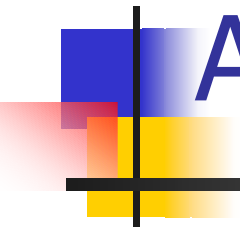
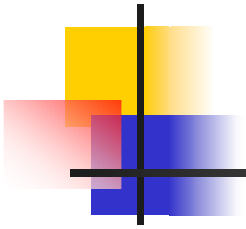


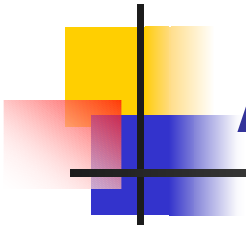
Association Mapping in Families

Gonçalo Abecasis
University of Oxford



Linkage Analysis

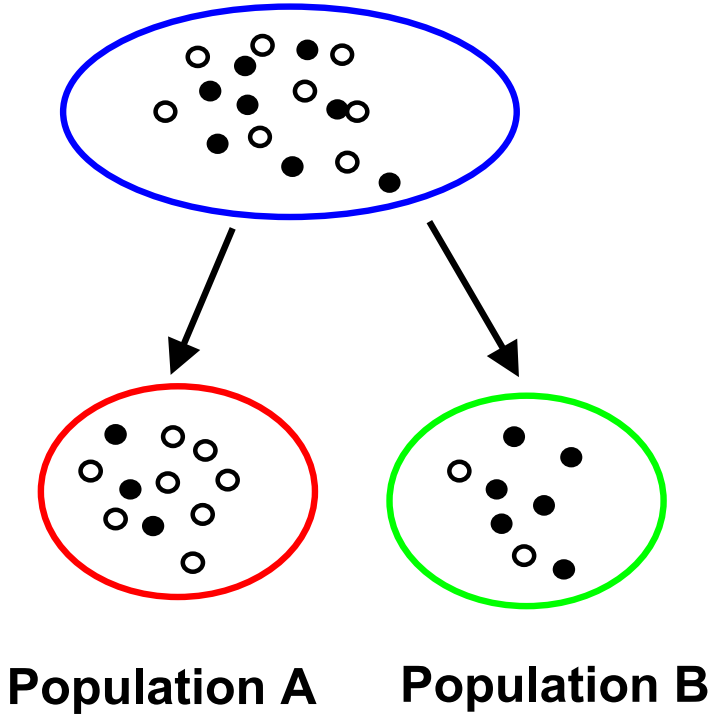
- Sharing between relatives
  - Identifies large regions
    - Include several candidates
- Complex disease
  - Scans on sets of small families popular
  - No strong assumptions about disease alleles
  - Low power
  - Limited resolution

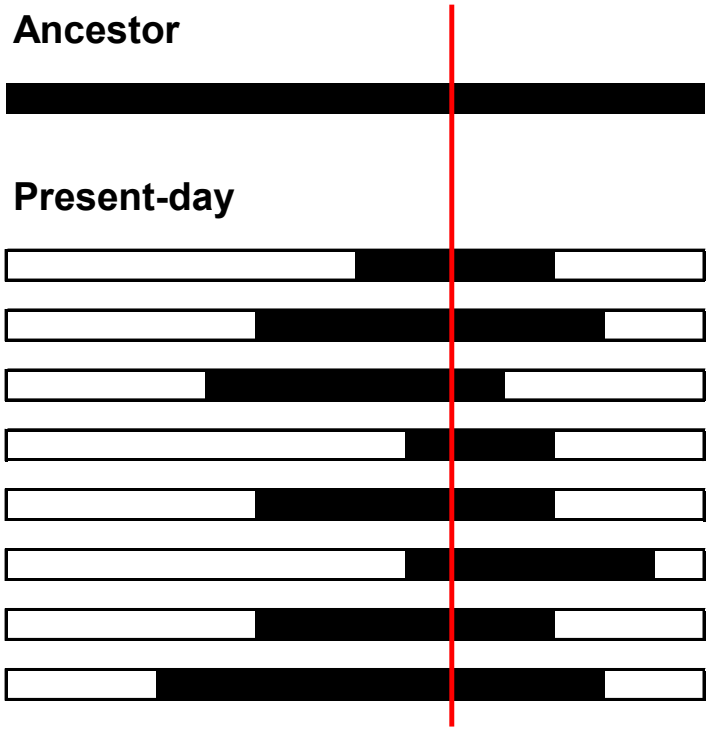


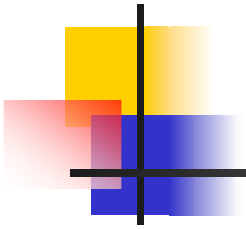
Association Analysis

- Sharing between *unrelated* individuals
  - Trait alleles originate in common ancestor
  - High resolution
    - Recombination since common ancestor
    - Large number of independent tests
- Powerful if assumptions are met
  - Same disease haplotype shared by many patients
- Sensitive to population structure

Stratification *vs* Disequilibrium







Disequilibrium Mapping

- Control for possible population structure
  - Distinguish linkage disequilibrium from other types of association
- Family-based association analysis
  - Using families collected for linkage mapping
- Powerful if assumptions are met
  - Same disease haplotype shared by many patients
- High-resolution

Essential Notation

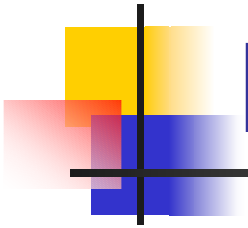
- i families
- $j = 1 .. n_i$ offspring in family i
- y_{ij} quantitative phenotype
- g_{ij} no. of '1' alleles at marker
- g_{iF}, g_{iM} parental genotypes, optional
- π_{ijk} IBD between j and k in family i
- φ_{ijk} kinship between j and k in family i

Controlling for Stratification

- If stratum were known…
  - For each individual genotype (g_{ij})
  - Average number of alleles in a strata (b_{ij})
  - Adjust for stratum differences ($w_{ij} = g_{ij} - b_{ij}$)

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

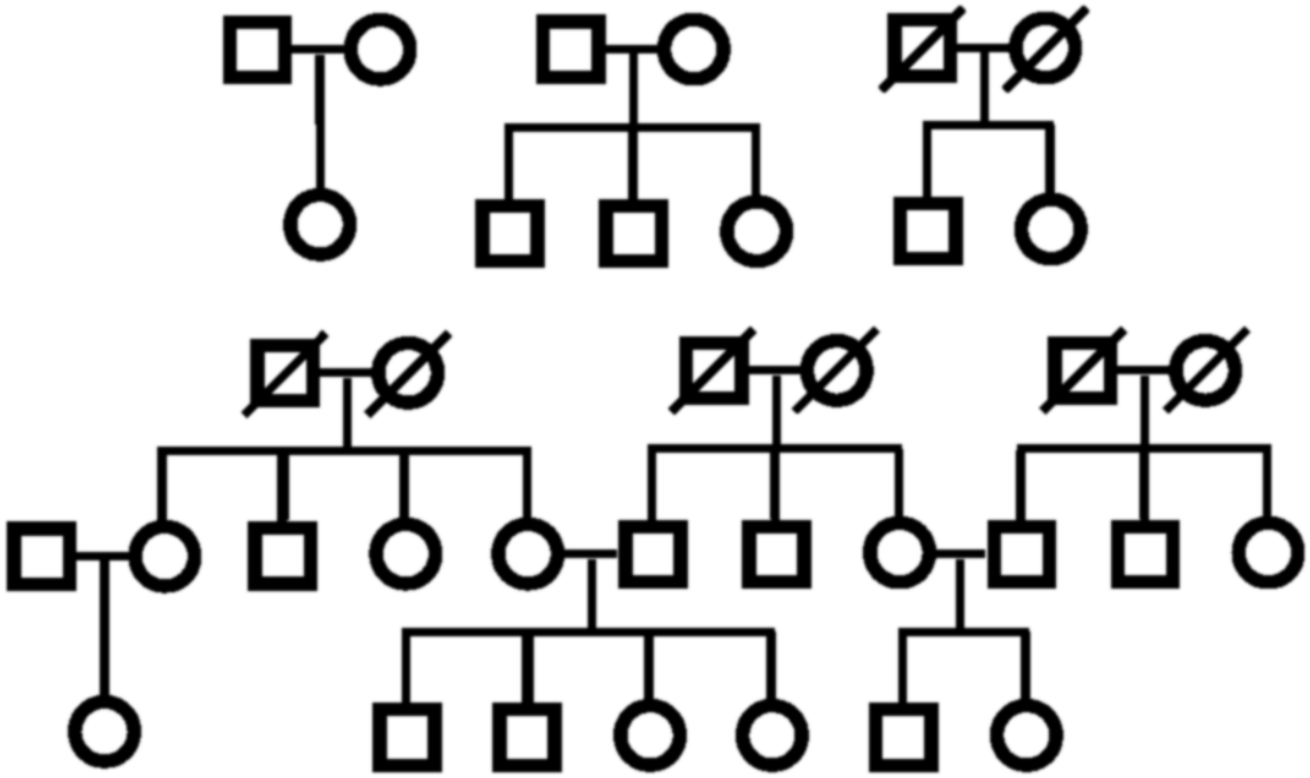
- How to define stratum then?
  - Use family data to estimate b_{ij}



b_{ij} as Family Control

- Expected genotype for each individual
  - Ancestors
  - Siblings
- Informative individuals
  - Genotype may differ from expected
  - Have heterozygous ancestor in pedigree

Allowable Family Structures



Nuclear Families

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

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

Extended Families

$$b_{ij} = \begin{cases} \dfrac{b_{iF_j} + b_{iM_j}}{2} & \text{average of parental controls} \\ \displaystyle\sum_{k}^{sibship} \frac{g_{ik}}{n_{sibs}} & \text{average of sibling genotypes} \\ g_{ij} & \text{self - genotype} \\ undefined & \text{otherwise} \end{cases}$$

Allowing for Related Data

- Similarities between individuals
  - Variance–covariance matrix
- Major gene, polygenic, environment

$$\Omega_{ijk} = \begin{cases} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \text{if } j = k \\ \pi_{ijk}\sigma_a^2 + 2\varphi_{ijk}\sigma_g^2 & \text{if } j \neq k \end{cases}$$

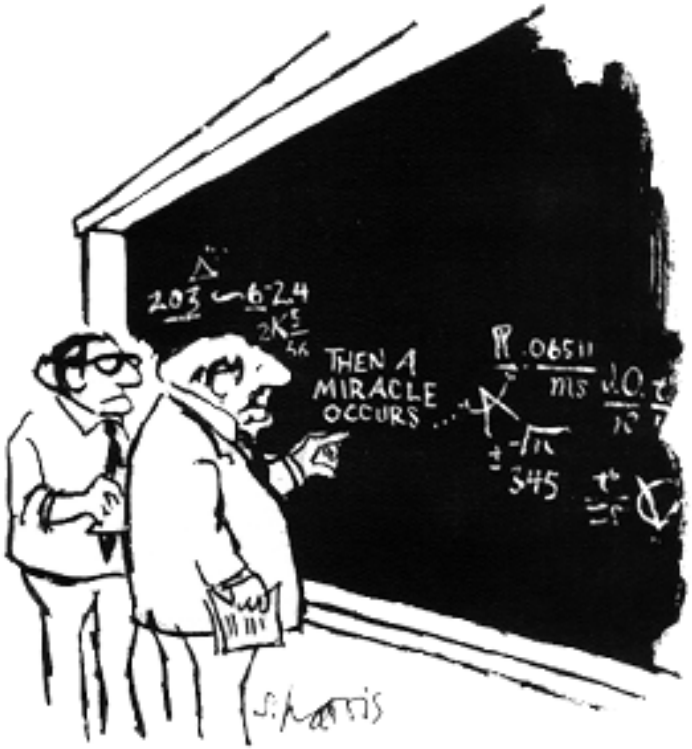
Likelihood function

- Multivariate Normal Distribution
- Defines asymptotic significance levels

$$L = \prod_{i} (2\pi)^{-n_i/2} \left| \hat{\boldsymbol{\Omega}}_{\mathbf{i}} \right|^{-1/2} e^{-1/2[(\mathbf{y_i} - \hat{\mathbf{y}}_{\mathbf{i}})' \hat{\boldsymbol{\Omega}}_{\mathbf{i}}^{-1} (\mathbf{y_i} - \hat{\mathbf{y}}_{\mathbf{i}})]}$$

$$\chi^2 = 2 \ln \left(\frac{L_{\text{full model}}}{L_{\text{sub-model}}} \right)$$

Parameter Derivations



"I think you should be more explicit here in step two."

$$Model = (\mu, \beta_b, \beta_w, \sigma_e^2, \sigma_g^2, \sigma_a^2)$$

$$\begin{bmatrix} \beta_b \\ \\ \beta_w \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_i n_i (p_i - q_i)\mu_i}{NV_b} + a \\ \\ a \end{bmatrix}$$

$$a = \frac{D}{pq} a_{QTL}$$

$$\sigma_a^2 = V_{QTL} - 2pqa$$

Exact Permutation Test

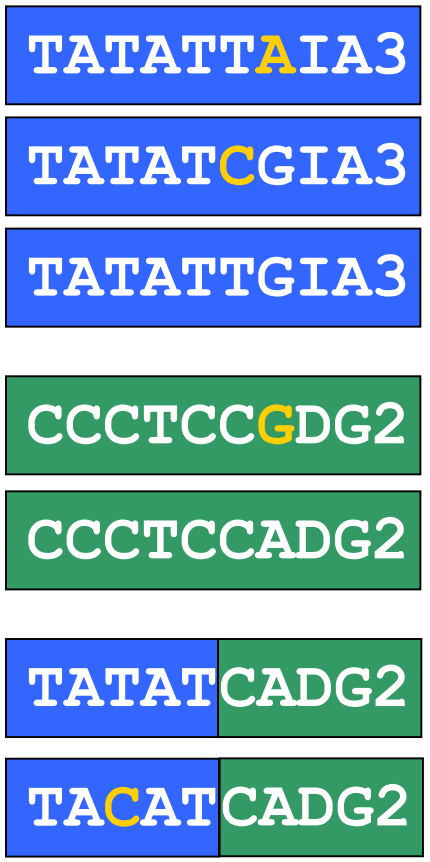
- In each family, $\mathbf{w_i} = [w_{i1}, w_{i2} \ldots]$ is the pattern of allelic transmission
  - $\mathbf{w_i}$ and $\mathbf{-w_i}$ are equally likely (H_o)
- Null distribution of the data
  - Randomly permute any set families by replacing each $\mathbf{w_i}$ with itself or $\mathbf{-w_i}$ with equal probability
  - The permuted data sets define the null distribution of the maximum likelihood statistic
- Empirical significance levels

Application: Angiotensin-1

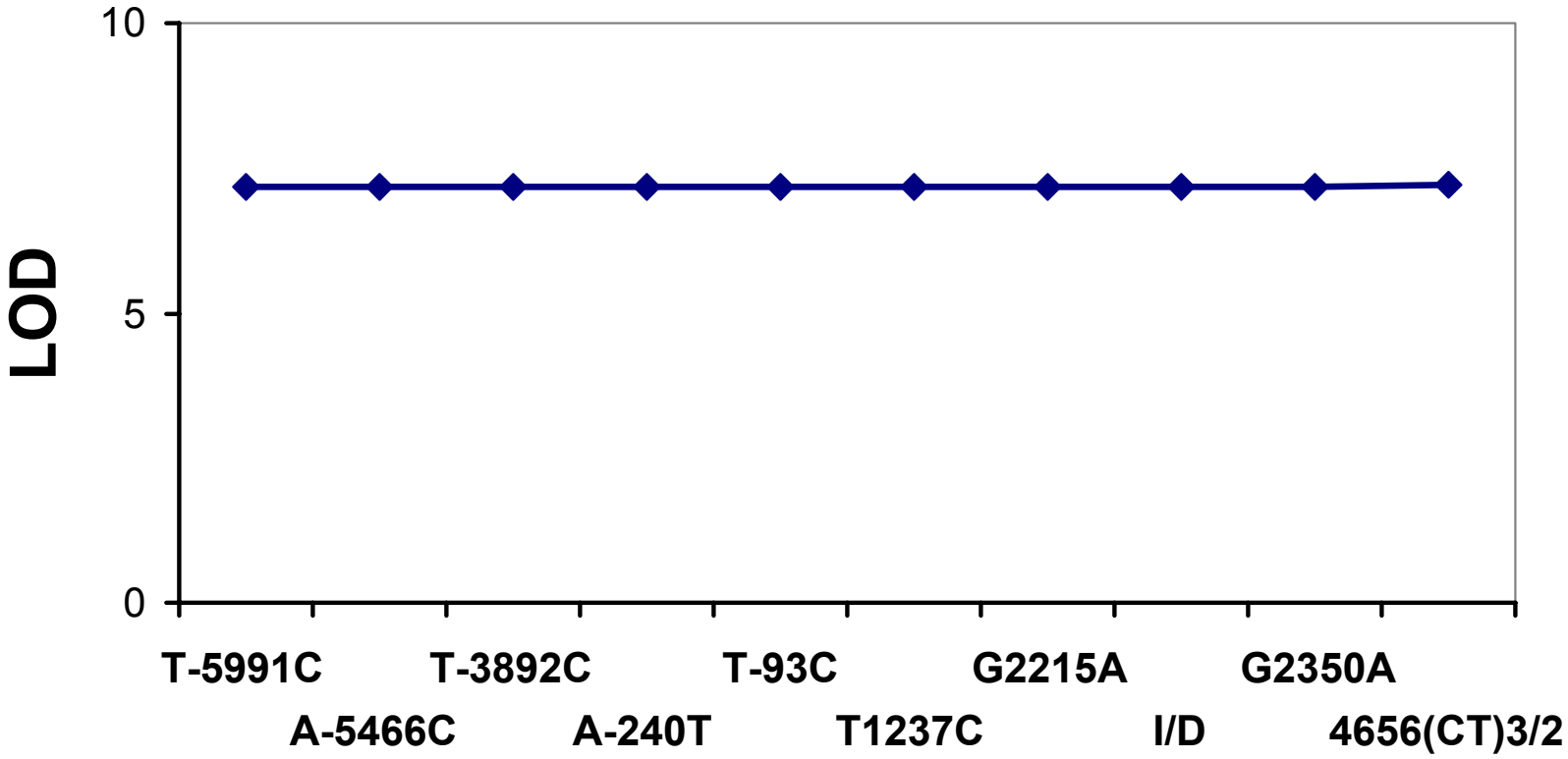
- British population
- Circulating ACE levels
  - Normalized separately for males / females
- 10 di-allelic polymorphisms
  - 26 kb
  - Common
  - In strong linkage disequilibrium
- Keavney et al, HMG, 1998

Haplotype Analysis

- 3 clades
  - All common haplotypes
  - >90% of all haplotypes
- "B" = "C"
  - Equal phenotypic effect
  - Functional variant on right
- Keavney et al (1998)

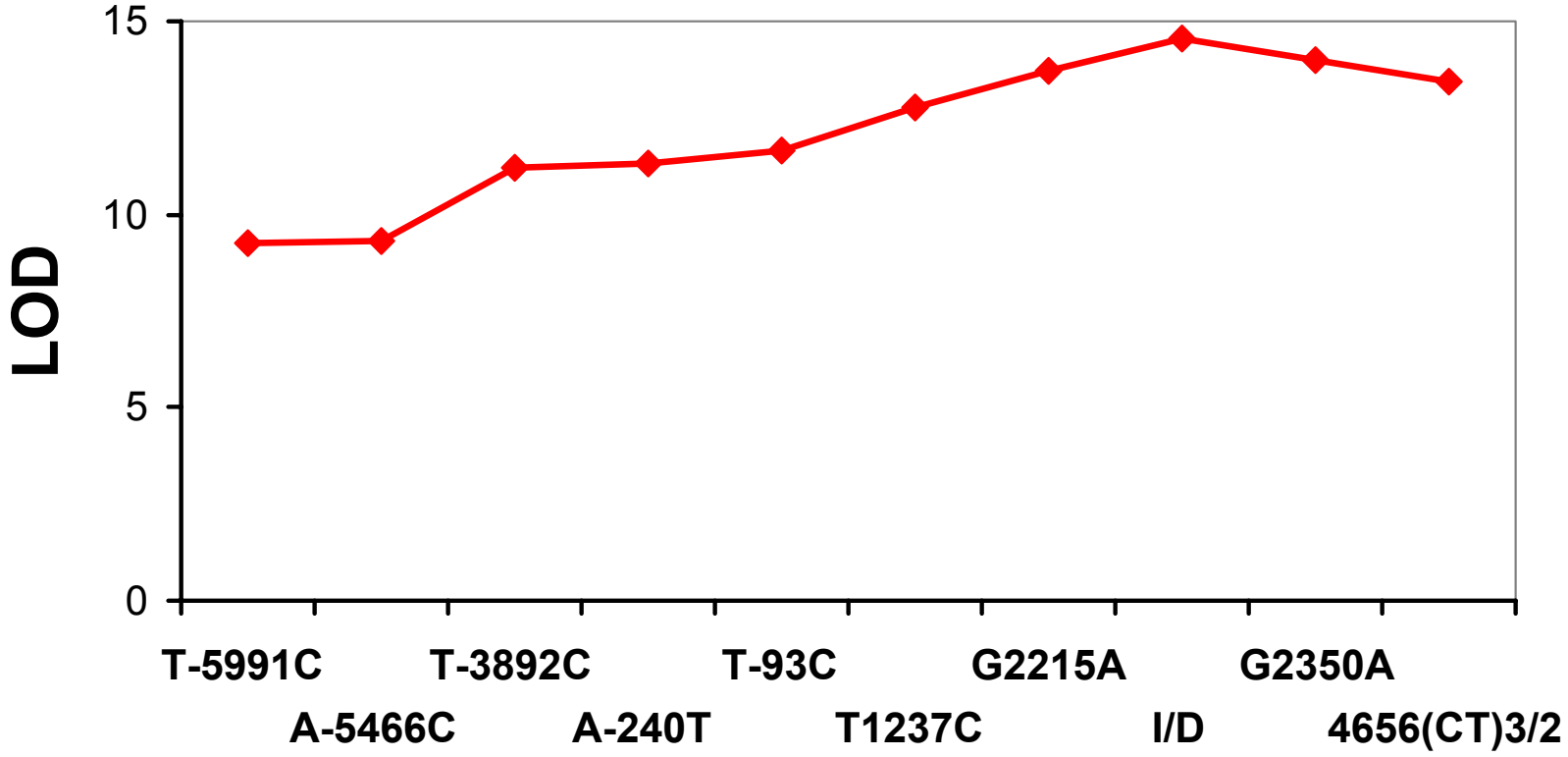


Evidence for Linkage



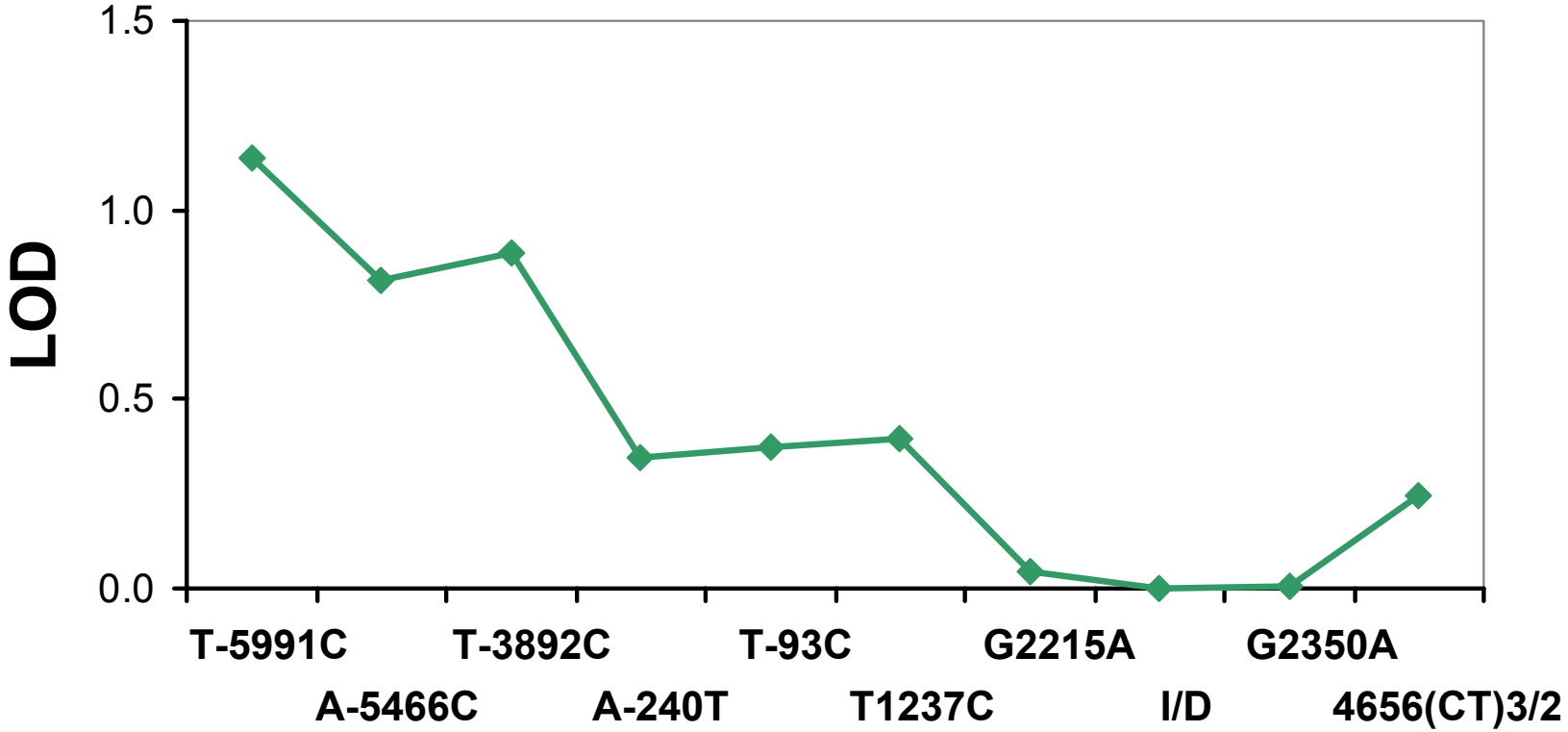
$$H_0 : (\mu, \sigma_g^2, \sigma_e^2) \qquad H_1 : (\mu, \sigma_g^2, \sigma_e^2, \sigma_a^2)$$

Evidence for Association



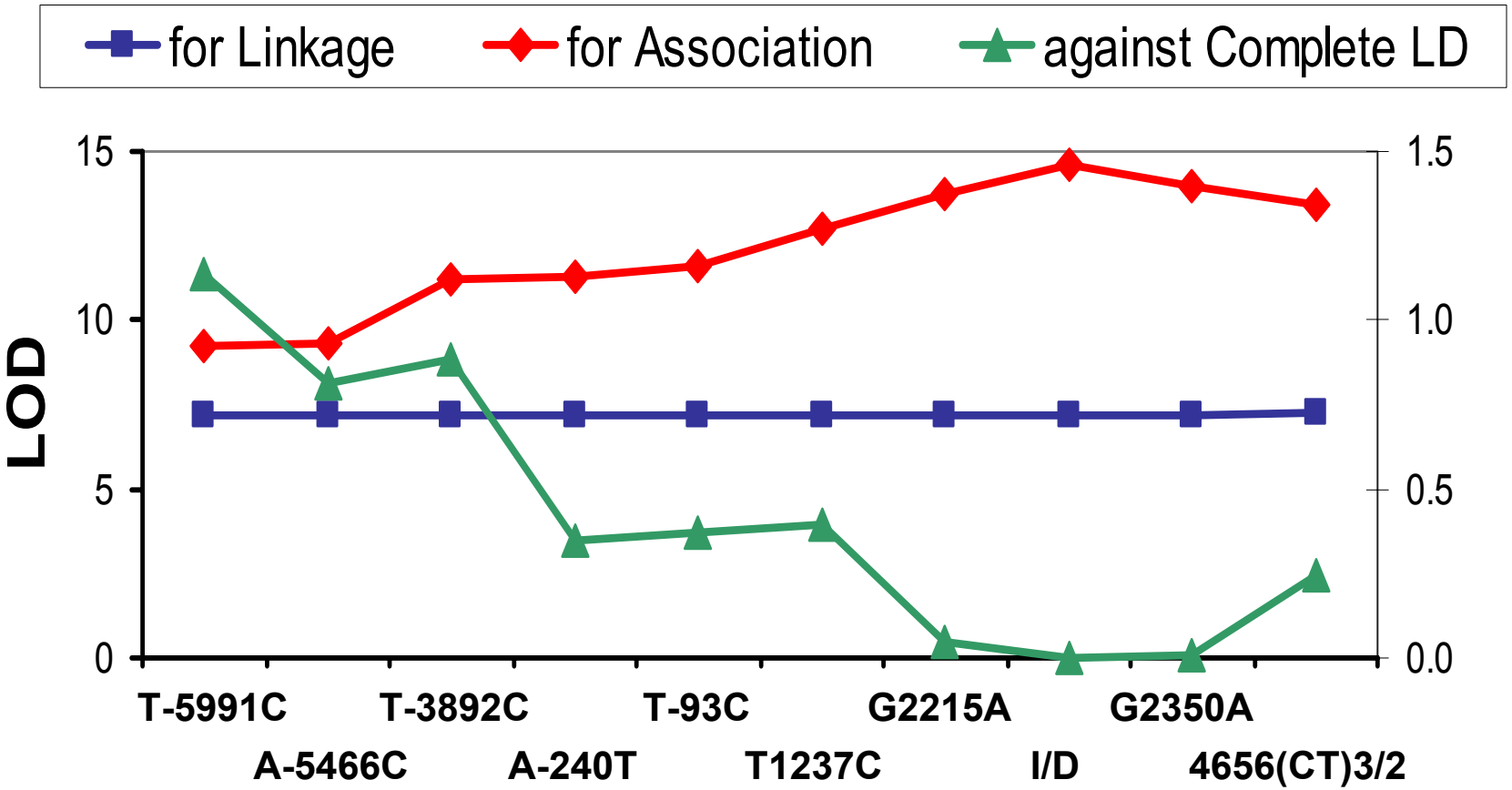
$$H_0 : (\mu, \sigma_g^2, \sigma_a^2, \sigma_e^2, \beta_b) \qquad H_1 : (\mu, \sigma_g^2, \sigma_a^2, \sigma_e^2, \beta_b, \beta_w)$$

Evidence Against Complete LD



$$H_0 : (\mu, \beta_b, \beta_w, \sigma_g^2, \sigma_e^2) \qquad H_1 : (\mu, \beta_b, \beta_w, \sigma_g^2, \sigma_e^2, \sigma_a^2)$$

Drawing Conclusions



Parameter Estimates

- Estimates
  - Total linkage (σ^2_a)
  - Linkage due to LD (σ^2_{a*LD})
  - Effect size at marker (β_w)
- Depend on
  - QTL allele frequencies (p,q)
  - QTL effect (a)
  - Disequilibrium (D)
  - Marker allele frequencies (r,s)

Useful diagnostics

- A bit of algebra
- Provide indicator of distance
  - Minimum D' (D'_{min})
- Select next markers
  - Range for QTL alleles (p_{min}, p_{max})

$$D'_{\min} = \sqrt{\sigma^2_{a \cdot LD} / \sigma^2_a}$$

$$p_{\min} = \frac{1}{1 + \sigma^2_a / (2\beta^2_w r^2)}$$

$$p_{\max} = 1 - \frac{1}{1 + \sigma^2_a / (2\beta^2_w s^2)}$$

Application to T-5991C

- LOD scores
  - ~7 linkage
  - ~9 association
  - ~1 linkage minus association
- Trait locus predictions
  - In greater than 78% disequilibrium
  - Minor allele frequency between .15 and .48
- Compare to I/D and neighbors

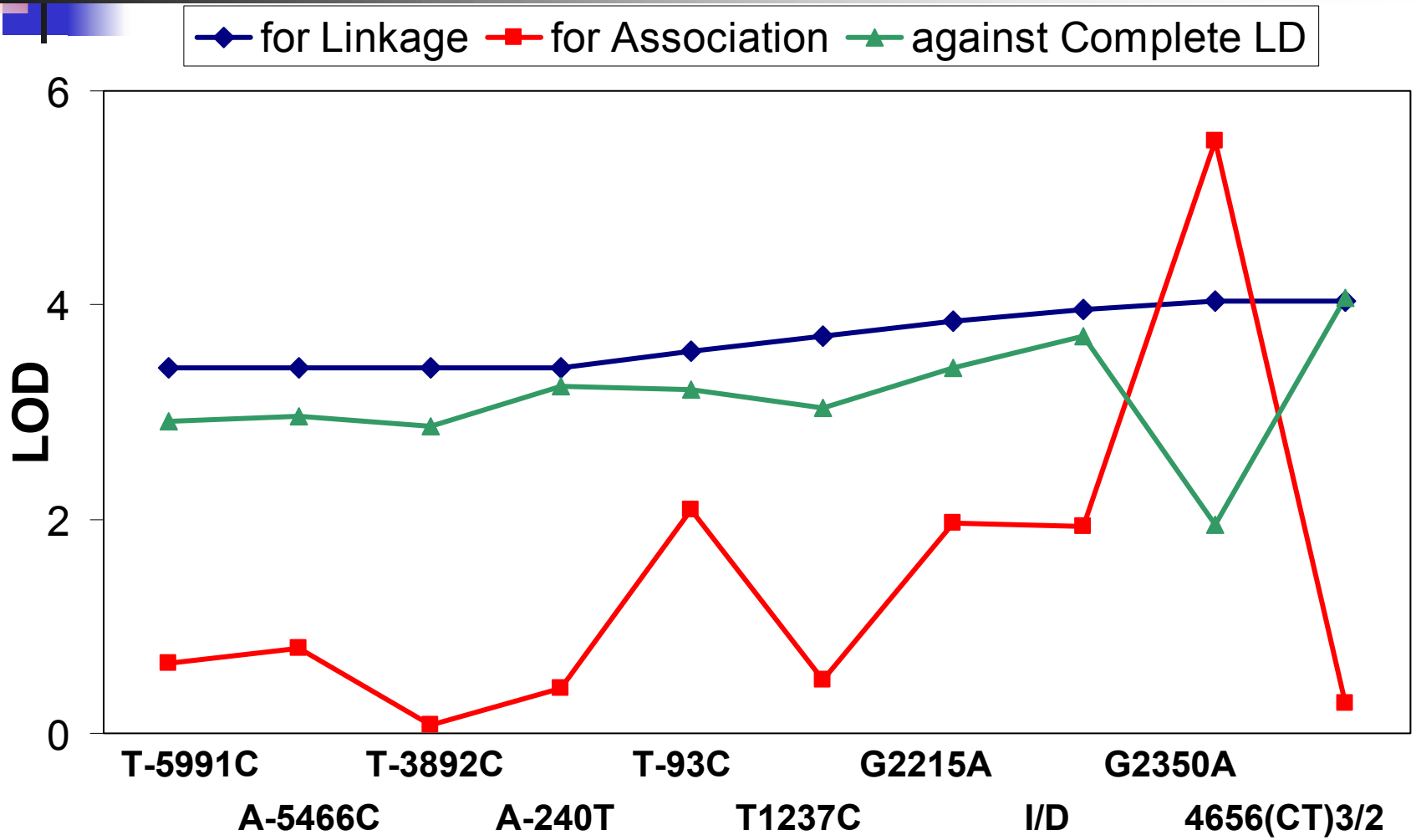
ACE: D'_{min}, p_{min} and p_{max}

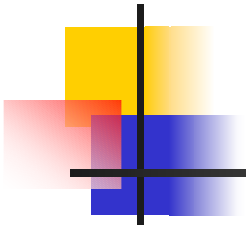
	Expected	Actual		
	T-5991C	**G2215A**	**I/D**	**G2350A**
D'	> 0.78	0.78	0.82	0.85
Minor allele	.15–.48	.45–.50	.45–.50	.45–.50

Finer mapping

- 3 mutations (inc. I/D) explain all linkage
  - UK population
- How to identify causal variant?
  - Population with more haplotype diversity
  - Jamaican sample
    - Colin McKenzie, University of West Indies, Jamaica
    - Same di-allelic polymorphisms

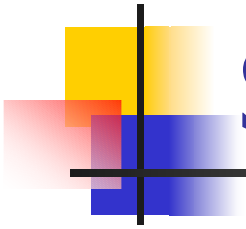
Jamaican Population Summary





Example Summary

- Agrees with haplotype analysis
- Distinguishes complete and incomplete disequilibrium
  - Measure of distance for incomplete LD
  - Indicator of trait allele frequencies
- Typical or fairy-tale?



Study design

- What markers?
  - Effect of disequilibrium
  - Effect of allele frequencies
- Family sample
  - What size families?
  - Parents or no parents?
- Effect of phenotypic selection

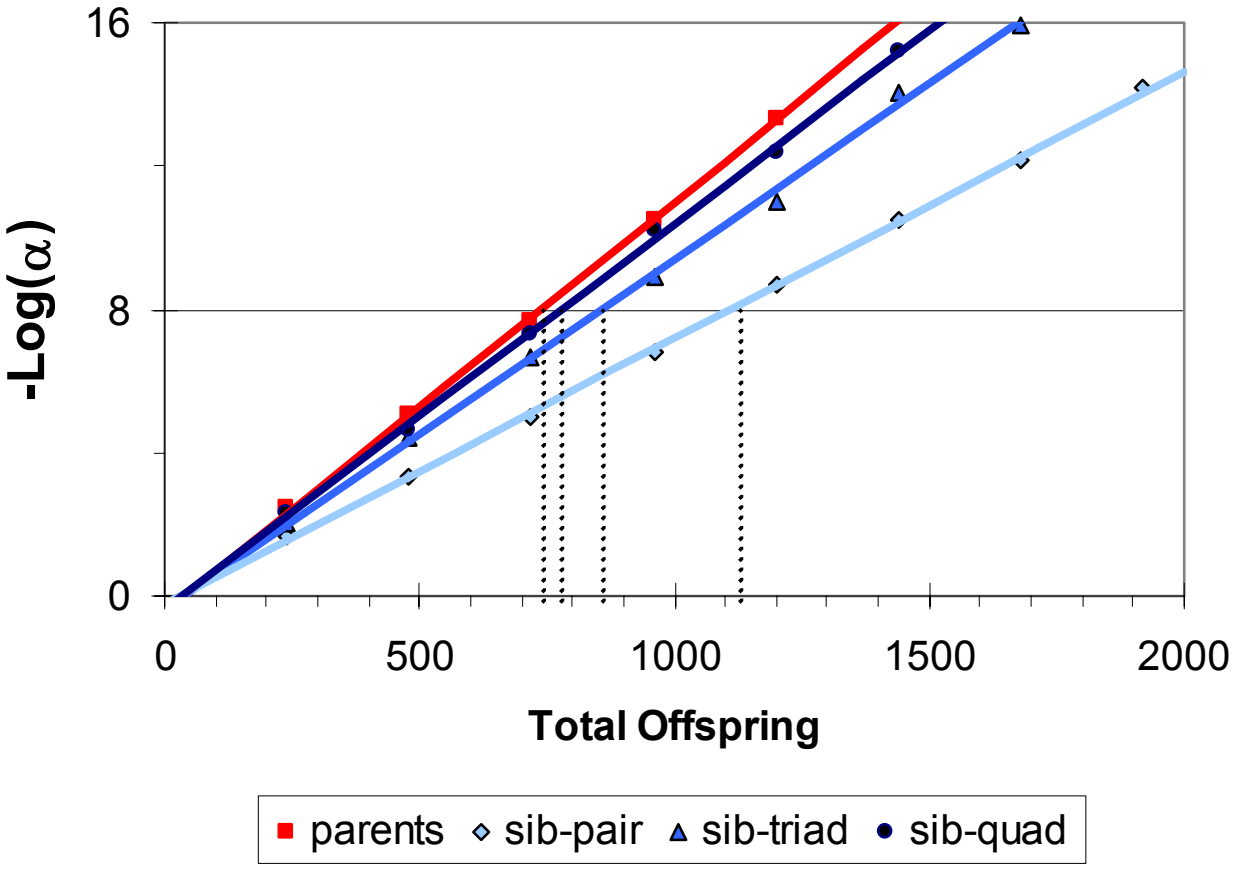
Sensitivity to Disequilibrium

	Amount of Disequilibrium				
	0%	**25%**	**50%**	**75%**	**100%**
480 triads	0	2	20	70	97
240 sib-pairs	0	2	23	73	98
120 sib-quads	0	3	27	76	98
Estimate of *a*	0	1.1	2.2	3.4	4.5

Power for α=0.001, h^2 = .1, s^2 = .3, θ = 0.

Average additive genetic value estimated at the marker.

Effect of Family Structure

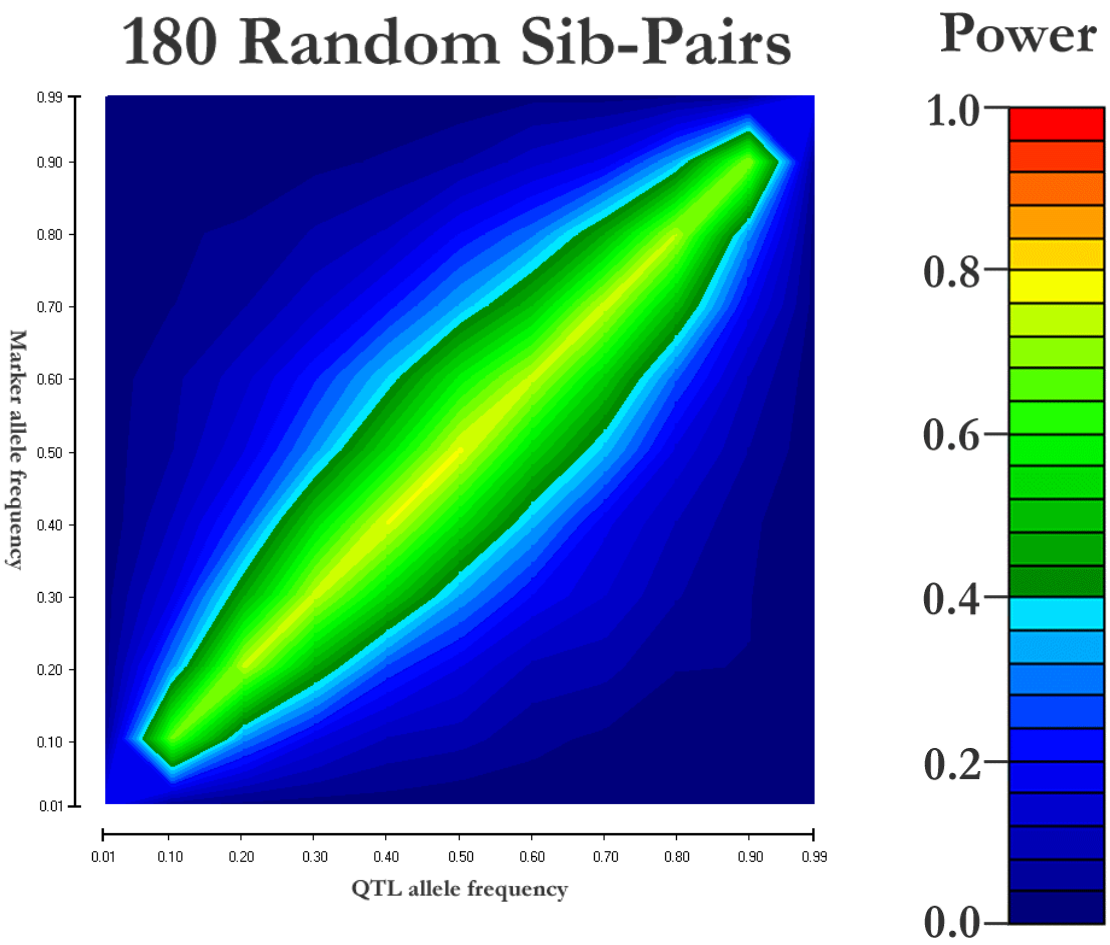


Trios For Genome-Wide Scan

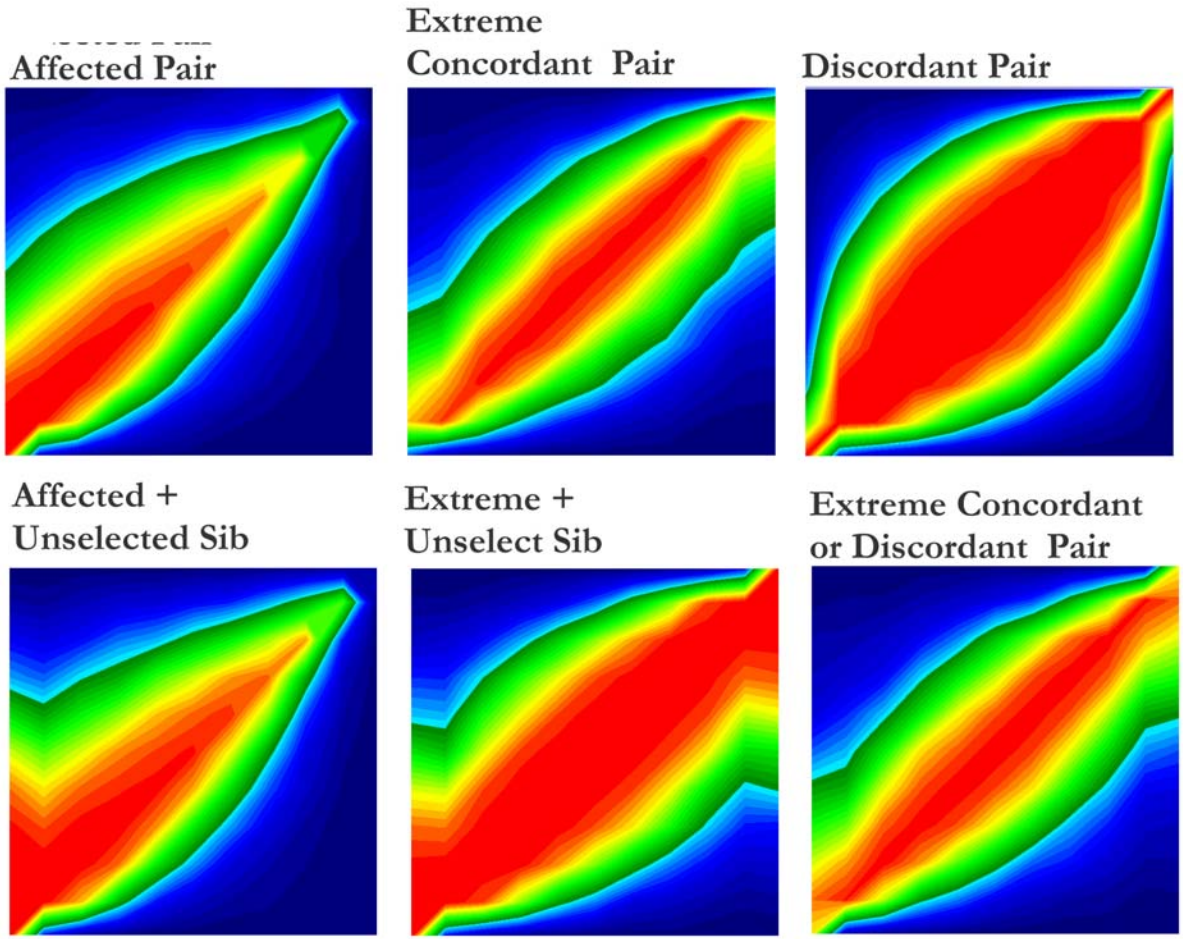
Disease Allele Frequency	Marker Allele Frequency				
	0.1	**0.3**	**0.5**	**0.7**	**0.9**
0.1	**248**	626	1306	2893	10830
0.3	1018	**238**	466	996	3651
0.5	2874	702	**267**	556	2002
0.7	9169	2299	925	**337**	1187
0.9	73783	18908	7933	3229	**616**

$\lambda s = 1.5$, $\alpha = 5 \times 10^{-8}$, Spielman TDT
(Müller-Myhsok and Abel, 1997)

Effect of Allele Frequencies



Effect of Selection



References

Fulker et al (1999) *Am J Hum Genet* **66**:259-267

Abecasis et al (2000) *Am J Hum Genet* **66**:279-292

Abecasis et al (2000) *Eur J Hum Genet* **8**:545-551

Cardon and Abecasis (2000) *Behav Genet* **30:**235-243

McKenzie et al (2001) *Hum Mol Genet* **10**:1077-84

Abecasis et al (2001) *Am J Hum Genet* **68:**1463-74

www.sph.umich.edu/csg/abecasis