

HUMAN LINKAGE AND ASSOCIATION ANALYSIS

Tuesday, September 20th, 2016

Other ways to estimate IBD

- Because, in reality, IBD cannot be “called” unambiguously;
- Not fully informative;
- Some alternatives exist:
 - Recurrence risk (without markers)
 - Using PMTs, disease MOI and penetrance

Other Ways of Estimating the Conditional IBD Probabilities, (Z_0, Z_1, Z_2) for ASP

Want to Compute : $Z_j = P(IBC = j | ASP)$, $j = 0, 1, 2$

$$P(IBC = j | ASP) = \frac{P(ASP | IBC = j)P(IBC = j)}{P(ASP)} \longrightarrow .25, .5, .25$$

$$P(ASP | IBC = j) = \sum_{k=1}^K P(ASP, MT_k | IBC = j) = \sum_{k=1}^K P(ASP | IBC = j, MT_k)P(MT_k)$$

$$P(ASP) = \sum P(ASP | MT_k)P(MT_k)$$

MT = parental mating types

Example

Calculate $P(\text{IBD}=0|\text{ASP})$ using Parental Mating Types

- Recessive Disease (A=disease allele)
- Assume: disease allele is rare, $p=0.001$; fully penetrant, no phenocopies:
 $(f_0, f_1, f_2) = (0, 0, 1)$

✓ $P(\text{IBD}=0) = 0.25$

Compute Numerator $P(ASP | IBD=0)$

$$P(ASP | IBD = 0) = \sum_{k=1}^6 P(ASP | IBD = 0, MT_k) P(MT_k)$$

This is a simplification:
assuming only 2 alleles

Remember: When dad and mom are not identical, multiply by 2!

k	PMT	P(MT _k)
1	aa,aa	q^4
2	Aa,aa	$4pq^3$
3	Aa,Aa	$4p^2q^2$
4	AA,aa	$2p^2q^2$
5	AA,Aa	$4p^3q$
6	AA,AA	p^4

Recessive Disease (A=disease allele)

Assume: disease allele is rare, $p=0.001$; fully penetrance, no phenocopies:

$$(f_0, f_1, f_2) = (0, 0, 1)$$

$$P(ASP | IBD=0) = 0 \cdot q^4 + 0 \cdot 4pq^3 + 0 \cdot 4p^2q^2 + 0 \cdot 2p^2q^2 + 0 \cdot 4p^3q + 1 \cdot p^4 = p^4$$

[Use frequencies in the MT matrix]

Example

Calculate $P(\text{IBD}=0|\text{ASP})$ using Parental Mating Types

- Recessive Disease (A=disease allele)
- Assume: disease allele is rare, $p=0.001$; fully penetrance, no phenocopies:
 $(f_0, f_1, f_2) = (0, 0, 1)$

✓ $P(\text{IBD}=0) = 0.25$

✓ $P(\text{ASP}|\text{IBD}=0)=p^4$

Calculate Denominator P(ASP)

$$P(ASP) = \sum_{k=1}^6 P(ASP | MT_k) P(MT_k)$$

$$= 0 \cdot q^4 + 0 \cdot 4pq^3 + \left(\frac{1}{4}\right)^2 4p^2q^2 + 0 \cdot 2p^2q^2 + \left(\frac{1}{2}\right)^2 4p^3q + 1 \cdot p^4$$

$$= p^4 + p^3q + \frac{1}{4}p^2q^2$$

<u>k</u>	<u>PMT</u>	<u>P(MT_k)</u>
1	aa,aa	q ⁴
2	Aa,aa	4pq ³
3	Aa,Aa	4p ² q ²
4	AA,aa	2p ² q ²
5	AA,Aa	4p ³ q
6	AA,AA	p ⁴

Example

Calculate $P(\text{IBD}=0|\text{ASP})$ using Parental Mating Types

- Recessive Disease (A=disease allele)
- Assume: disease allele is rare, $p=0.001$; fully penetrance, no phenocopies:
 $(f_0, f_1, f_2) = (0, 0, 1)$

✓ $P(\text{IBD}=0) = 0.25$

✓ $P(\text{ASP}|\text{IBD}=0)=p^4$

✓ $P(\text{ASP})=p^4 + p^3q + 0.25p^2q^2$

Putting it all together...

$$P(ASP) = \sum_{k=1}^6 P(ASP | MT_k) P(MT_k)$$

$$= 0 \cdot q^4 + 0 \cdot 4pq^3 + \left(\frac{1}{4}\right)^2 4p^2q^2 + 0 \cdot 2p^2q^2 + \left(\frac{1}{2}\right)^2 4p^3q + 1 \cdot p^4$$

$$= p^4 + p^3q + \frac{1}{4}p^2q^2$$

$$P(IBD = 0 | ASP) = \frac{P(ASP | IBD = 0)P(IBD = 0)}{P(ASP)}$$

$$= \frac{0.25p^4}{p^4 + p^3q + 0.25p^2q^2} = \frac{p^2}{(1+p)^2}$$

$$P(IBD = 1 | ASP) = \frac{2p}{(1+p)^2} \quad P(IBD = 2 | ASP) = \frac{1}{(1+p)^2}$$

<u>k</u>	<u>PMT</u>	<u>P(MT_k)</u>
1	aa,aa	q ⁴
2	Aa,aa	4pq ³
3	Aa,Aa	4p ² q ²
4	AA,aa	2p ² q ²
5	AA,Aa	4p ³ q
6	AA,AA	p ⁴

If p=0.001, then P(IBD=0|ASP) = 0.000 --z0

P(IBD=1|ASP) = 0.002 --z1

P(IBD=2|ASP) = 0.998 --z2

Example

Given a marker with 2 alleles, A and a.

Find **$\Pr(\text{PMT}=\text{Aa} \times \text{Aa} \mid \text{offspring}=\text{aa})$**

$$= \Pr(\text{offspring}=\text{aa} \mid \text{PMT}=\text{Aa} \times \text{Aa})\Pr(\text{PMT}=\text{Aa} \times \text{Aa}) / \Pr(\text{offspring}=\text{aa})$$

3 possible PMTs that will give an 'aa' offspring with:

- a priori probabilities (parental population in equilibrium)
Aa x Aa ($4p^2q^2$), Aa x aa ($4pq^3$), and aa x aa (q^4)
- conditional probabilities, (i.e., $\Pr(\text{aa}|\text{PMT})$) $\frac{1}{4}$, $\frac{1}{2}$, 1, resp.

Therefore, $\Pr(\text{offspring}=\text{aa}) = \frac{1}{4} \times 4p^2q^2 + \frac{1}{2} \times 4pq^3 + 1 \times q^4 = q^2$

$$\Pr(\text{PMT}=\text{Aa} \times \text{Aa} \mid \text{offspring}=\text{aa}) = \frac{1}{4} \times 4p^2q^2 / q^2 = p^2$$

The other 2 parental matings can be found in the same way.

Back to non-parametric linkage using
genetic data...

IBD Methods Using Pedigree Data

NPL Analysis and Scoring Functions

- “Allele sharing statistics” in pedigrees are calculated based on the “pattern of gene flow” in the pedigree.
- The calculation of the “NPL” statistic can be thought of in 2 parts:
 - Calculation of the inheritance distribution
 - Deciding upon a scoring function that determines if the inheritance information is indicative of linkage.
- Briefly, the statistic is then transformed and compared to a normal distribution for statistical significance.

The Scoring Function: NPL_{pairs} , NPL_{all}

Two statistics to 'measure' the degree of IBD similarity among affected individuals in the pedigrees.

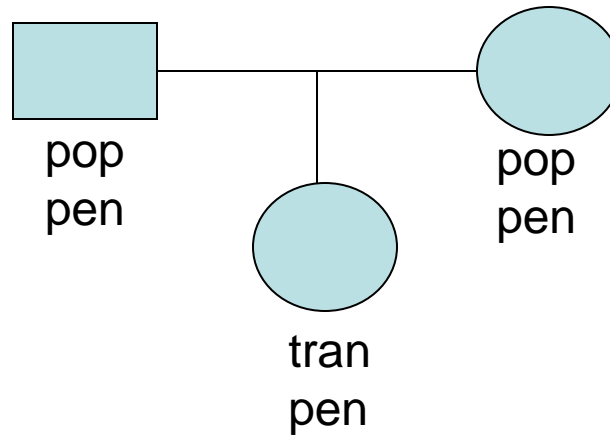
- (1) NPL_{pairs} : Essentially '**counts**' the number of alleles shared IBD by affected pairs and **sums** over individuals.
- (2) NPL_{all} : Considers, **jointly**, all relatives who share the same single allele IBD and **puts extra weight** on families with 3 or more individuals sharing the allele.

So, $NPL(\text{pairs}) = NPL(\text{all})$ for sibships with 2 affected individuals.

Whittamore & Halpern: Biometrics 50:109-117

Whittamore & Halpern: Biometrics 50:118-127

Elston-Stewart Algorithm



- Peeling /Clipping algorithm:
 - considers 2 parents and one offspring at a time;
 - Sums over all sets of 3, accounting for linking individuals;
 - Great for large pedigrees!
 - Problem: Considers ALL genotypes per summation - imagine 400 markers typed across 300 multiplex pedigrees!
 - Improvements in speed exist (e.g. 2-point)

Lander-Green Algorithm

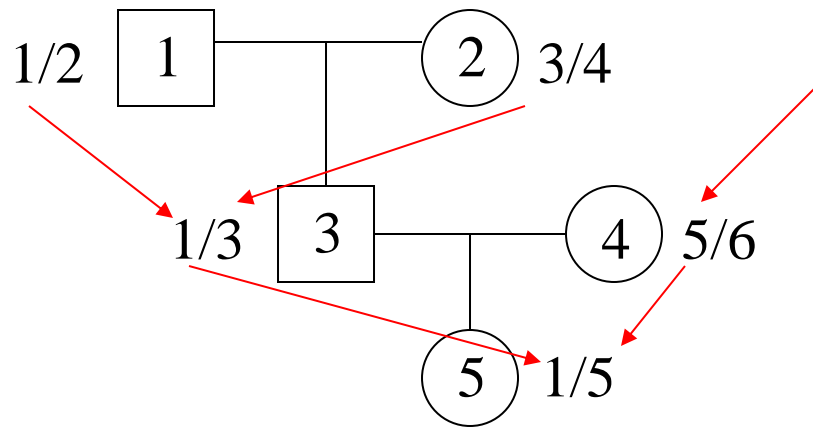
- Inheritance vectors (v): At a single locus, linkage information of each **non-founding** member can be captured by just two binary (0/1) digits, one for status of the allele in the paternal haplotype and one for the allele in the maternal haplotype.

$$P(x|v) = \sum_G P(x|g) P(g|v)$$

Notes

- The inheritance vector specifies which of the “founder’s” (f) alleles are inherited by each “non-founder”.
 - Each inheritance vector contains $2(n-f)$ ‘bits’
where n = total number of individuals in the pedigree.
And f = number of founders (no parental data)
 - Therefore, there are $2^{2(n-f)}$ possible vectors
- Difficult to determine the true inheritance vectors in a pedigree; genotype data often gives partial information
 - Untyped pedigree members
 - Limited heterozygosity
- In the absence of any genotype information, all inheritance vectors are equally likely, i.e. the ‘priors’ are all equal.

Basis of IBD Methods Using Pedigree Data



Inheritance vector, v , at locus $x = [p_1, m_1, p_2, m_2, \dots, p_n, m_n]$

where $n-f$ = the number of “non founders” and

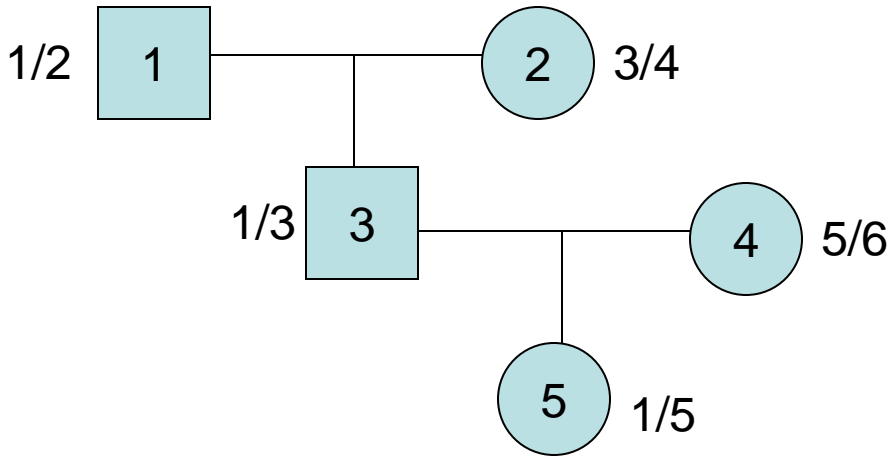
$$p_i = \begin{cases} 0 & \text{if grand-father's allele transmitted} \\ 1 & \text{if grand-mother's allele transmitted} \end{cases}$$

in the i th non founder.

Need to track Person 3 [$\frac{1}{3}$] and Person 5 [$\frac{1}{5}$]

Assume **phase is known**. Then $v(x) = [0, 0, \underline{0}, 0]$

IBD Methods Using Pedigree Data



If phase unknown:

Then $V(x) = [?, ?, \mathbf{0}, ?]$

Inheritance Vector	Prior	True	Posterior
0000	1/16	1	1/8
0001	1/16	0	1/8
0010	1/16	0	0
0011	1/16	0	0
0100	1/16	0	1/8
0101	1/16	0	1/8
0110	1/16	0	0
0111	1/16	0	0
1000	1/16	0	1/8
1001	1/16	0	1/8
1010	1/16	0	0
1011	1/16	0	0
1100	1/16	0	1/8
1101	1/16	0	1/8
1110	1/16	0	0
1111	1/16	0	0

The Number of Possible Inheritance Vectors

<u>Number of non-founders</u>	<u>Number of vectors</u>	<u>Prior p</u>
1	4	1/4
2	16	1/16
3	64	1/64
:	:	:
7	16,384	1/16,384
8	65,536	1/65,536
:	:	:
15	1,073,741,824	1/1,073,741,824

- Thus Merlin and Allegro (and to a greater extent GH) are hindered by the size of the pedigrees analyzed.

Computing “nbits”

- Compute the following number from your pedigree
 $\text{nbits} = 2(n-f)$
- The default for GH is $2(n-f) \leq 16$, although this may be made a bit larger.
- Thus, for moderately sized pedigrees (and lots of markers), MERLIN, GH, or Allegro are the programs of choice.
- For larger pedigrees (and fewer markers) the LINKAGE program, FASTLINK, is preferred. (See Strauch et al.: Hum Hered v55, 2003)

MERLIN

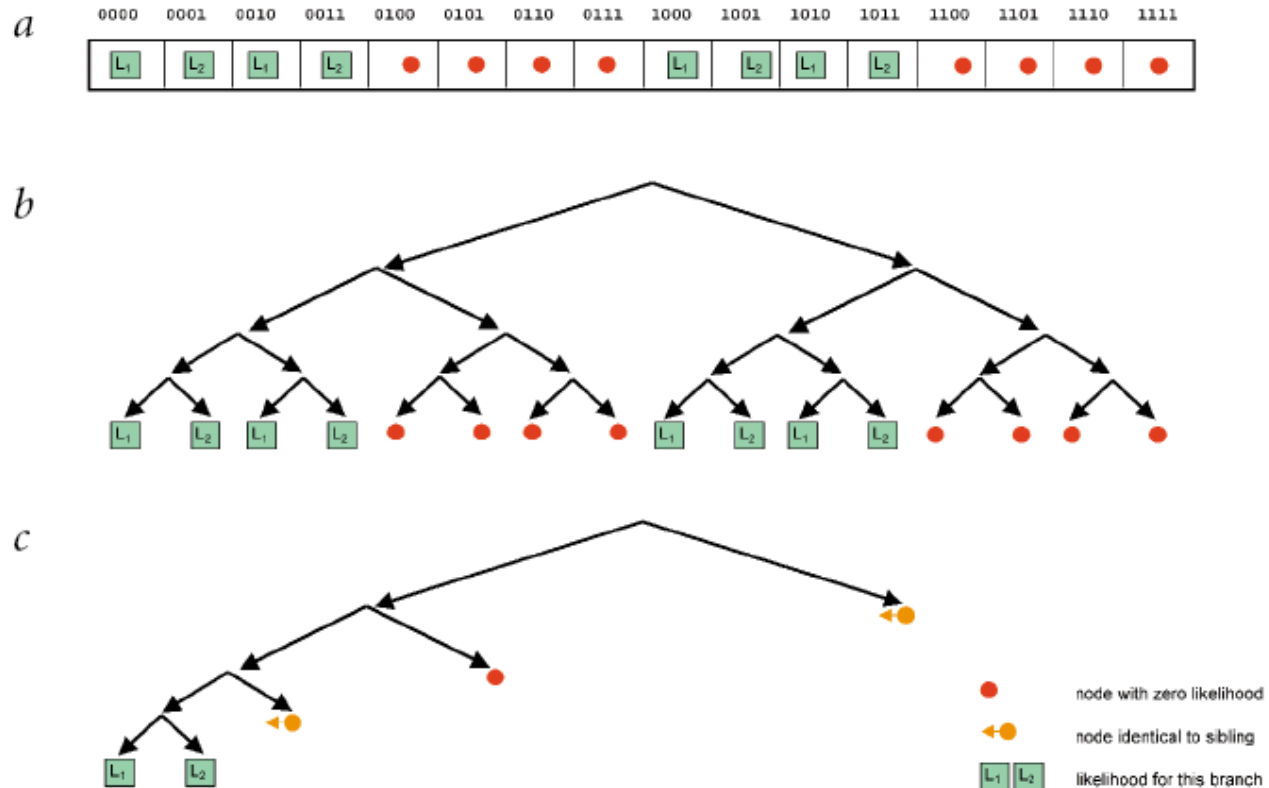
Merlin—rapid analysis of dense genetic maps using sparse gene flow trees

Gonçalo R. Abecasis^{1,2}, Stacey S. Cherny¹, William O. Cookson¹ & Lon R. Cardon¹

Published online: 3 December 2001, DOI: 10.1038/ng786

- Multipoint Engine for Rapid Likelihood Interference
- Lander Green inheritance vectors
- Meiosis trees marking gene flow – branches and nodes
- Some meioses have identical outcomes and can be collapsed into premature nodes – sparse trees – increase efficiency

Gene Flow Trees



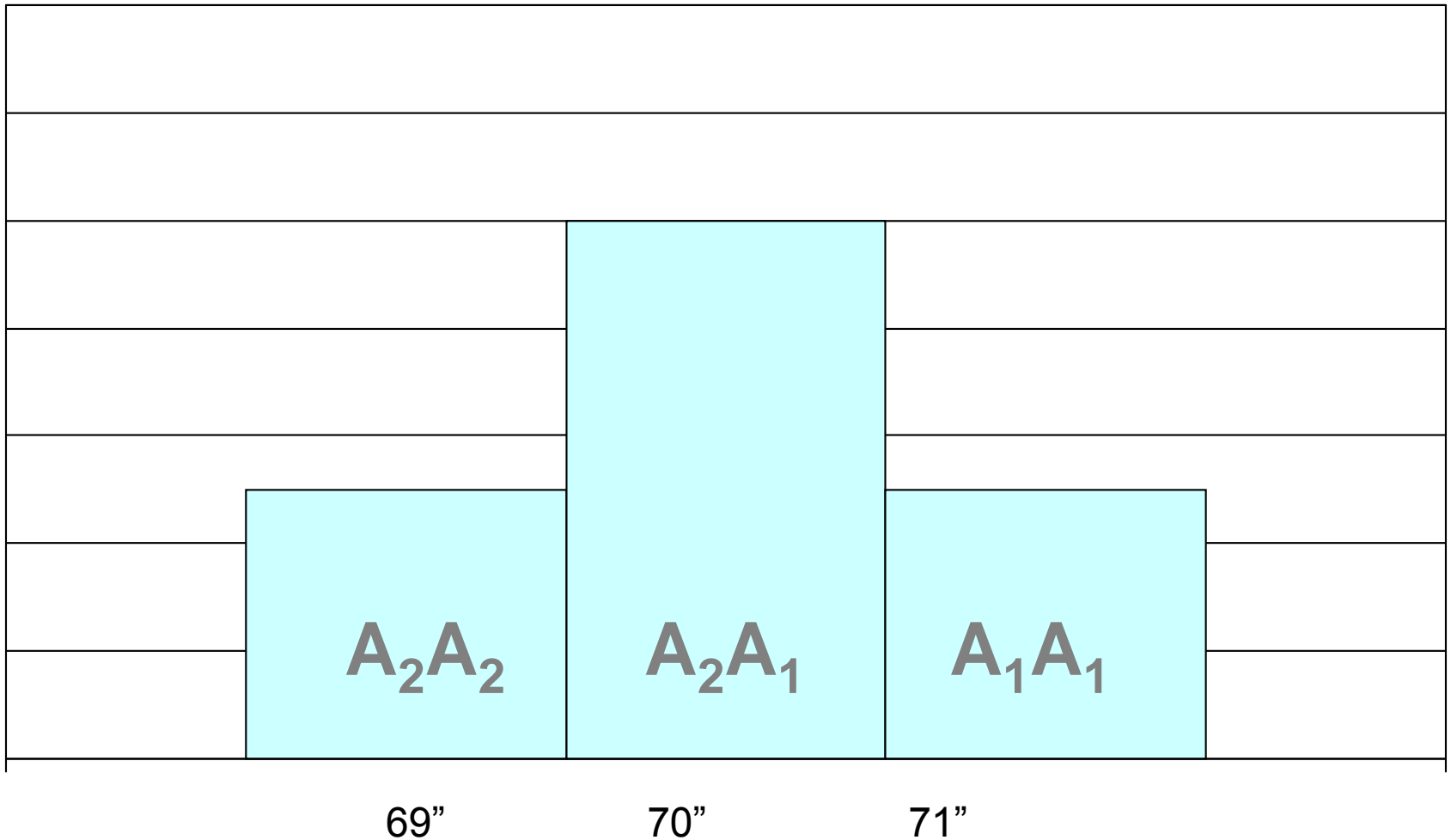
VARIANCE COMPONENTS APPROACH

Single-locus Completely Additive Model

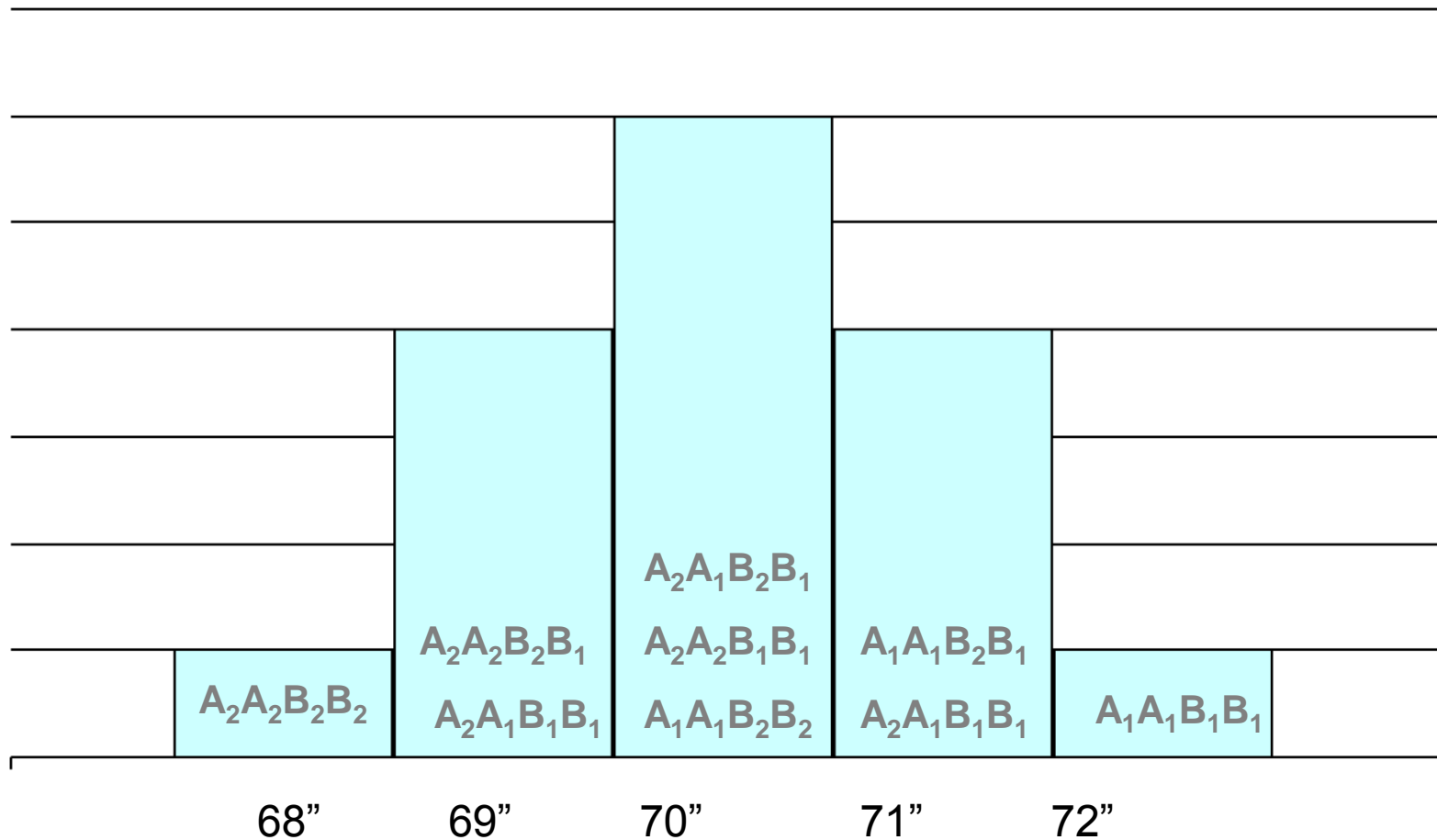
Genotype:	A_1A_1	A_1A_2	A_2A_2
Phenotype:	71"	70"	69"

$$P = G$$

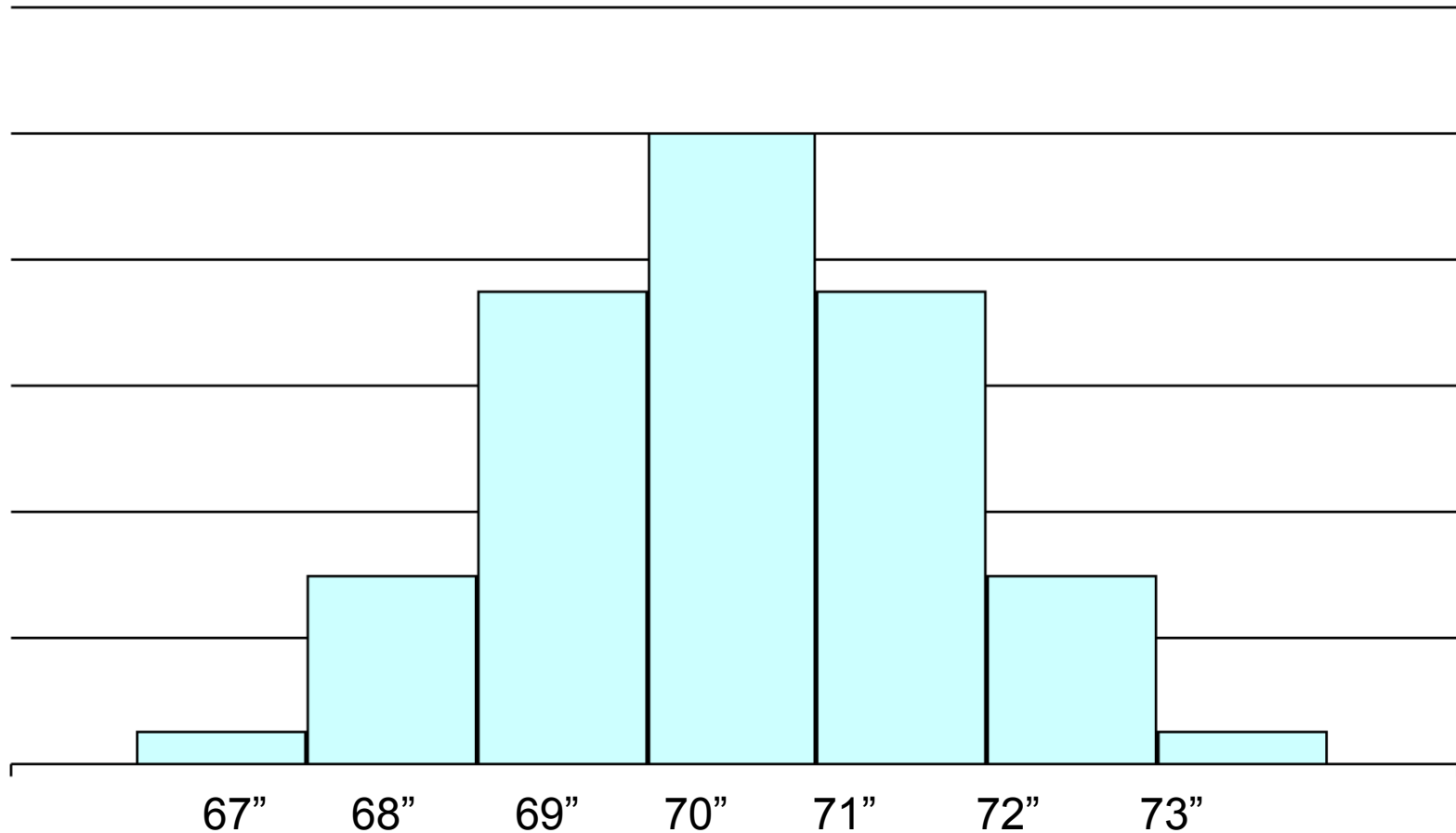
Single-Locus Phenotypic Distribution



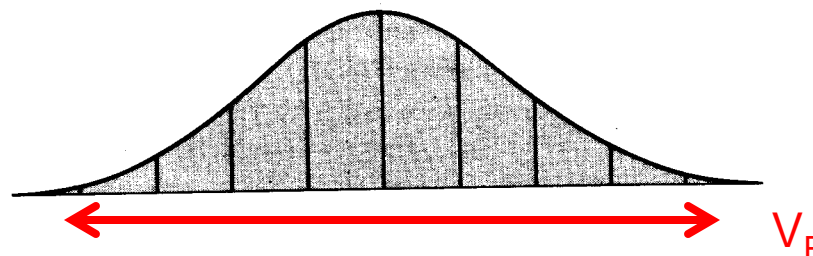
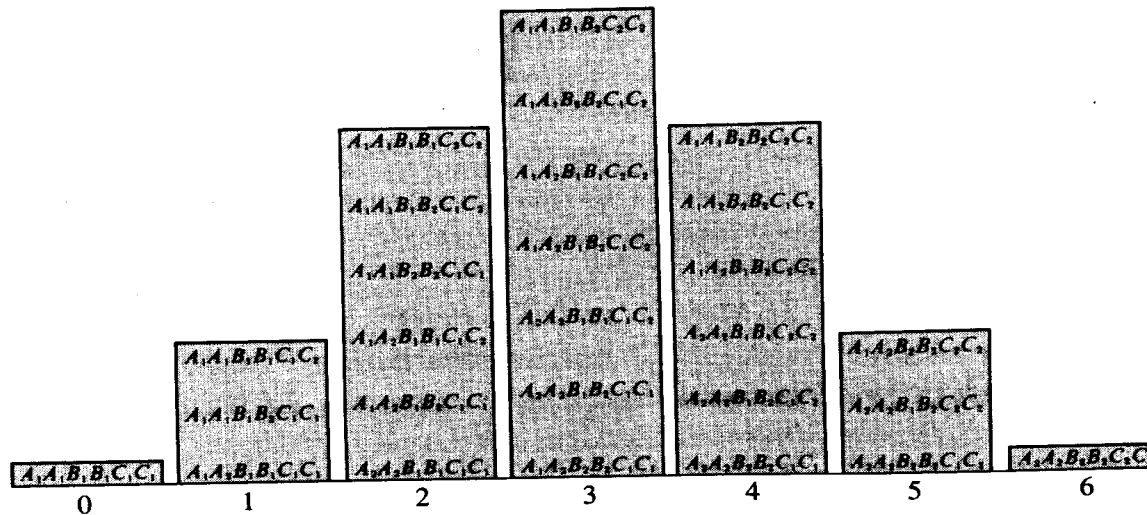
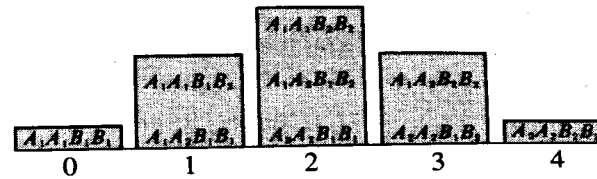
Two-Locus Phenotypic Distribution



Three-Locus Phenotypic Distribution



Infinite Number of Loci Model



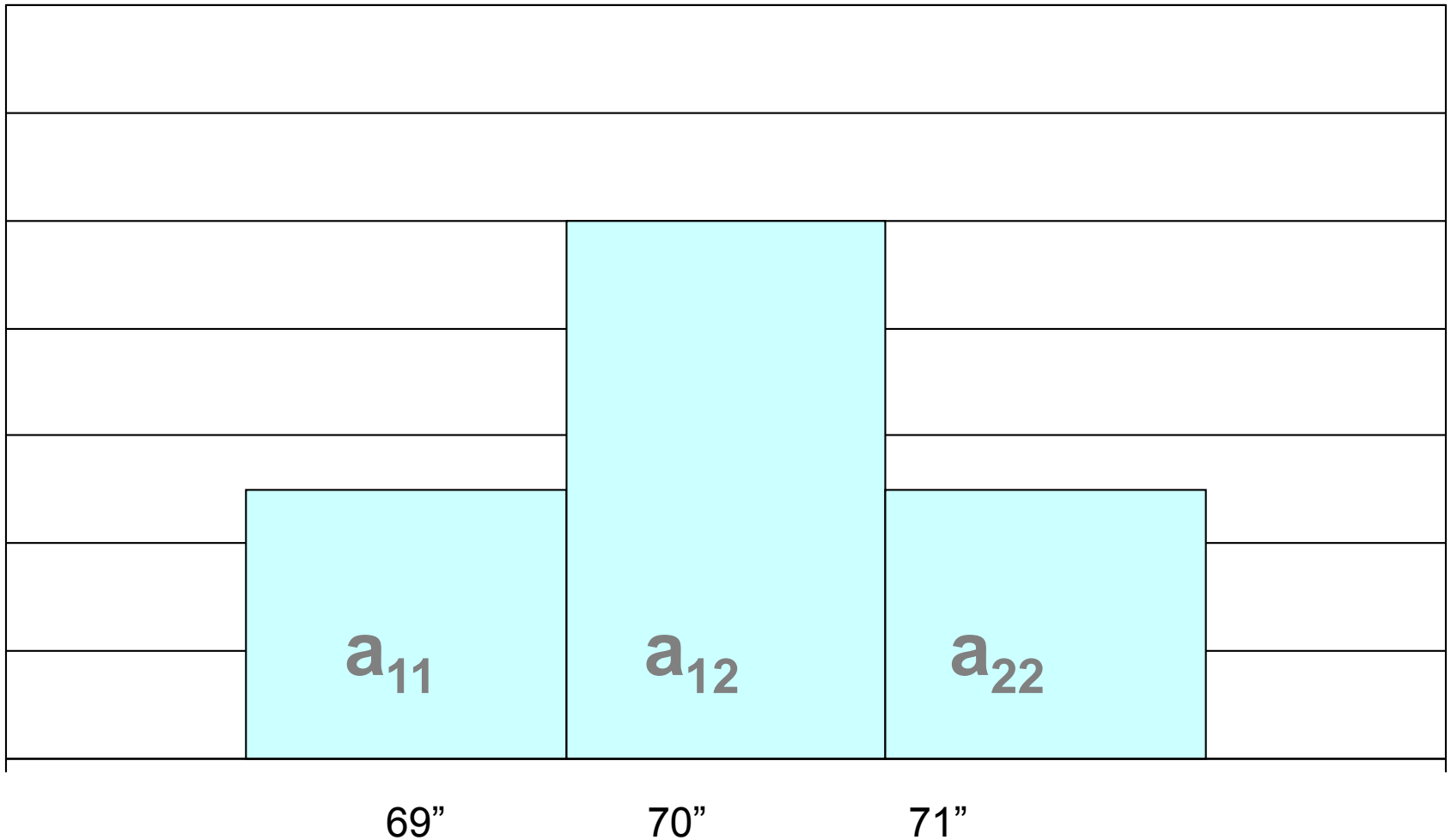
- Polygenic Model:

$$P = G + E$$

$$V_P = V_G + V_E$$

- can estimate V_G from correlations (between relatives)
- usually done in a variance components (VC) framework

Single-Locus Phenotypic Distribution



The Biometrical Model

- Consider 1 locus / 2 alleles (A_1 & A_2) , with population frequencies p_1 & p_2

$$y_{11} = \bar{a} + a_{11}$$

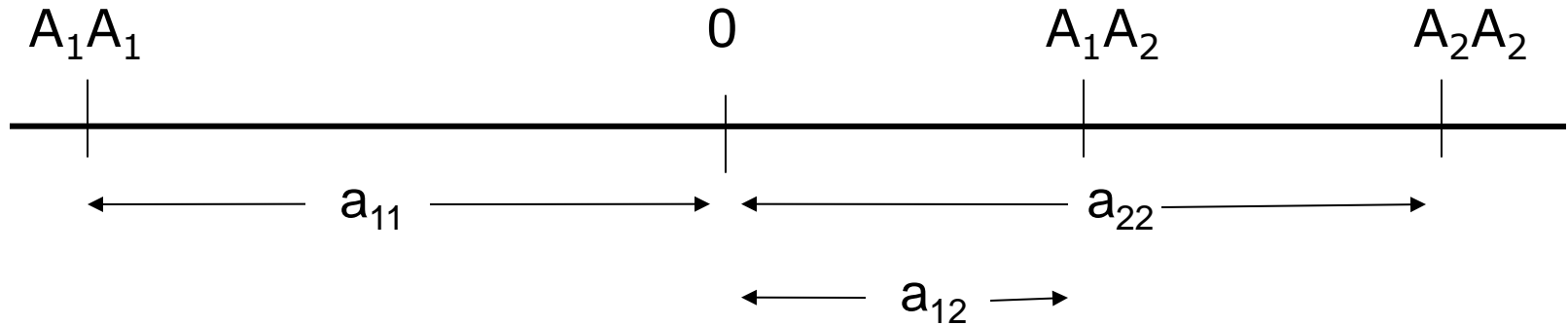
$$y_{12} = \bar{a} + a_{12}$$

$$y_{22} = \bar{a} + a_{22}$$

where,

\bar{a} = population mean

a_{ij} = deviations introduced by gene substitution



If population mean is standardized to zero ($\bar{a} = 0$),

- Mean effect of the locus:

**N.B. H-W
assumption:
 $p_1 + p_2 = 1$**

(p_2 same as q)

$$\mu_g = p_1^2 a_{11} + 2 p_1 p_2 a_{12} + p_2^2 a_{22}$$

- Variance:

$$\sigma_g^2 = p_1^2 a_{11}^2 + 2 p_1 p_2 a_{12}^2 + p_2^2 a_{22}^2$$

Obtain:

$$\sigma_a^2 = 2p_1p_2 [p_1(a_{11} - a_{12}) + p_2 (a_{12} - a_{22})]^2$$

THE ADDITIVE GENETIC VARIANCE!

Subtract σ_a^2 from σ_g^2 :

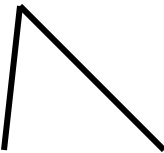
$$(\sigma_g^2 = p_1 a_{11}^2 + 2 p_1p_2a_{12}^2 + p_2 a_{22}^2)$$

$$\sigma_d^2 = p_1^2p_2^2 (a_{11} - 2a_{12} + a_{22})^2$$

THE DOMINANCE VARIANCE!

When taken over a large number of loci:

$$V_P = V_G + V_E$$


$$V_P = V_A + V_D + V_E$$

Broad sense heritability - the proportion of variance due to genetic (familial) effects:

$$H = V_G / V_P$$

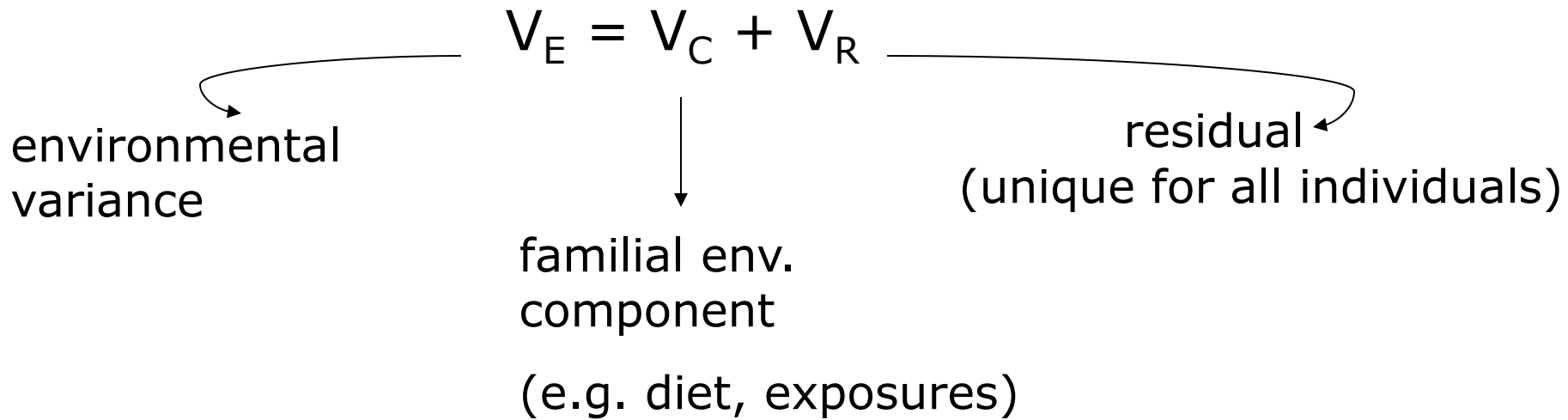
Narrow sense heritability – the proportion of variance due *strictly to additive effects*:

$$h^2 = V_A / V_P$$

Additivity assumed in most estimation models

- no dominance effects
- no interactions, i.e. $G \times E$, $A \times A$, $A \times D$, $D \times D$

Can expand this to accommodate alternative models:
e.g.



$$V_P = V_G + V_C + V_R$$

- Must have special information to dissect genetic and familial environmental components
- In general, V_G will pick up all / most additive familial effects, i.e. pseudopolygenic (really multifactorial & inflated with non-additive familial effects)

So, how do we estimate these components of variance?

- We do this on the basis of **correlations between relatives of a pair** (e.g. sib, p-o, GP-GC, etc)
- The expected correlations for relatives of a particular type is a function of the degree to which they share common genotypes, and the latent additive and dominance factors.

IDENTITY RELATIONS BETWEEN RELATIVES

- We begin with the expected sharing of alleles **IIDENTICAL BY DESCEND (IBD)**

To derive the expected correlations between relatives, we need the probabilities of sharing 0,1,2 alleles IBD for each type of relative pair.

These were derived by Cotterman (1940)

“K-coefficients”

Consider 2 related individuals i, j

$$k_0 = \Pr (\text{no alleles are IBD})$$

$$2k_1 = \Pr (1 \text{ allele IBD})$$

$$k_2 = \Pr (\text{both alleles IBD})$$

- Can get these by complete enumeration (e.g. sibs)
- Cotterman developed expressions to compute these as functions of the coefficients of consanguinity f_{ij} ’s

IDENTITY BY DESCENT (IBD) Distribution for sibs by complete enumeration

Consider the following fully informative mating: AB x CD

Sibship Genotype		IBD		
		0	1	2
AC	AC			X
AC	AD		X	
AC	BC		X	
AC	BD	X		
AD	AC		X	
AD	AD			X
AD	BC	X		
AD	BD		X	
BC	AC		X	
BC	AD	X		
BC	BC			X
BC	BD		X	
BD	AC	X		
BD	AD		X	
BD	BC		X	
BD	BD			X
TOTAL		4	8	4
		$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

The expectation is that sib pairs share, on average, $\frac{1}{2}$ of their genes identical by descent

Measurement of Phenotypic Similarity

- **Continuous phenotypes:**
 - All individuals are potentially informative
 - *Deviation* – squared difference in phenotypes
 - *Covariance* – product of differences from the mean
- **Disease phenotypes:**
 - Concordance (affected, affected)
 - Amount of information is greatest for affected
 - affected pairs, so they are mostly used.

Measurement of Phenotypic Similarity (more on this later . . .)

- **Continuous phenotypes:**
 - All individuals are potentially informative
 - *Deviation* – squared difference in phenotypes
 - *Covariance* – product of differences from the mean
- **Disease phenotypes:**
 - Concordance (affected, affected)
 - Amount of information is greatest for affected
 - affected pairs, so they are mostly used.

The **expected phenotypic correlation** between trait values for any pair of relatives is:

$$\text{cov} (y_1, y_2) = E [(y_1 - \mu)(y_2 - \mu)]$$

$$= \sum [(1/2k_{1i} + k_{2i}) \sigma^2_{qi}]$$

summed over $i=1$ to n loci

where k_j are Cotterman's k-coefficients, the probability of sharing 0,1,or 2 alleles IBD

Standardize by dividing with the phenotypic variance:

$$\rho (y_1, y_2) = \sum [(1/2k_{1i} + k_{2i}) h^2_{qi}],$$

where h^2_{qi} is the QTL-specific heritability

(of course, if the sibs share no alleles IBD k_0 , there is no contribution to the covariance)

When the QTLs are not specifically measured (by IBD), we model the residual polygenic component using the **expectation** over the genome of the k_{ji} probabilities to obtain approximation:

$$\text{cov} (y_1, y_2) = \mathbf{2 \Phi} \sigma^2_g$$

where Φ is *kinship coefficient* (1/2 for sibs), and 2Φ is the coefficient of relationship

$$\sigma^2_g = \sum \sigma^2_{qi}, \text{ and}$$

$$\Phi = 1/2 E [k_{1i} / 2 + k_{2i}]$$

(see slide from lecture 1 on kinship coefficients)

Let's say we are focusing on the i^{th} QTL – the effect of any remaining QTLs can be folded into the residual genetic component, i.e.,

$$\text{cov}(y_1, y_2) = \pi_i \sigma_{qi}^2 + 2 \Phi \sigma_g^2$$

$\pi_i \sigma_{qi}^2$ is the contribution of the i^{th} QTL

$2 \Phi \sigma_g^2$ is the residual “polygenic” component

$\pi_i = k_{1i} / 2 + k_{2i}$ (proportion of alleles IBD at QTL i),
is **estimated from genetic marker data** and
information on the genetic map.

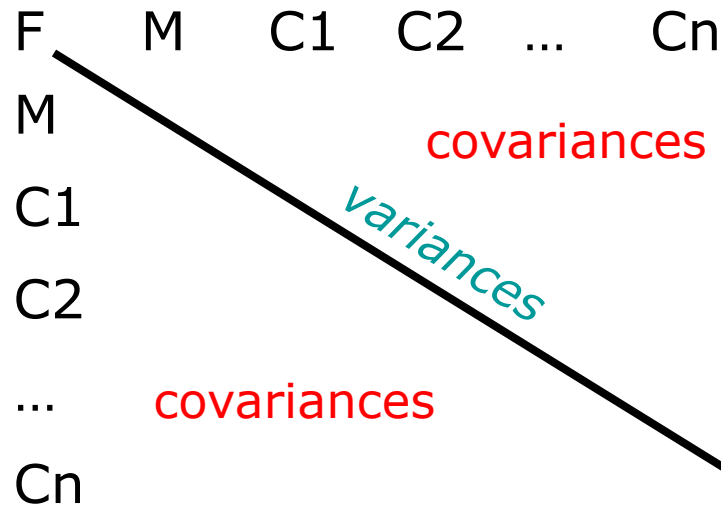
π_i and 2Φ both are **structuring parameters** for the
covariance, enabling estimation of their respective
variance components.

This simple model can be extended to a general pedigree.
The covariance for the relative pair is replaced by the covariance matrix for the pedigree:

$$\text{cov} \rightarrow \mathbf{\Omega} = \mathbf{\Sigma} \pi_i \sigma_{q_i}^2 + \overbrace{2 \mathbf{\Phi}}^{\text{Coef of relationship}} \sigma_g^2 + \mathbf{I} \sigma_e^2 \rightarrow \text{Environmental variance}$$

Matrix IBD prob effect of i^{th} QTL Polygenic variance

Covariance matrix:



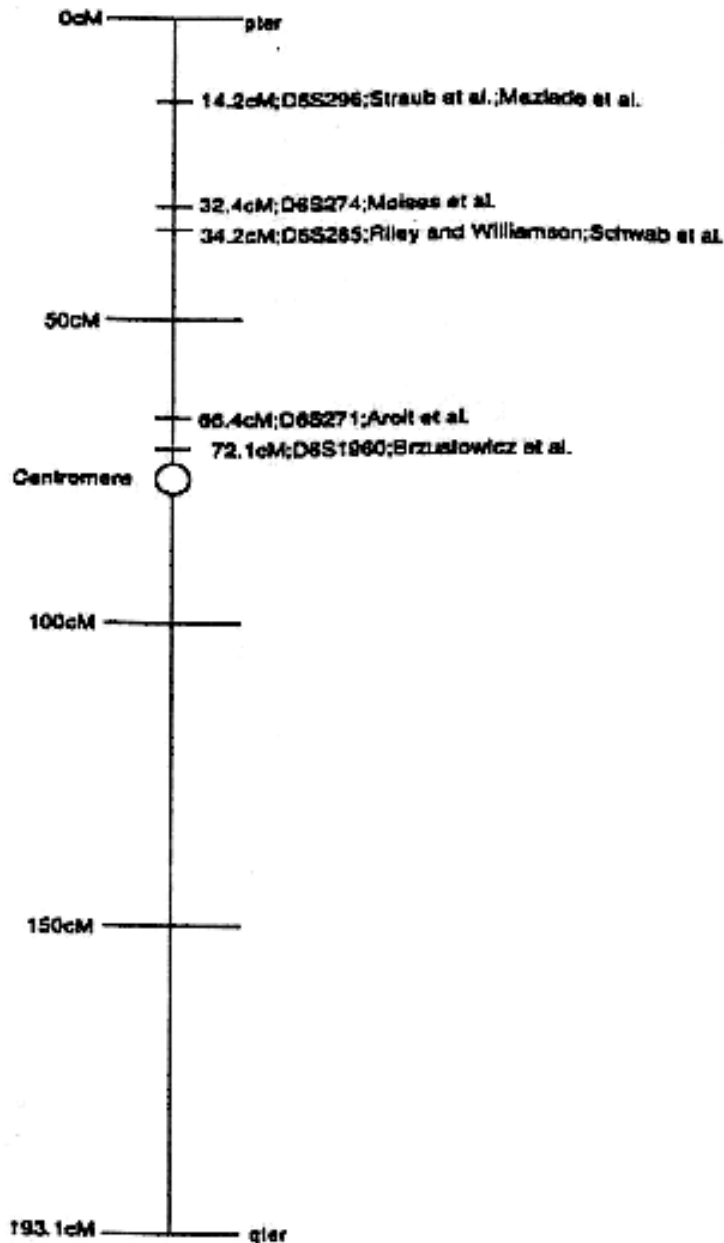
Why do we prefer extended families to sib pairs?

- ❖ Families are much more **information-rich** than a single type of relative pair.
- ❖ A **diversity of different types of relationship** and, thus, different levels of expectation for genetic similarity lends additional **power and accuracy** for estimating genetic effects.
 - wrt allele-sharing at specific loci, there is greater power for **gene detection**, and greater accuracy in estimating locus-specific heritability
 - wrt average allele-sharing, **polygenic variance is much more accurately estimated**, with the estimate truly approaching polygenic variance uncontaminated by environmental sources of resemblance.

How accurately can a gene be localized using linkage analysis?

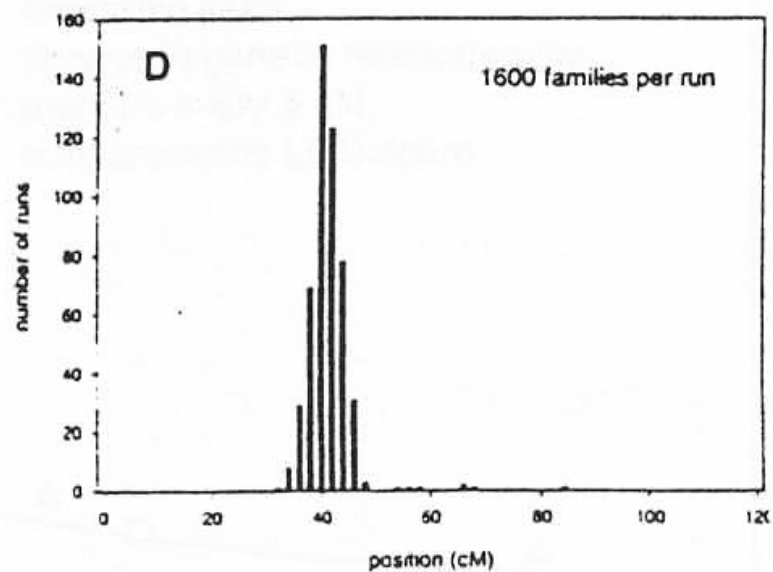
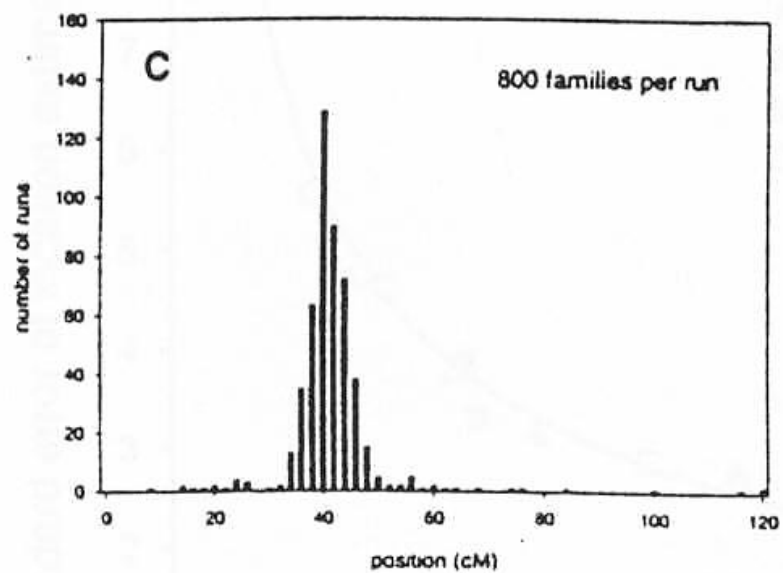
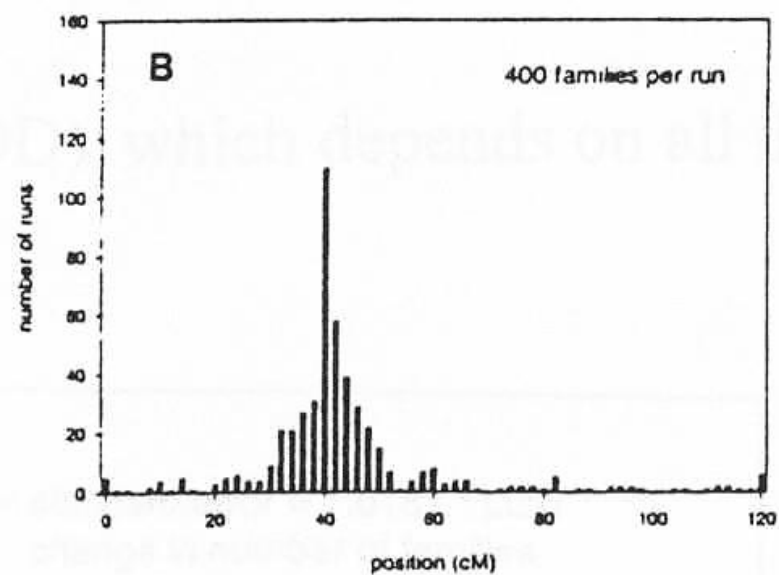
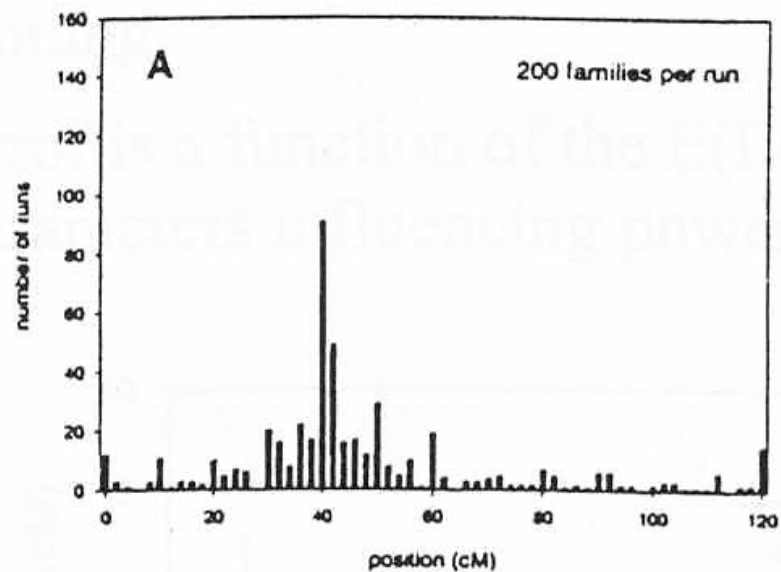
Roberts et al. (1999)

- Motivated by the question: *In considering results from separate studies, how can you decide whether a given study has replicated an initial study's finding given there is considerable variation in the location estimate?*
- Consider the example of multiple studies aimed at identifying loci influencing schizophrenia:

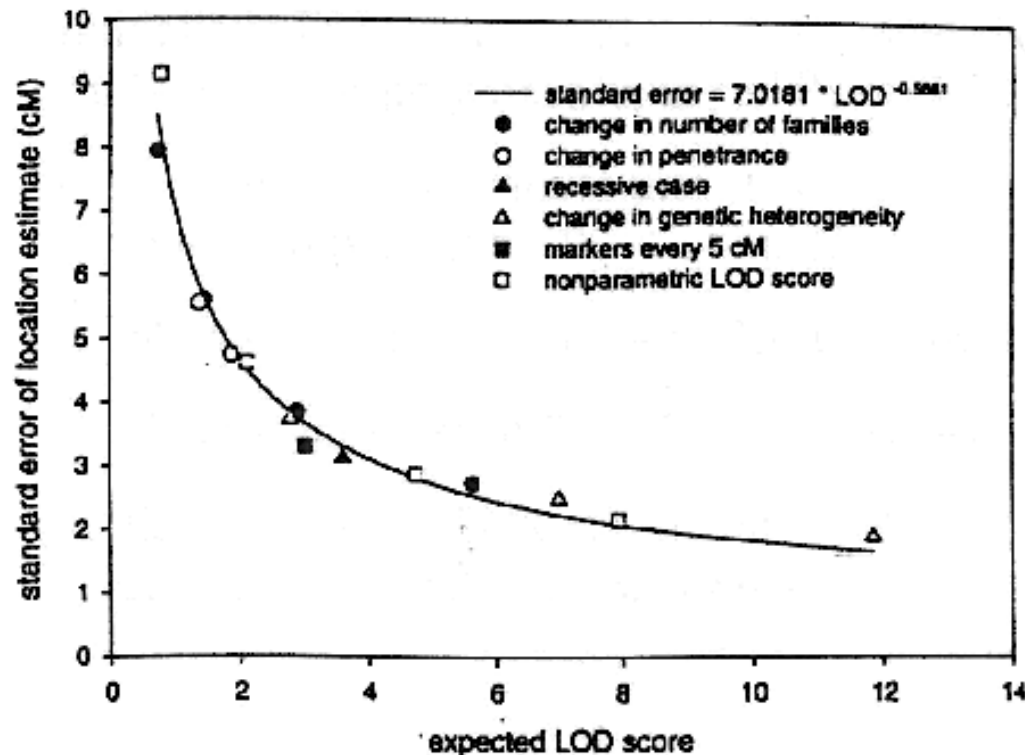


- **2 studies found evidence for linkage to schizophrenia-related phenotypes.**
 - Arolt et al. used *eye tracking dysfunction*, an endophenotype
 - Brzustowicz et al. used severity of positive psychotic symptoms.
- Location of all signals on chr 6p scattered over a **50-60 cM region**.
- Could these findings have resulted from the **same susceptibility locus?**
- Is the 6p linkage **replicated??**
- Depends on the expected variability in location estimates, and the **resolution of linkage methods**.

- Roberts et al. investigated the **sampling distributions for location estimates** by computer simulation (500 reps).
 - *Simulate data, marker and trait locus (including incomplete penetrance, heterogeneity, and phenocopies)*
 - *Analyze the data*
 - *Record location of peak lod score*
- Examined **different sample sizes** (N=200, 400, 800, 1600 nuclear families) to reflect increasing information content and power.
 - *More information, better localization*
 - *True location of disease locus @ 41.4 cM*



- Error decreases with increasing sample size.
- Even with a large number of families, the 95% CI spans **10's of cM** – this is a huge genetic distance to search over or for positional cloning.
- Error is a function of the **E(LOD)**, which depends on all the usual parameters influencing power (penetrance / location).



Results from Duggirala et al. (nuclear families)

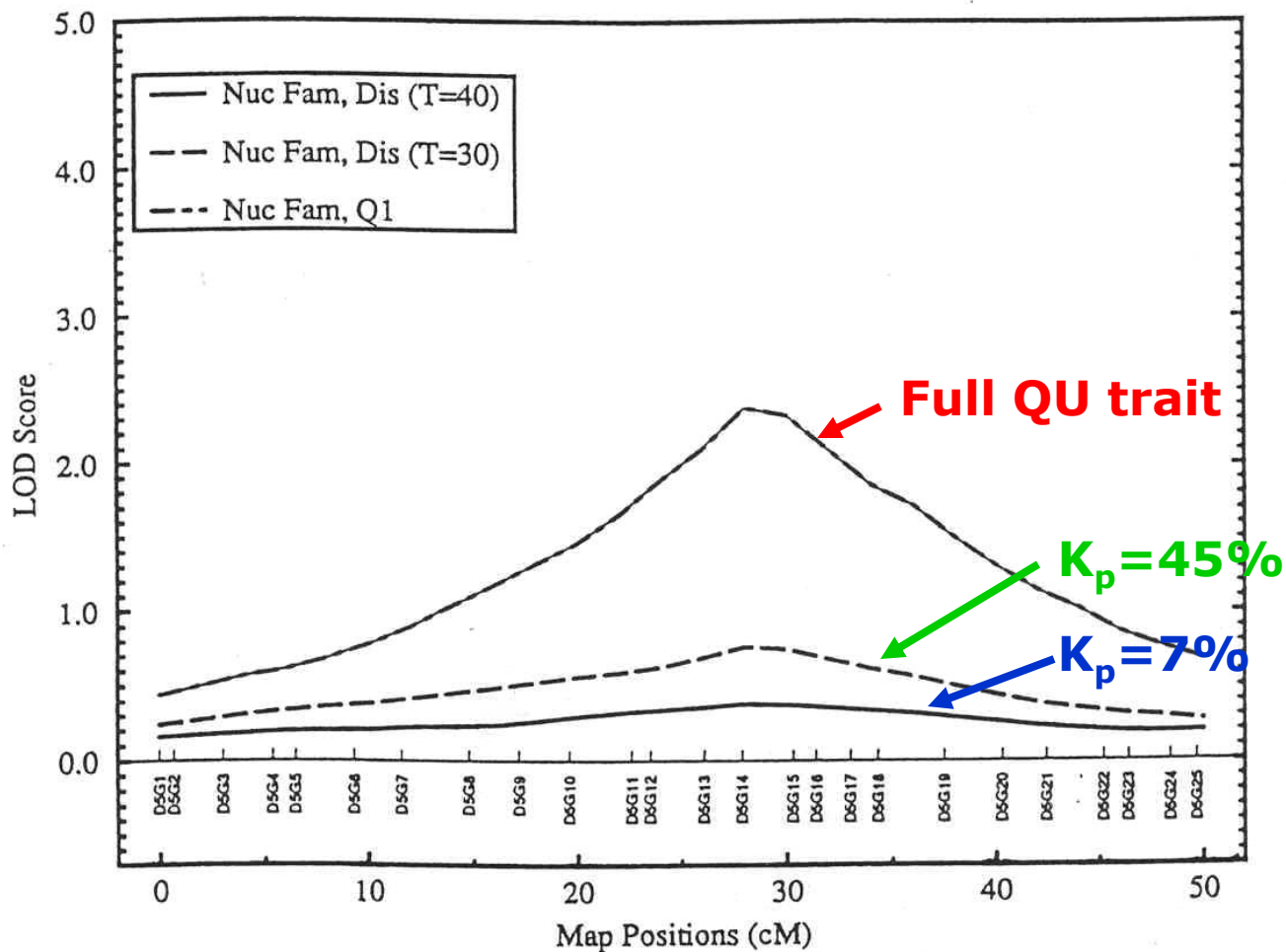
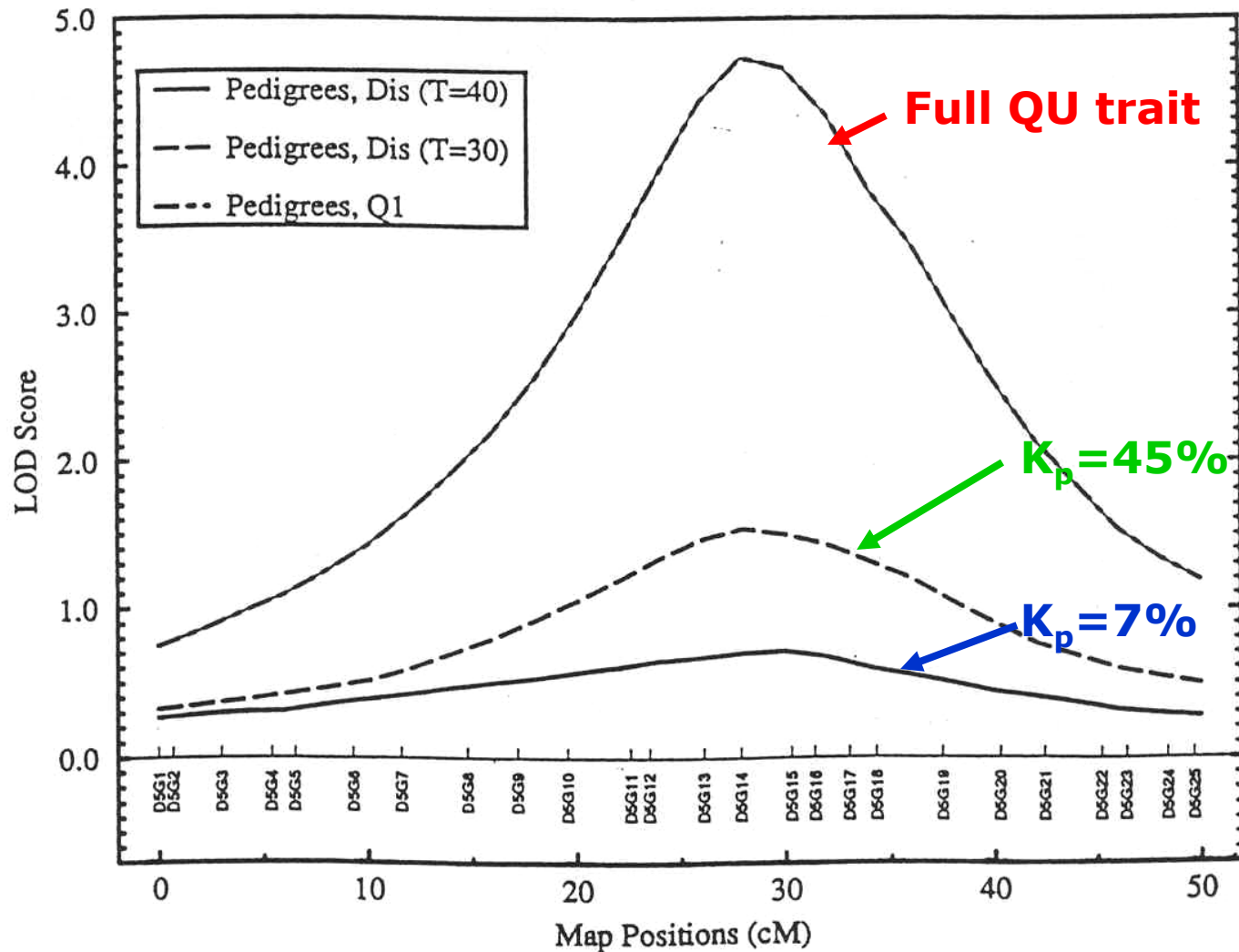


Fig. 1. Plot of lod scores obtained from discrete and quantitative trait analyses using nuclear families vs. map positions on chromosome 5.

Results from Duggirala et al. (pedigrees)



Covariate Adjustments

In approaching the analysis of a quantitative trait, it is important to **adjust for concomitant variation**.

Consider the model:

$$V_P = V_g + V_G + V_E$$

$$V_P = V_g + V_G + V_{cov} + V_E$$

Where, V_{cov} can be measured. Say,

$$100\% = 20\% + 30\% + \mathbf{35\%} + 15\%$$

Estimating Covariate Effects

- Adjustment for concomitants can be done prior to genetic analysis, or simultaneously estimating covariate effects in the context of the genetic analysis.
- Typical factors to adjust for include age, sex, race, and then if possible, measurable relevant factors, e.g. medication use, smoking, diet, etc.
- Adjustment for covariate effects is important because it reduces the residual variation, thereby increasing the relative variance due to the effects we seek, providing increased power for statistical detection

$$100\% = 20\% + 30\% + 50\%$$

Locus-specific heritability

$$h^2_g = 20\%$$

Total heritability

$$h^2_T = 50\%$$

Adjust for the measured covariates:

$$100\% = 20\% + 30\% + \del{35\%} + 15\%$$

$$0.65 = 0.20 + 0.30 + 0.15$$

Locus-specific heritability

$$h^2_g = 0.20/0.65 = 31\%$$

Total heritability

$$h^2_T = 0.50/0.65 = 77\%$$

***Can improve the
relative QTL effect***

- **Conclusions:**

- ❖ Pedigrees yield higher average lod scores than do nuclear families
- ❖ Using all quantitative variation yields roughly 2-5 times the average lod scores than the dichotomized data.
- ❖ VC threshold model does better with common endpoint than a “rare” one.

***Analysis of quantitative variation is more powerful than discrete traits any day!
(However, note that other approaches are better for rare disease traits!)***

Summary

- Using PMTs to estimate IBD – another way to get at degree of relatedness;
- Inheritance vectors; L-G versus E-S algorithm;
- Biometrical approach to locus identification;
- Variance components analysis