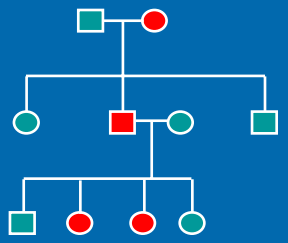
Family Based Association Studies

Gonçalo Abecasis
University of Michigan

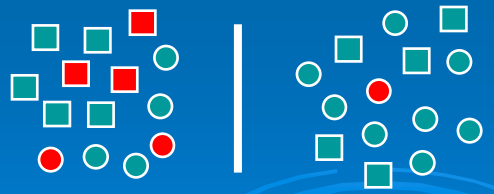
Design of Gene Mapping Studies

How to find disease susceptibility alleles?



Rare, high penetrance mutations – use linkage

Common, low penetrance variants – use association



Magnitude of effect

Frequency in population

Mark Daly

Genetic Linkage Studies

- Search for long stretches of chromosome shared between close relatives with similar phenotype
- Identify variants with relatively large contributions to disease risk
- Require only a coarse measurement of genetic variation
  - 400 – 800 microsatellites or a few thousand SNPs extract most of the linkage information in typical pedigrees
- High-throughput SNP genotyping has already sped up and facilitated these studies
  - Data analysis methods must select subset of independent SNPs or model disequilibrium between markers

Genetic Association Studies

- Search for short stretches of chromosome shared between distantly related individuals with similar phenotype
- Identify genetic variants with relatively small individual contributions to disease risk
- Require detailed measurement of genetic variation
  - >8,000,000 catalogued genetic variants, so …
  - Until recently, limited to candidate genes or regions
    - A hit-and-miss approach…
- SNP resources and decreasing assay costs now make it possible to conduct comprehensive genome-wide scans

Association Studies in Families

- Majority of association studies genotype unrelated individuals and nearly all linkage studies genotype related individuals…
- However, tests for genetic association can use family data when relatedness between individuals is modeled appropriately (e.g. George and Elston, 1987)

Linear Model for Association

- Model association using a model such as:

$$E(y_i) = \mu + \beta_g g + \beta_c c + \ldots$$

- y_i is the phenotype for individual i
- g_i is the phenotype for individual i
  - Simplest coding is to set g_i = number of copies of allele *'1'*
- c_i is a covariate for individual i
  - Covariates could be estimated ancestry, environmental factors…

- β coefficients are estimated effects of genotype, covariates

Allowing for Related Data

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

$$\Omega_{ij} = \begin{cases} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & i = j \\ 2\varphi_{marker(ij)}\sigma_a^2 + 2\varphi_{ij}\sigma_g^2 & i \neq j \end{cases}$$

Where,

φ_{ij} is the kinship coefficient for the individuals i and j

$\varphi_{marker(ij)}$ depends on the number of alleles shared IBD

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

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

Nuclear Families

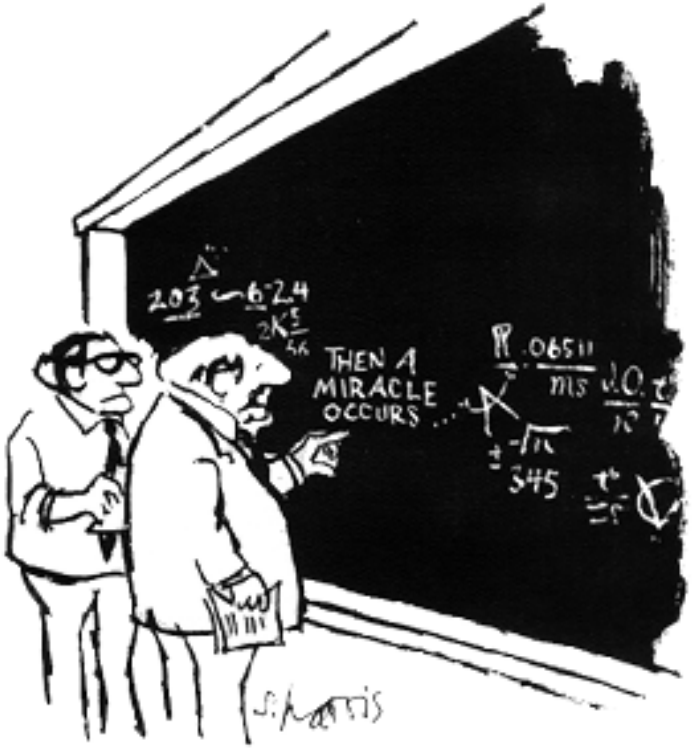
$$b_{ij} = b_i = \begin{cases} \dfrac{g_{iF} + g_{iM}}{2} & \text{average of parental genotypes} \\ \displaystyle\sum_{k}^{sibship} \dfrac{g_{ik}}{n_{sibs}} & \text{average of sibling genotypes} \end{cases}$$

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

Extended Families

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

Parameter Derivations



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

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

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

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

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

QTDT Practical

- We'll use QTDT to:
  - Evaluate whether a trait is heritable
  - Evaluate evidence for linkage
  - Evaluate evidence for association
- We'll use the Bernard Keavney's classic ACE data as an example…

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
  - goncaloa\keavney

Variance Components (QTDT)

- Allows customized variance-covariance matrices
  - Key options are **–w and –v**
  - Describe two alternative models for variances
- Can build means model for association tests
  - The **–a** option and **–m** options
  - Using observed marker genotypes
- Reads in IBD matrix generated by Merlin

Common Variance Components in QTDT

Value	Description
e	**Non-shared Environment.** Environmental effects that are unique to each family member and measurement error.
g	**Polygenic.** These effects are a function of relatedness between family members and may be due to polygenes.
a	**Additive Major Gene Effect.** This represents the additive effect of linkage to a major gene. The pi-hat component.
t	**Twin Environment.** This represents the environment shared by twins, but not other types of relatives.
c	**Common Environment.** This represents the environment shared by all relatives.

The **–w** option specifies variances under the null. E.g. **–we** switch specifies that only environmental effects should be modeled. The **–v** switch specifies variances under the alternative. E.g. **–veg** models environmental and polygenic effects.

For other options and components, see documentation.

Testing Heritability With QTDT

- Disable association models
  - Use option `-a-`
- Specify two models for variances
  - Use options `-we` and `-veg`
- Typical command line:

```
qtdt -d keavney.dat -p keavney.ped
     -a- -we -veg
```

Typical Output: Testing Heritability

- Summary output appears on screen

```
The following models will be evaluated...
  NULL MODEL
     Means = Mu
 Variances = Ve
  FULL MODEL
     Means = Mu
 Variances = Ve + Vg

Testing trait:                                    ACE
 Allele    df(0) -LnLk(0)   df(V) -LnLk(V)   ChiSq      p
    N/A      403   573.67     402   544.77   57.80 3e-14 (405 probands)
```

- In this case the trait is highly heritable ($p < 10^{-13}$)
- Additional information, including variance component estimates output to "regress.tbl"

Estimating IBD Coefficients

- QTDT has very limited IBD engine
  - More sophisticated estimates from Merlin are better
- Relevant Merlin options
  - Option `--ibd` generates *merlin.ibd* output file
  - Option `--markerNames` to include marker labels
- Use the following Merlin command line

```
merlin --ibd --markerNames
   -d keavney.dat -p keavney.ped -m keavney.map
```

Testing Linkage With QTDT

- Disable association models
  - Use option `-a-`
- Specify two models for variances
  - Use options `-we` and `-veg`
- Provide an IBD file
  - Use option `-i ibdfile`
- Typical command line:

```
qtdt -d keavney.dat -p keavney.ped -i keavney.ibd
     -a- -weg -vega
```

Typical Output: Testing Linkage

The following models will be evaluated...

```
  NULL MODEL
     Means = Mu
 Variances = Ve + Vg

  FULL MODEL
     Means = Mu
 Variances = Ve + Vg + Va
```

Testing trait: ACE

Testing marker: T-5491C

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
All	402	544.77	401	528.22	33.09	9e-09	(405 probands)

Testing marker: A-5466C

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
All	402	544.77	401	528.22	33.09	9e-09	(405 probands)

(… output continues at each marker …)
(… additional information in regress.tbl file …)

Simple Association Model

- Each copy of allele changes trait by a fixed amount
  - Use covariate counting copies for allele of interest
- Results in estimate of additive genetic value
  - Evidence for association when a ≠ 0

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

$$\mathrm{E}(\mathrm{y_i}) = \mu + \beta_X X_i$$

X is the number of copies for allele of interest.
β_x is the estimated effect of each copy (the additive genetic value).

Using QTDT to Test Association

- Select association model
  - If population is homogenous, use option `-at`
- Specify model for variances
  - Use option `-we` if sample consists of unrelated individuals
  - Use option `-weg` if sample consists of trios only
  - Use option `-wega` under the null if families in general
- Provide an IBD file
- Typical command line:

```
qtdt -d keavney.dat -p keavney.ped -i merlin.ibd
     -at -wega
```

Typical Output: Testing Association

The following models will be evaluated...

```
  NULL MODEL
     Means = Mu
 Variances = Ve + Vg + Va

  FULL MODEL
     Means = Mu + X
 Variances = Ve + Vg + Va
```

Testing trait: ACE

Testing marker: T-5491C

Allele	df(0)	-LnLk(0)	df(X)	-LnLk(X)	ChiSq	p	
1	382	503.12	381	471.97	62.29	3e-015	(386 probands)
2	382	503.12	381	471.97	62.29	3e-015	(386 probands)

Testing marker: A-5466C

Allele	df(0)	-LnLk(0)	df(X)	-LnLk(X)	ChiSq	p	
1	380	503.11	379	471.07	64.09	1e-015	(384 probands)
2	380	503.11	379	471.07	64.09	1e-015	(384 probands)

(… output continues at each marker …)
(… additional information in regress.tbl file …)

Using QTDT to Test Association II

- Select association model
  - For the between-within model, use option `-ao`
- Specify model for variances
  - Use option `-we` if sample consists of unrelated individuals
  - Use option `-weg` if sample consists of trios only
  - Use option `-wega` under the null if families in general
- Provide an IBD file
- Typical command line:

```
qtdt -d keavney.dat -p keavney.ped -i merlin.ibd
     -ao -wega
```

Typical Output: Between-Within Association Model

The following models will be evaluated...

```
  NULL MODEL
     Means = Mu + B
 Variances = Ve + Vg + Va

  FULL MODEL
     Means = Mu + B + W
 Variances = Ve + Vg + Va
```

Testing trait: ACE

Testing marker: T-5491C

Allele	df(0)	-LnLk(0)	df(T)	-LnLk(T)	ChiSq	p	
1	381	490.93	380	470.41	41.04	1e-010	(180/386 probands)
2	381	490.93	380	470.41	41.04	1e-010	(180/386 probands)

Testing marker: A-5466C

Allele	df(0)	-LnLk(0)	df(T)	-LnLk(T)	ChiSq	p	
1	379	489.59	378	470.22	38.73	5e-010	(186/384 probands)
2	379	489.59	378	470.22	38.73	5e-010	(186/384 probands)

(… output continues at each marker …)
(… additional information in regress.tbl file …)

Can association explain linkage?

- If linkage and association signals are present…
  - Using the previous two tests
- Decide whether locus has been mapped
  - Are there other associated alleles to be found?
- Can the small region where LD extends …
- … account for the covariance between relatives who share the surrounding stretch of DNA?

Using QTDT to Test If Association Can Explain Linkage

- Select an association model
  - For example, use option `-at`
- Specify models for variances as in linkage test
  - Use option `-weg` under the null
  - Use option `-vega` under the alternative
- Provide an IBD file
- Typical command line:

```
qtdt -d keavney.dat -p keavney.ped -i merlin.ibd
     -at -weg -vega
```

Typical Output: Linkage after modeling association

The following models will be evaluated...

```
  NULL MODEL
     Means = Mu + X
 Variances = Ve + Vg

  FULL MODEL
     Means = Mu + X
 Variances = Ve + Vg + Va
```

Testing trait: ACE

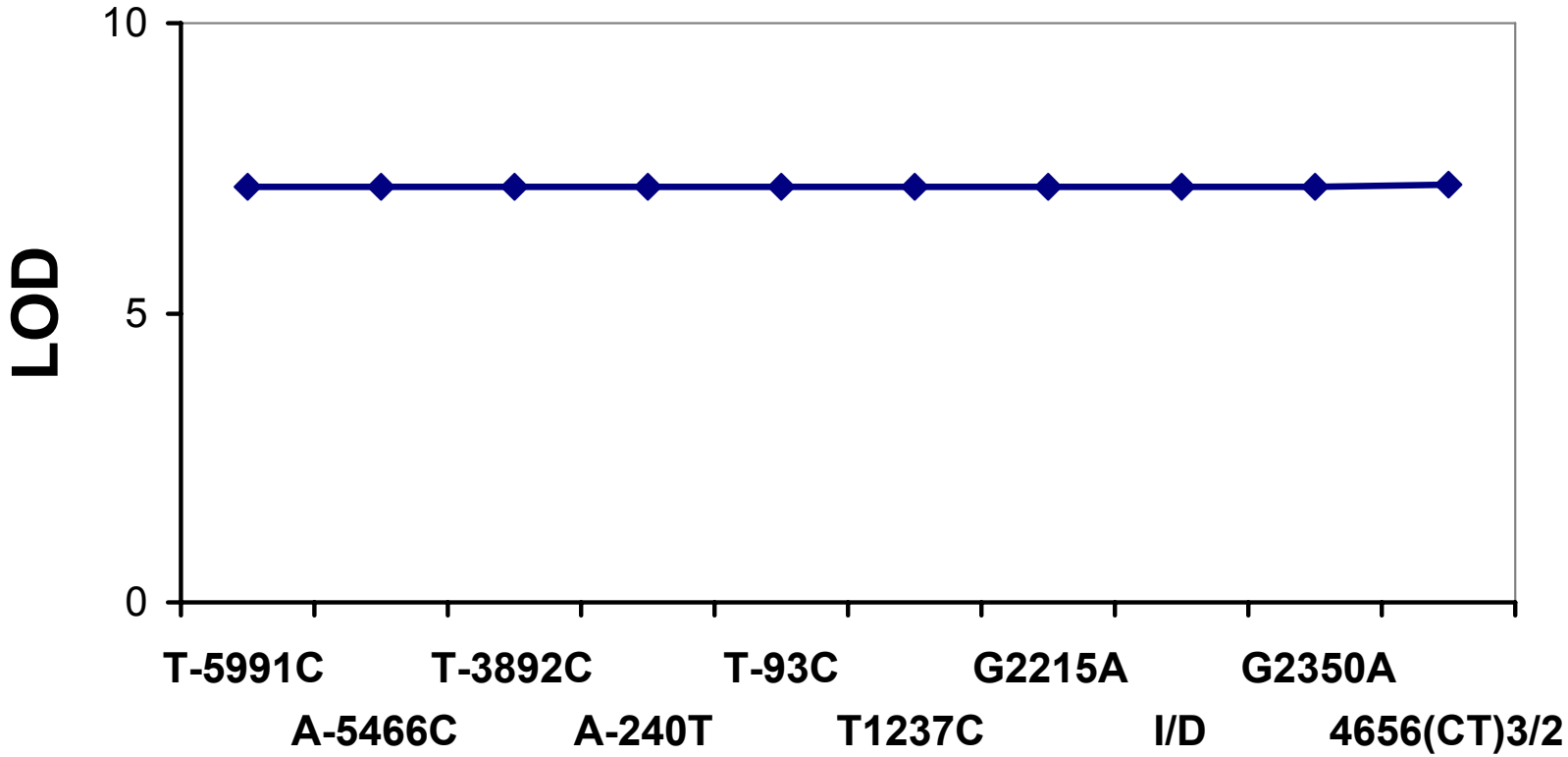
Testing marker: T-5491C

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
1	391	486.44	390	481.40	10.08	0.0015	(395 probands)
2	391	486.44	390	481.40	10.08	0.0015	(395 probands)

Testing marker: A-5466C

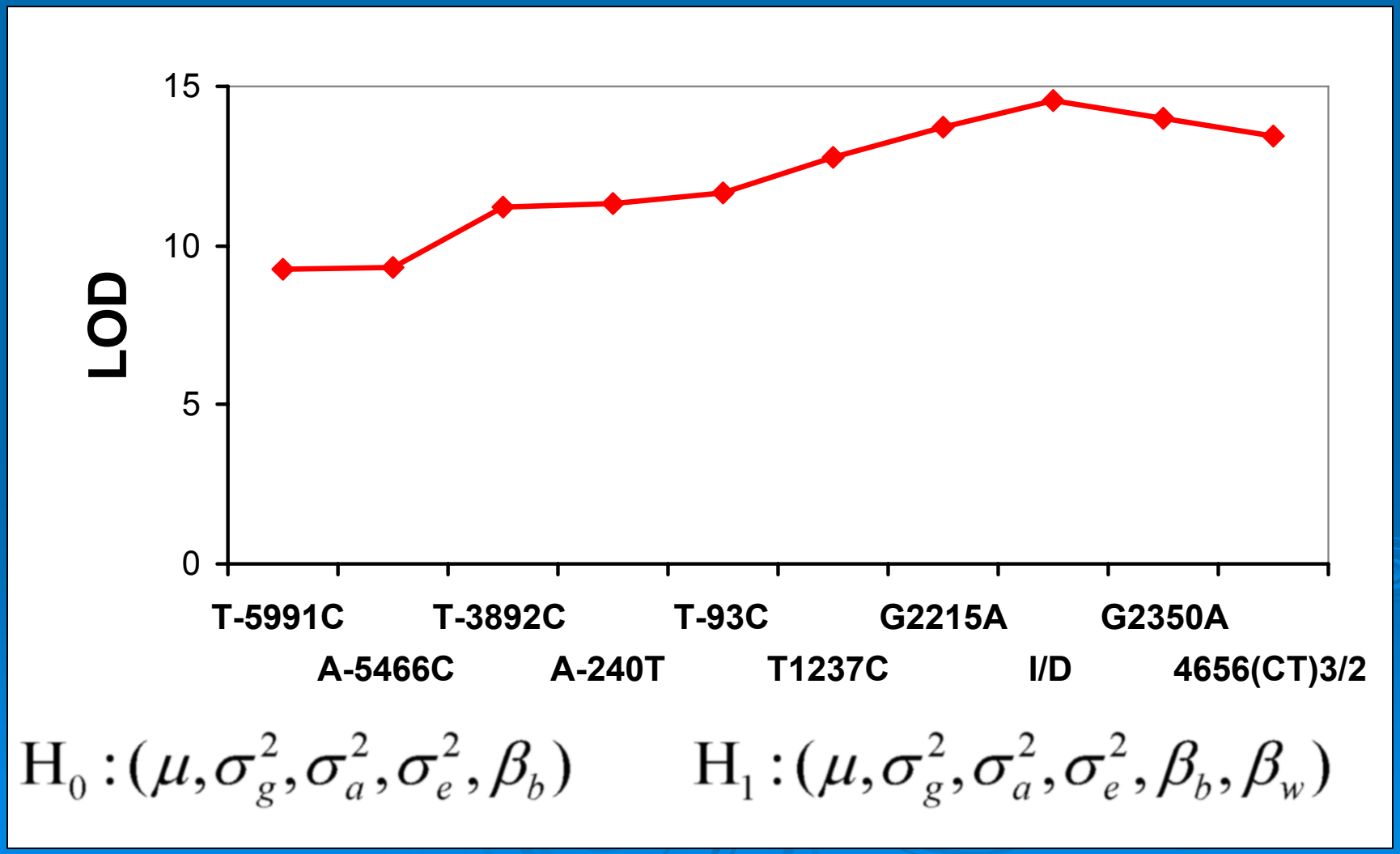
Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
1	388	483.20	387	479.50	7.41	0.0065	(392 probands)
2	388	483.20	387	479.50	7.41	0.0065	(392 probands)

Evidence for Linkage



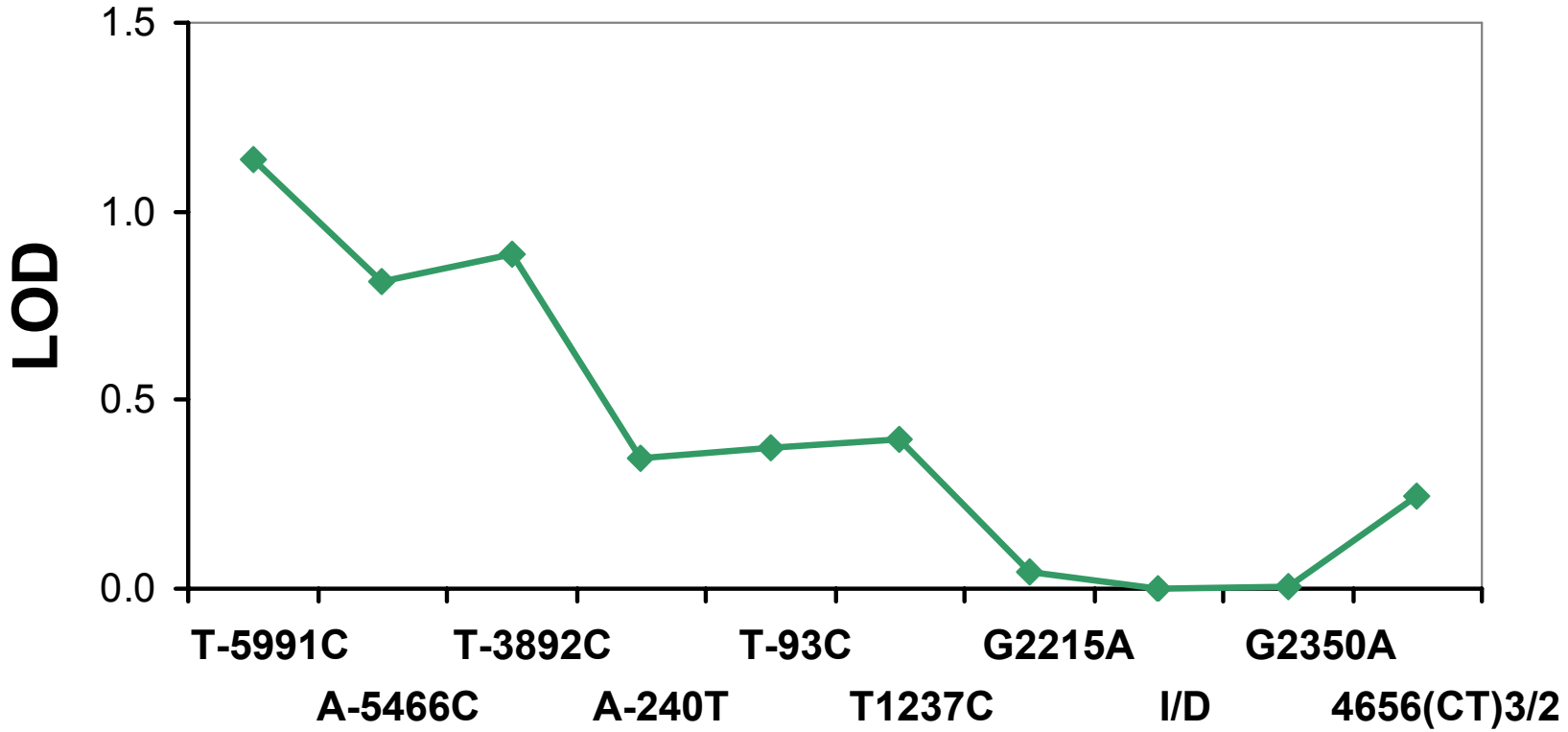
$$H_0 : (\mu, \sigma_g^2, \sigma_e^2) \qquad H_1 : (\mu, \sigma_g^2, \sigma_e^2, \sigma_a^2)$$

Evidence for Association



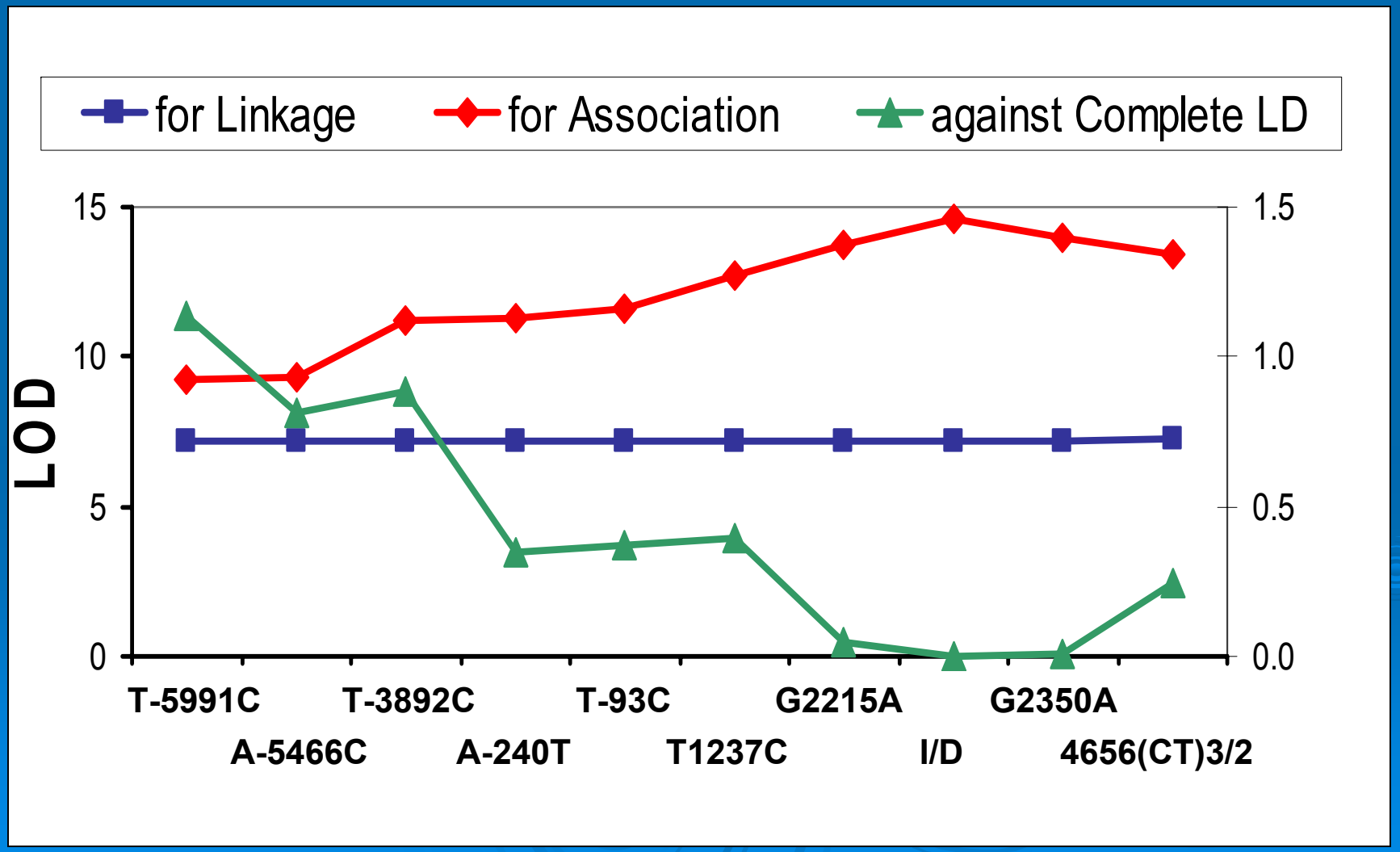
$$H_0 : (\mu, \sigma_g^2, \sigma_a^2, \sigma_e^2, \beta_b) \qquad H_1 : (\mu, \sigma_g^2, \sigma_a^2, \sigma_e^2, \beta_b, \beta_w)$$

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



References

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

Neale et al (1999) *Behav Genet* **29**: 233—244

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

Sham et al (2000) *Am J Hum Genet* ***66:****1616-1630*

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

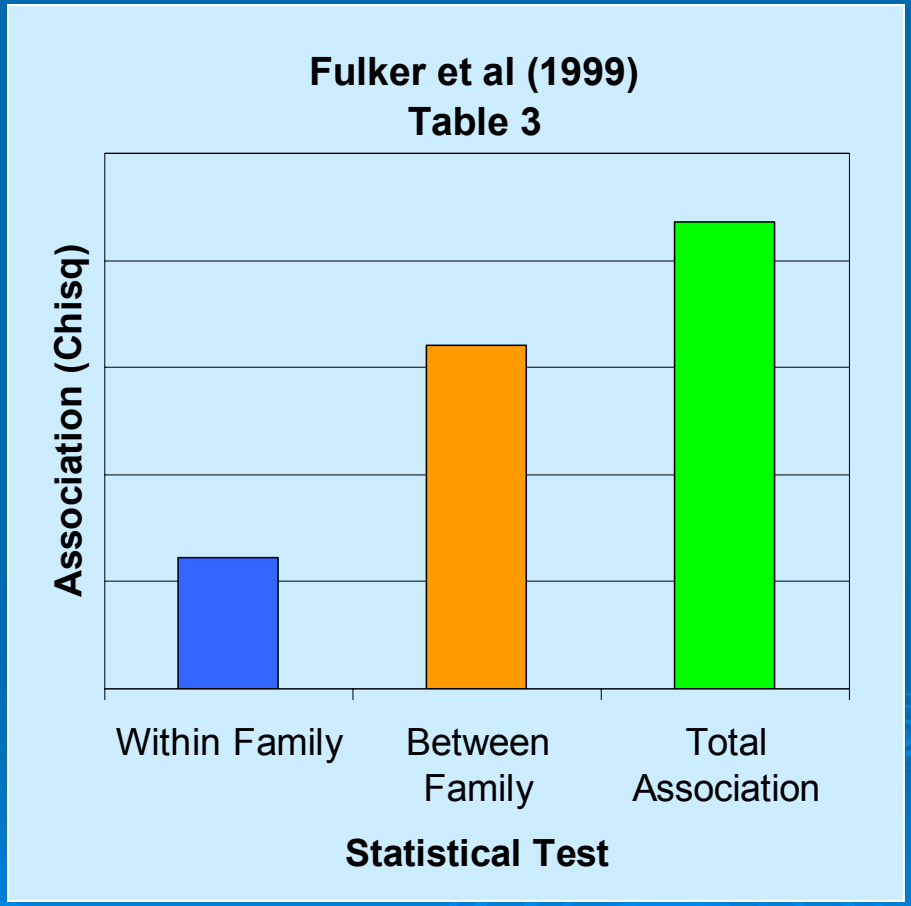
Cardon and Abecasis (2000) *Behav Genet* **30:**235-243

Family Based Association Studies Have Been Getting a Bad Rap…

- The most popular methods for association mapping in families rely on transmission disequilibrium
  - TDT, sib-TDT, QTDT, PDT, ...
- Out of an abundance of caution, TDT-like tests use very conservative approach to guard against population structure
  - Reduced power on per genotype basis
- Given thousands of genotyped markers, better methods for guarding against stratification are available
  - Genomic Control (Devlin and Roeder, 1999)
  - Structured Association Mapping (Pritchard et al, 2000)
  - Avoid heavily stratified samples

Powerful Association Tests in Families

- Do not focus on within family association (e.g. transmission disequilibrium)
- Evaluate the effect of each genotype, adjusting for familial correlations (as in George and Elston 1987)
- Fulker et al (1999) showed that focusing on alleles transmitted from heterozygous parents discards 50-75% of information in a family sample

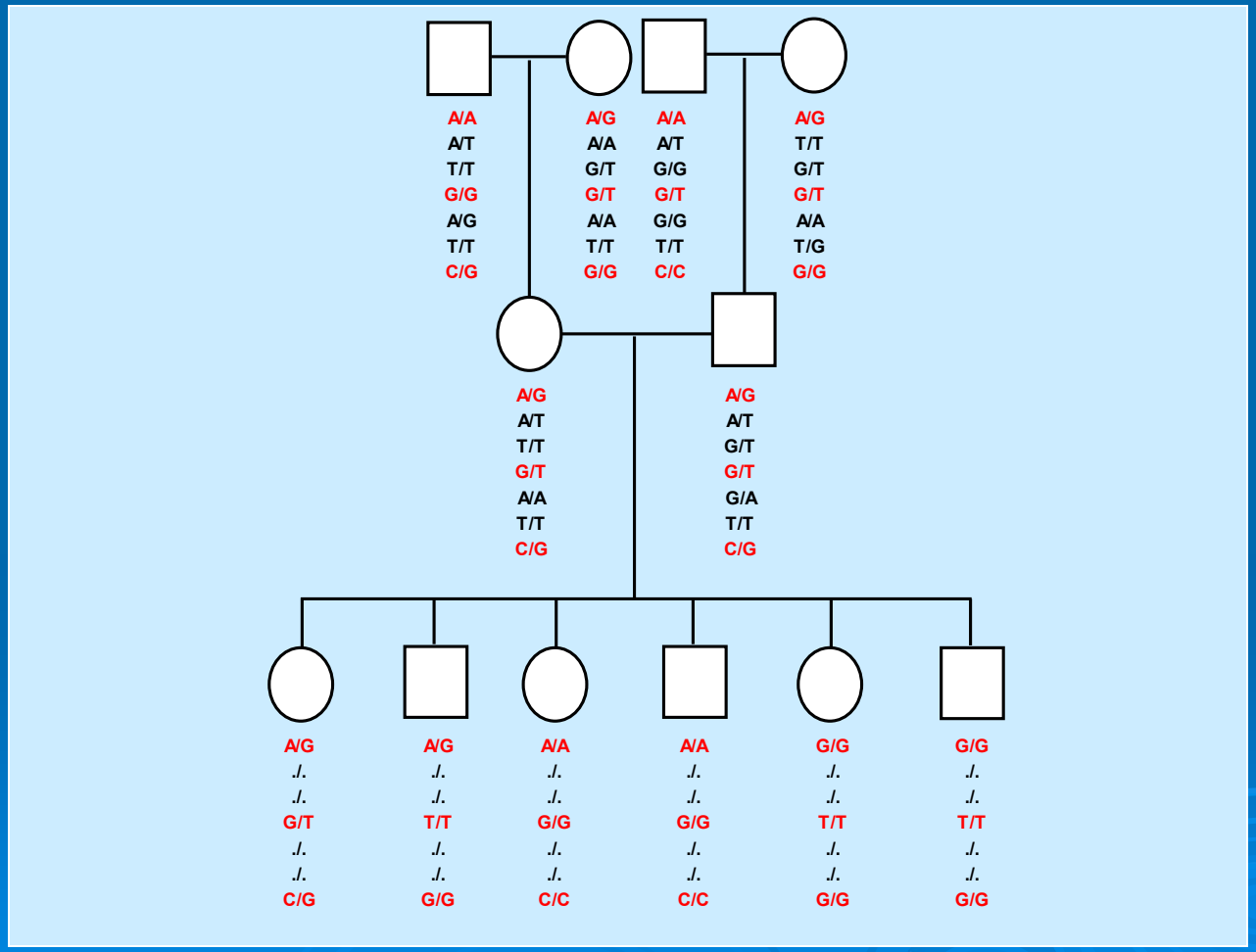


Transmission disequilibrium tests focus on within family component of association

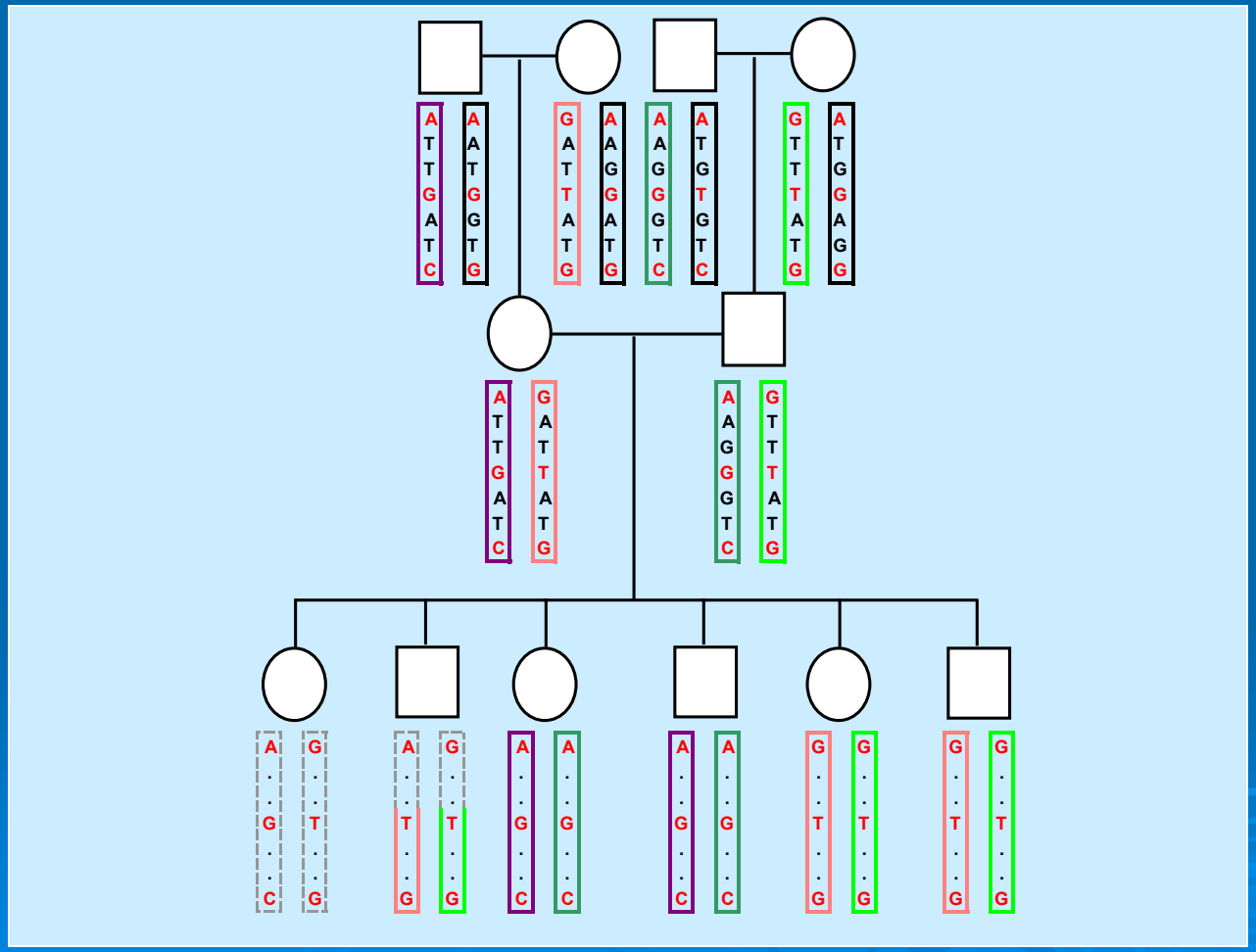
Incorporating Family Information in Genome Wide Studies

- Family members will share large segments of chromosomes
- If we genotype many related individuals, we will effectively be genotyping a few chromosomes many times
- In fact, we can:
  - genotype a few markers on all individuals
  - use high-density panel to genotype a few individuals
  - infer shared segments and then estimate the missing genotypes

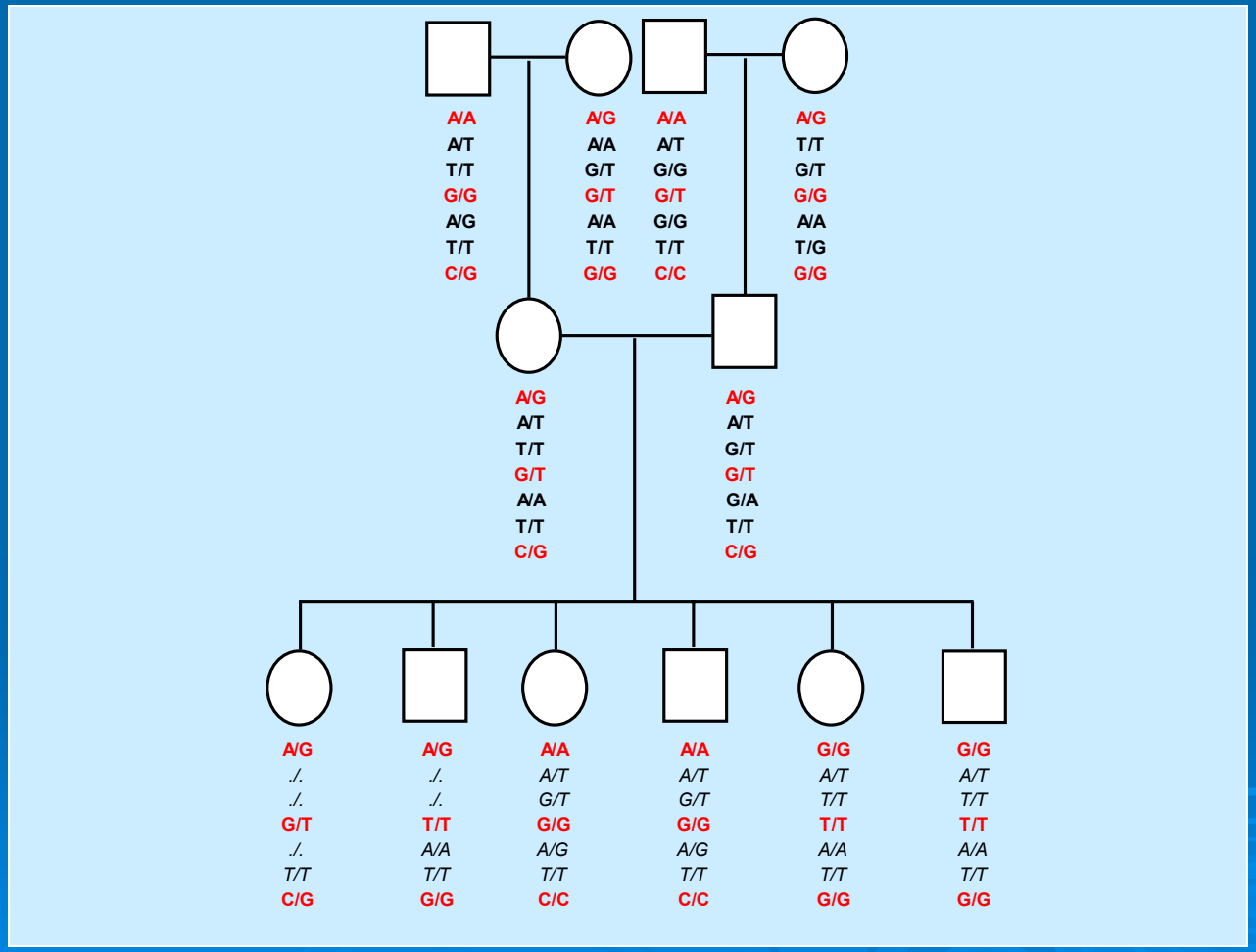
Genotype Inference
Part 1 – Observed Genotype Data



Genotype Inference
Part 2 – Inferring Allele Sharing



Genotype Inference
Part 3 – Imputing Missing Genotypes



Our Approach

- Consider full set of observed genotypes G
- Evaluate pedigree likelihood L for each possible value of each missing genotype g_{ij}
- Posterior probability for each missing genotype

$$P(g_{ij} = x \mid G) = \frac{L(G, g_{ij} = x)}{L(G)}$$

- Implemented both using Elston-Stewart (1972) and Lander-Green (1987) algorithms

Standard Linear Model for Genetic Association

- Model association using a model such as:

$$E(y_i) = \mu + \beta_g g + \beta_c c + \ldots$$

- y_i is the phenotype for individual i
- g_i is the genotype for individual i
  - Simplest coding is to set g_i = number of copies of allele *'1'*
- c_i is a covariate for individual i
  - Covariates could be estimated ancestry, environmental factors…

- β coefficients are estimated covariate, genotype effects
- Model is fitted in variance component framework

Model With Inferred Genotypes

- Replace genotype score g with its expected value:

$$E(y_i) = \mu + \beta_g \bar{g} + \beta_c c + \ldots$$

- Where

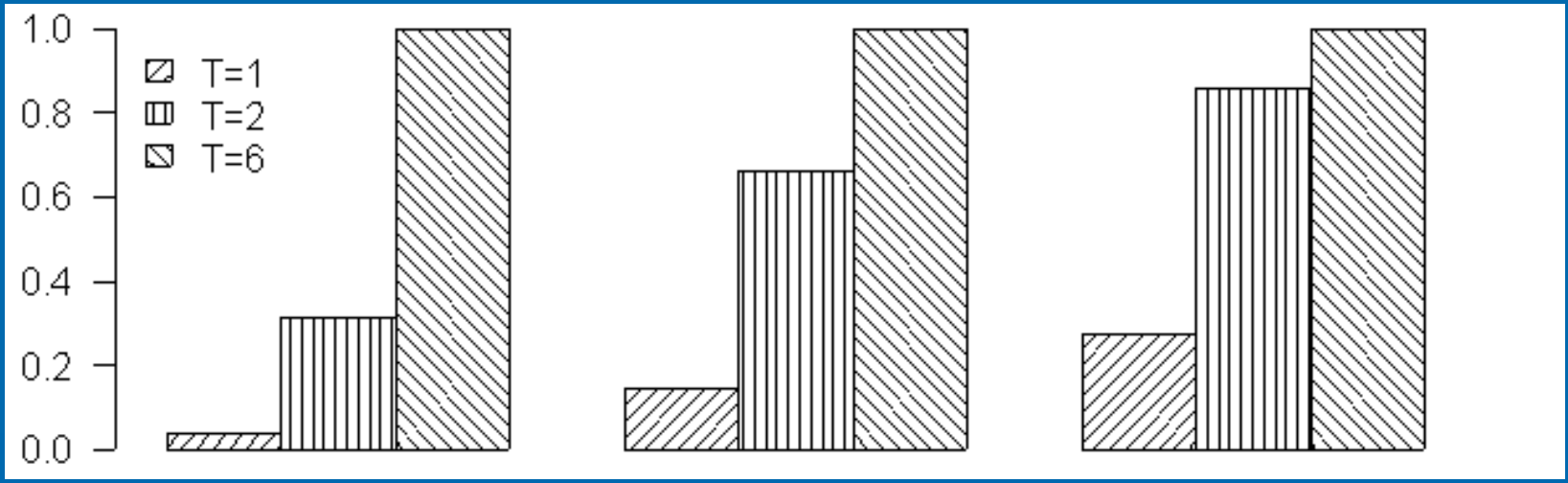
$$\bar{g}_i = 2P(g_i = 2 \mid G) + P(g_i = 1 \mid G)$$

- Association test can then be implemented as a score test or as a likelihood ratio test

- Alternatives would be to
  - (a) impute genotypes with large posterior probabilities; or
  - (b) integrate joint distribution of unobserved genotypes in family

Power in Sibships of Size 6

Without Parental Genotype Data



Analyze Observed Data

Impute when Posterior >.99

Using Expected Genotype Score

T is the number of genotyped offspring.
QTL explains 5% of variance, polygenes explain 35%,
250 sibships, $\alpha = 0.001$.

Application: Gene Expression Data

- Cheung et al (2005) carried out a genome wide association with 27 expression levels as traits
- Measured in grandparents and parents of CEPH pedigrees and took advantage of HapMap I genotypes
- TSC genotypes also available for ~6000 SNPs in the offspring of each CEPH family

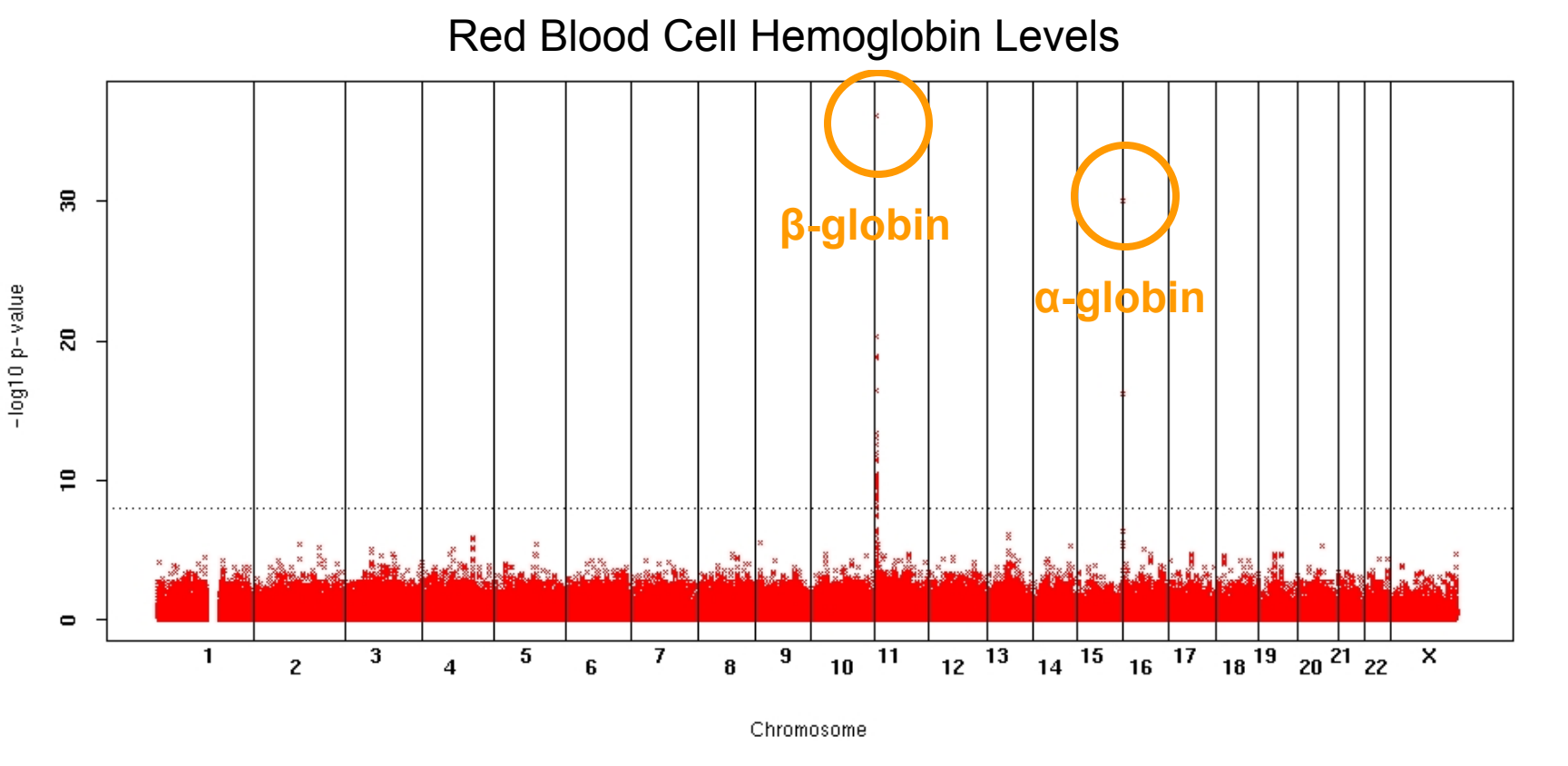
Results: Gene Expression Data

- Using observed genotypes, the most significant association mapped in *cis* for 15 of 27 traits
  - 12 of these reach $p < 5 * 10^{-8}$
- Using inferred genotypes, the most significant association mapped in *cis* for 20 of 27 traits
  - 15 of these reach $p < 5 * 10^{-8}$
  - 1 trans linkage also reaches $p < 5 * 10^{-8}$

Sardinia

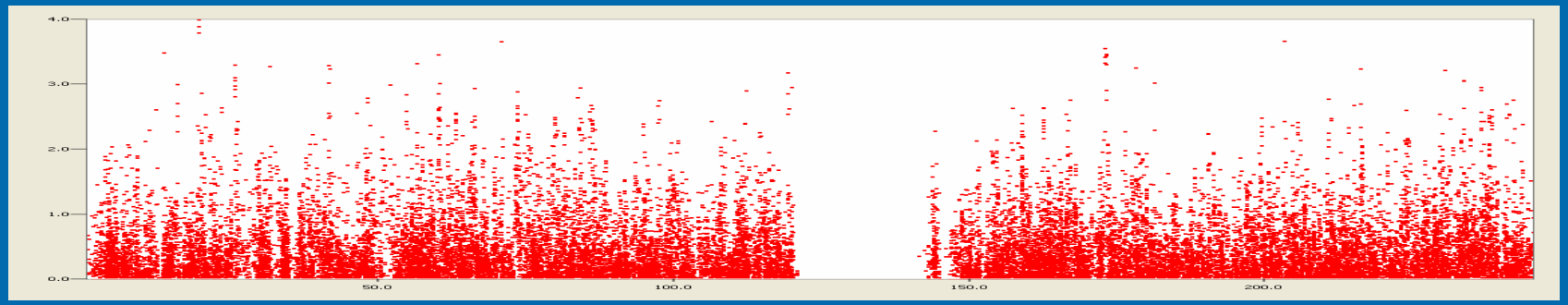
- 6,148 Sardinians from 4 towns in Ogliastra
- Measured 98 aging related quantitative traits
- Genotyping:
  - Affymetrix 10K chip in 4,500 individuals
  - Affymetrix 500K chip in 1,500 individuals
- Large pedigrees, computationally challenging
  - Preliminary results

Preliminary Results: Sardinia (after ~900 500K arrays)

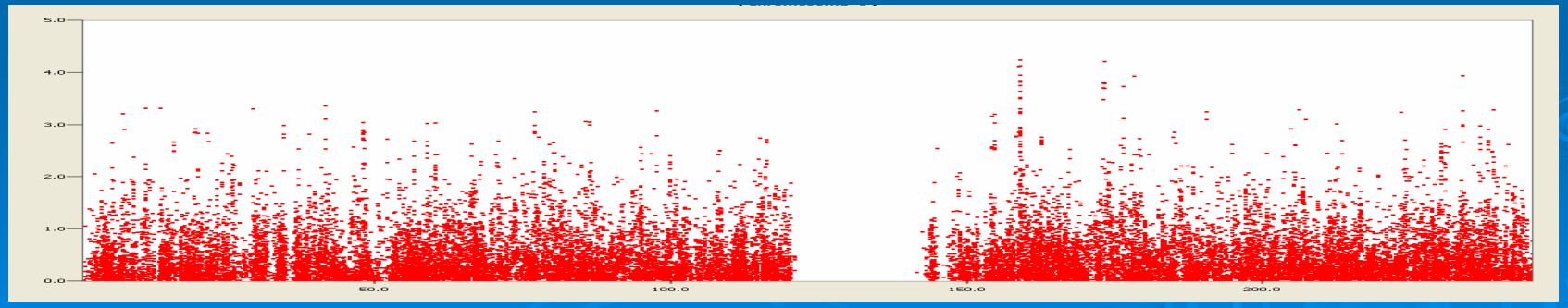


Preliminary Results from Sardinia QT interval, Chromosome 1

Before imputation



After imputation



Position (in Mb) Along Chromosome 1

Acknowledgements

- Weimin Chen, Serena Sanna
- Sardinia Investigators, led by:
  - David Schlessinger, Manuela Uda, Antonio Cao, Edward Lakatta, Paul Costa
- Gene Expression Data:
  - Vivian Cheung, Josh Burdick

goncalo@umich.edu