelic Locus with Alleles E and e in Sib-pairs

Genotype		Genotypic effect		Additive		Dominance		Partitioned genotypic effects	
<i>Sib 1</i>	<i>Sib 2</i>	<i>Sib 1</i>	<i>Sib 2</i>	<i>Mean</i>	<i>Difference/2</i>	<i>Mean</i>	<i>Difference/2</i>	<i>Sib 1</i>	<i>Sib 2</i>
EE	EE	a	a	a_b	0	0	0	a_b	a_b
EE	Ee	a	d	$\frac{1}{2}a_b$	$\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$-\frac{1}{2}d_w$	$(\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$	$(\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$
EE	ee	a	-a	0	a_w	0	0	a_w	$-a_w$
Ee	EE	d	a	$\frac{1}{2}a_b$	$-\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$\frac{1}{2}d_w$	$(\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$	$(\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$
Ee	Ee	d	d	0	0	d_b	0	d_b	d_b
Ee	ee	d	-a	$-\frac{1}{2}a_b$	$\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$\frac{1}{2}d_w$	$(-\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$	$(-\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$
ee	EE	-a	a	0	$-a_w$	0	0	$-a_w$	a_w
ee	Ee	-a	d	$-\frac{1}{2}a_b$	$-\frac{1}{2}a_w$	$\frac{1}{2}d_b$	$-\frac{1}{2}d_w$	$(-\frac{1}{2}a_b - \frac{1}{2}a_w) + (\frac{1}{2}d_b - \frac{1}{2}d_w)$	$(-\frac{1}{2}a_b + \frac{1}{2}a_w) + (\frac{1}{2}d_b + \frac{1}{2}d_w)$
ee	ee	-a	-a	$-a_b$	0	0	0	$-a_b$	$-a_b$

Table III. Partitioning of Additive and Dominance Genotypic Effects into Between and Within Components for a Diallelic Locus with Alleles E and e Conditional on Parental Genotypes

Parental mating type ↓	Offspring						Family mean (<i>Between</i>)
	Probability	EE <i>a</i> Deviation from family mean (<i>Within</i>)	Probability	Ee <i>d</i> Deviation from family mean (<i>Within</i>)	Probability	ee <i>-a</i> Deviation from family mean (<i>Within</i>)	
EE × EE	1	0	0	n.p.	0	n.p.	a_b
EE × Ee	$\frac{1}{2}$	$\frac{1}{2}a_w - \frac{1}{2}d_w$	$\frac{1}{2}$	$-\frac{1}{2}a_w + \frac{1}{2}d_w$	0	n.p.	$\frac{1}{2}a_b + \frac{1}{2}d_b$
EE × ee	0	n.p.	1	0	0	n.p.	d_b
Ee × Ee	$\frac{1}{4}$	$a_w - \frac{1}{2}d_w$	$\frac{1}{2}$	$\frac{1}{2}d_w$	$\frac{1}{4}$	$-a_w - \frac{1}{2}d_w$	$\frac{1}{2}d_b$
Ee × ee	0	n.p.	$\frac{1}{2}$	$\frac{1}{2}d_w + \frac{1}{2}a_w$	$\frac{1}{2}$	$-\frac{1}{2}d_w - \frac{1}{2}a_w$	$\frac{1}{2}d_b - \frac{1}{2}a_b$
ee × ee	0	n.p.	0	n.p.	1	0	$-a_b$

Note: The expectation for an individual sibling is the sum of the *Between* and *Within* components.
n.p., not possible.

Nuclear Families

$$b_{ij} = b_i = \begin{cases} \frac{g_{iF} + g_{iM}}{2} & \text{average of parental genotypes} \\ \sum_k^{sibship} \frac{g_{ik}}{n_{sibs}} & \text{average of sibling genotypes} \end{cases}$$

$$w_{ij} = g_{ij} - b_{ij}$$

Combined Linkage & association

Implemented in **QTDT** (*Abecasis et al., 2000*) and **Mx** (*Posthuma et al., 2004*)

Association and Linkage modeled simultaneously:

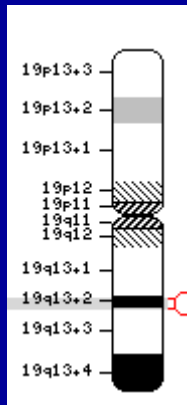
- Association is modeled in the means
- Linkage is modeled in the (co)variances

Testing for linkage in the presence of association provides information on whether or not the polymorphisms used in the association model explain the observed linkage or whether other polymorphisms in that region are expected to be of influence

QTDT: simple, quick, straightforward, but not so flexible in terms of models

Mx: can be considered less simple, but highly flexible

Example: The ApoE-gene



- Three alleles have been identified: e2, e3, and e4
- e3-allele is most common
- e2 and e4 are rarer and associated with pathological conditions

The apoE-gene is localized on chromosome 19 (q12-13.2)

Six combinations of the apoE alleles are possible

The 3 alleles (e2, e3, and e4) code for different proteins (isoforms), but may also relate to differences in transcription

ApoE isoforms

Apo E2



Apo E3 (wild type)



Apo E4



-  : Signal Peptide
-  : LDL Receptor Binding Site (140-160)
-  : Lipid Binding Site (202-299)

APOE $\epsilon 2/\epsilon 3/\epsilon 4$ gene and apoE plasma levels

- 148 Adolescent twin pairs
- 202 Adult twin pairs

TABLE I. Descriptive statistics of the adolescent and adult Dutch monozygotic and dizygotic twin-pair samples of which both members provide data on apoE level and APOE genotype

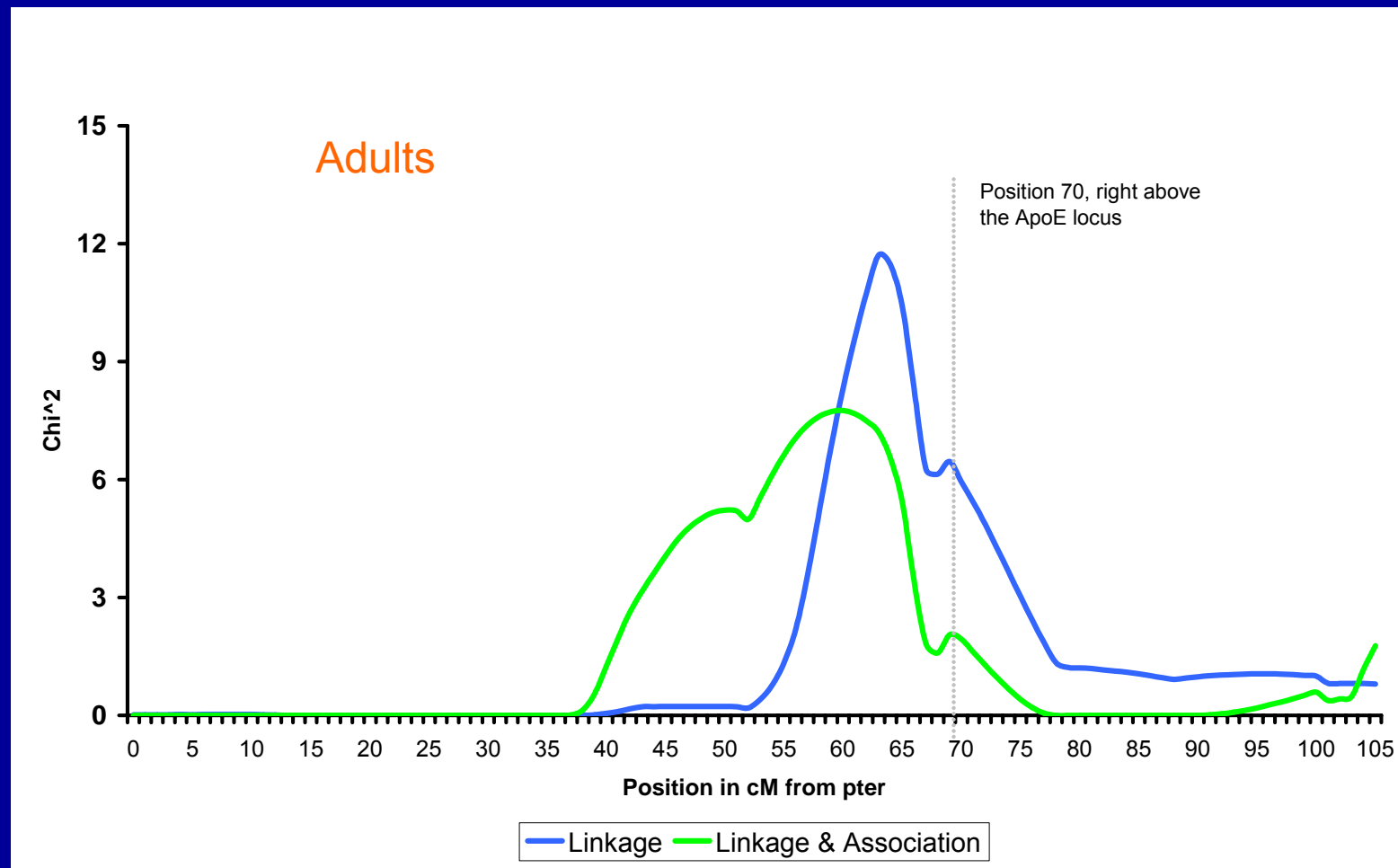
	Adolescent twins		Adult twins	
	MZ	DZ	MZ	DZ
Number of pairs	65	83	88	114
Percentage males	46.2	49.4	47.7	48.2
Age, years—mean (range)	16 (13–22)	17 (13–22)	44 (34–62)	44 (34–59)
ApoE, mg/dL—mean (SD)	7.12 (2.42)	6.59 (2.39)	2.53 (1.03)	2.54 (1.02)
Ln(apoE)—mean (SD)	1.91 (0.34)	1.82 (0.36)	0.86 (0.39)	0.86 (0.37)
Correlation	0.88	0.37	0.87	0.41

Genetic Epidemiology 26: 328–337 (2004)

Combined Association and Linkage Analysis Applied to the APOE Locus

Marian Beekman,^{1*} Daniëlle Posthuma,² Bastiaan T. Heijmans,¹ Nico Lakenberg,¹ H. Eka D. Suchiman,¹ Harold Snieder,^{3,4} Peter de Knijff,⁵ Rune R. Frants,⁵ Gert Jan B. van Ommen,⁵ Cornelis Klufft,⁶ George P. Vogler,⁷ P. Eline Slagboom,¹ and Dorret I. Boomsma²

Linkage on chrom. 19 and association with APOE $\epsilon 2/\epsilon 3/\epsilon 4$ for apoE plasma levels



Implementation in Mx

```
#define n 3                                ! number of alleles is 3, coded 1, 2, 3
G1: calculation group between and within effects
Data Calc
Begin matrices;
  A Full 1 n free                          ! additive allelic effects within
  C Full 1 n free                          ! additive allelic effects between
  D Sdiag n n free                        ! dominance deviations within
  F Sdiag n n free                        ! dominance deviations between
  I Unit 1 n                              ! one's
End matrices;

Specify A 100 101 102
Specify C 200 201 202
Specify D 800 801 802
Specify F 900 901 902
```

$$K = (A'@I) + (A@I')$$

$$L = D + D'$$

$$W = K+L$$

! Within effects, additive

! Within effects, dominance

! Within effects total

$$I = [1 \ 1 \ 1], \quad A = [a1 \ a2 \ a3]$$

$$D = \begin{bmatrix} 0 & 0 & 0 \\ d21 & 0 & 0 \\ d31 & d32 & 0 \end{bmatrix}$$

$$K = (A'@I) + (A@I') =$$

$$\begin{bmatrix} a1 \\ a2 \\ a3 \end{bmatrix} @ [1 \ 1 \ 1] + [a1 \ a2 \ a3] @ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} =$$

$$\begin{bmatrix} a1 & a1 & a1 \\ a2 & a2 & a2 \\ a3 & a3 & a3 \end{bmatrix} + \begin{bmatrix} a1 & a2 & a3 \\ a1 & a2 & a3 \\ a1 & a2 & a3 \end{bmatrix} = \begin{bmatrix} a1a1 & a1a2 & a1a3 \\ a2a1 & a2a2 & a2a3 \\ a3a1 & a3a2 & a3a3 \end{bmatrix}$$

$$W = K+L =$$

$$\begin{bmatrix} a1a1 & a1a2 & a1a3 \\ a2a1 & a2a2 & a2a3 \\ a3a1 & a3a2 & a3a3 \end{bmatrix} + \begin{bmatrix} 0 & d21 & d31 \\ d21 & 0 & d32 \\ d31 & d32 & 0 \end{bmatrix} = \begin{bmatrix} a1a1 & a1a2d21 & a1a3d31 \\ a2a1d21 & a2a2 & a2a3d32 \\ a3a1d31 & a3a2d32 & a3a3 \end{bmatrix}$$

$$L = D + D' =$$

$$\begin{bmatrix} 0 & 0 & 0 \\ d21 & 0 & 0 \\ d31 & d32 & 0 \end{bmatrix} + \begin{bmatrix} 0 & d21 & d31 \\ 0 & 0 & d32 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & d21 & d31 \\ d21 & 0 & d32 \\ d31 & d32 & 0 \end{bmatrix}$$

$$M = (C'@I) + (C@I')$$

$$N = F + F'$$

$$B = M+N$$

! Between effects, additive

! Between effects, dominance

! Between effects - total

$$W = \begin{bmatrix} a1a1 & a1a2d21 & a1a3d31 \\ a2a1d21 & a2a2 & a2a3d32 \\ a3a1d31 & a3a2d32 & a3a3 \end{bmatrix}$$

$$B = \begin{bmatrix} c1c1 & c1c2f21 & c1c3f31 \\ c2c1f21 & c2c2 & c2c3f32 \\ c3c1f31 & c3c2f32 & c3c3 \end{bmatrix}$$

- We have a sibpair with genotypes 1,1 and 1,2.
- To calculate the between-pairs effect, or the mean genotypic effect of this pair, we need matrix B: $((c1c1) + (c1c2f21)) / 2$
- To calculate the within-pair effect we need matrix W and the between pairs effect:

For sib1: $(a1a1) + ((c1c1) + (c1c2f21)) / 2$

For sib2: $(a1a2d21) - ((c1c1) + (c1c2f21)) / 2$

Specify K apoe_11 apoe_21 apoe_11 apoe_21
! allele1twin1 allele2twin1 allele1twin1 allele2twin1 , used for \part

Specify L apoe_12 apoe_22 apoe_12 apoe_22
! allele1twin2 allele2twin2 allele1twin2 allele2twin2 , used for \part

$V = (\text{\part(B,K)} + \text{\part(B,L)}) \%S ;$
! Calculates sib genotypic mean (= Between effects)

$C = (\text{\part(W,K)} + \text{\part(W,L)}) \%S ;$
! Calculates sib genotypic mean, used to derive deviation from this
mean below (Within effects)

Means $G + F * R' + V + (\text{\part(W,K)} - C) | G + I * R' + V + (\text{\part(W,L)} - C);$

$$W = \begin{bmatrix} a1a1 & a1a2d21 & a1a3d31 \\ a2a1d21 & a2a2 & a2a3d32 \\ a3a1d31 & a3a2d32 & a3a3 \end{bmatrix}$$

$$B = \begin{bmatrix} c1c1 & c1c2f21 & c1c3f31 \\ c2c1f21 & c2c2 & c2c3f32 \\ c3c1f31 & c3c2f32 & c3c3 \end{bmatrix}$$

Sibpair with genotypes: 1,1 and 1,2

Specify K apoe_11 apoe_21 apoe_11 apoe_21 = 1 1 1 1

Specify L apoe_12 apoe_22 apoe_12 apoe_22 = 1 2 1 2

$$V = (\text{\part}(B,K) + \text{\part}(B,L)) \%S ; \quad (c1c1 + c1c2f21)/2$$

$$C = (\text{\part}(W,K) + \text{\part}(W,L)) \%S ; \quad (a1a1 + a1a2d21)/2$$

Means $G + F^*R' + V + (\text{\part}(W,K)-C) \mid G + I^*R' + V + (\text{\part}(W,L)-C); =$

$$G + F^*R' + (c1c1 + c1c1f21)/2 + (a1a1 - (a1a1 + a1a2d21)/2) \mid$$

$$G + I^*R' + (c1c1 + c1c1f21)/2 + (a2a1 - (a1a1 + a1a2d21)/2)$$

Constrain sum additive allelic within effects = 0

Constraint ni=1

Begin Matrices;

 A full 1 n = A1

 O zero 1 1

End Matrices;

Begin algebra;

 B = \sum(A) ;

End Algebra;

Constraint O = B ;

end

Constrain sum additive allelic between effects = 0

Constraint ni=1

Begin Matrices;

 C full 1 n = C1 !

 O zero 1 1

End Matrices;

Begin algebra;

 B = \sum(C) ;

End Algebra;

Constraint O = B ;

end

!1.test for linkage in presence of full association

Drop D 2 1 1

end

!2.Test for population stratification:

!between effects = within effects.

Specify 1 A 100 101 102

Specify 1 C 100 101 202

Specify 1 D 800 801 802

Specify 1 F 800 801 802

end

!3.Test for presence of dominance

Drop @0 800 801 802

end

!4.Test for presence of full association

Drop @0 800 801 802 100 101

end

!5.Test for linkage in absence of association

Free D 2 1 1

end

Practical

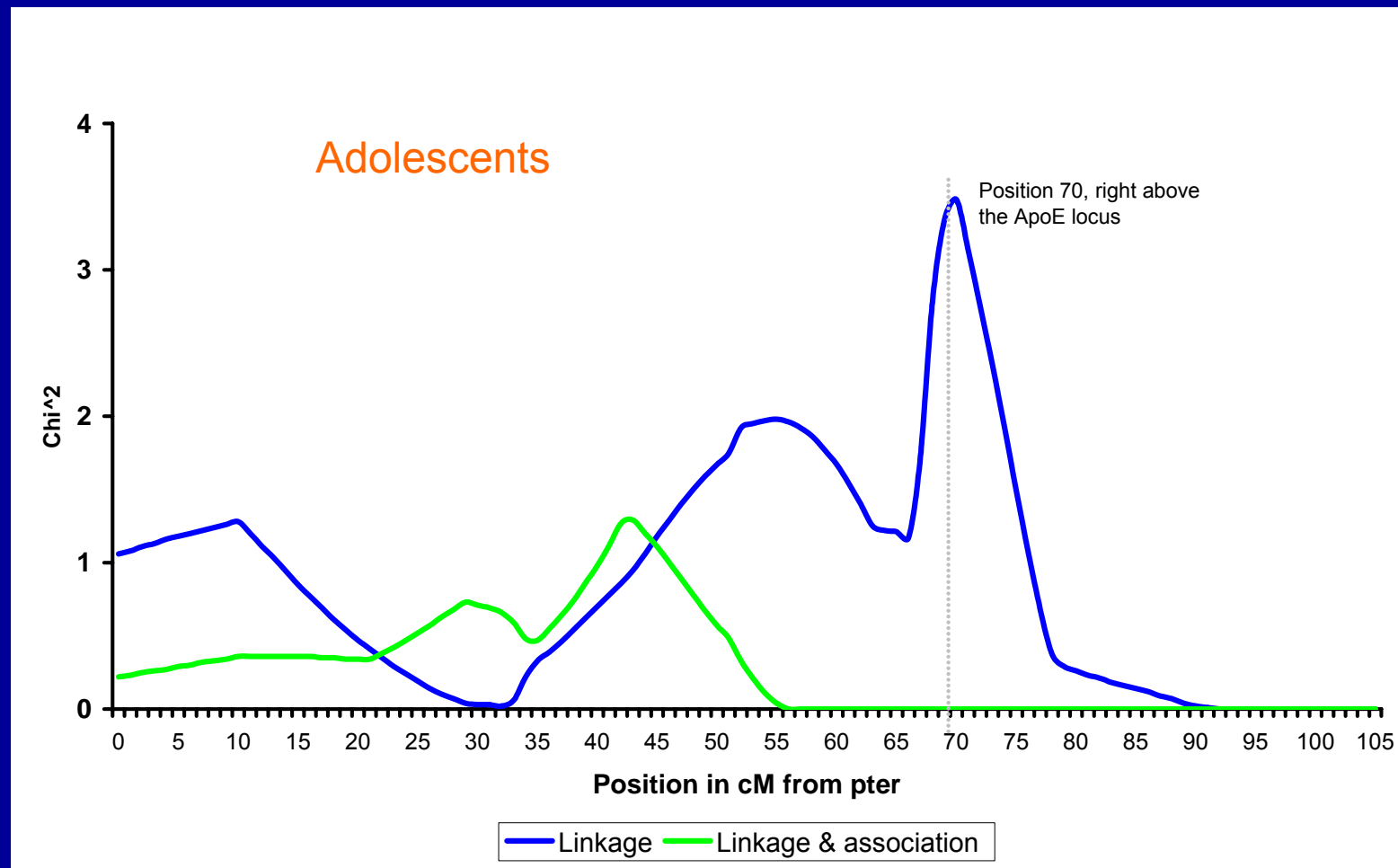
- We will run a combined linkage and association analysis on Dutch adolescents for apoe-level on chrom 19 using the apoe-gene in the means model, and will test for population stratification

Practical

- Open LinkAsso.mx, run it, fill out the table on the next slide and answer these questions:
- Is there evidence for population stratification?
- Does the apoe gene explain the linkage completely? Partly? Not at all?
- Is there association of the apoe gene with apoelevel?
- If you get bored: script LinkAsso.mx has several typos and mistakes in it: find all

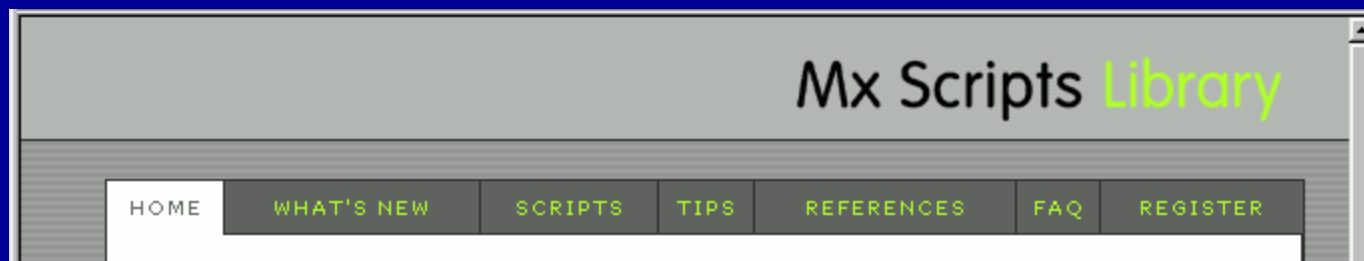
Model	Test	-2ll	df	Vs model	Chi ²	Df-diff	P-value
0				-	-		-
1	Linkage in presence of association						
2	B=W						
3	Dominance						
4	Full association						
5	Linkage in absence of association						

Linkage on chrom. 19 and association with APOE $\epsilon 2/\epsilon 3/\epsilon 4$ for apoE plasma levels



If there is time / Homework

- Take the table from Posthuma et al 2004 (ie Fulker model including dominance), and the biometrical model, and try to derive the within and between effects
- More scripts (ie including parental genotypes: Mx scripts library (<http://www.psy.vu.nl/mxbib>))



Funded by the [GenomEUtwin](#) project

(European Union Contract No. QLG2-CT-2002-01254)