Practical With Merlin

Gonçalo Abecasis

MERLIN Website

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

- Reference
- FAQ
- Source
- Binaries
- Tutorial
  - Linkage
  - Haplotyping
  - Simulation
  - Error detection
  - IBD calculation
  - Association Analysis

QTL Regression Analysis

- Go to Merlin website
  - Click on tutorial (left menu)
  - Click on regression analysis (left menu)

- What we'll do:
  - Analyze a single trait
  - Evaluate family informativeness

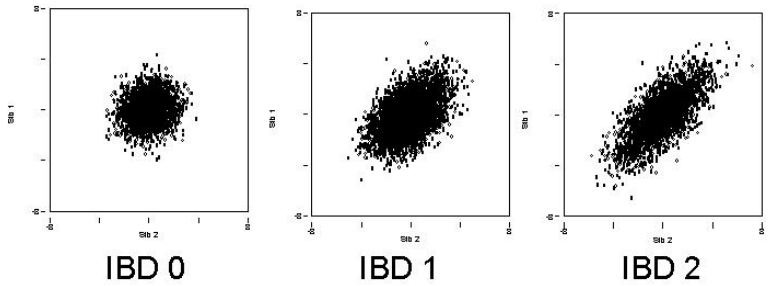
Rest of the Afternoon

- Other things you can do with Merlin …
  - Checking for errors in your data
  - Dealing with markers that aren’t independent
  - Affected sibling pair analysis

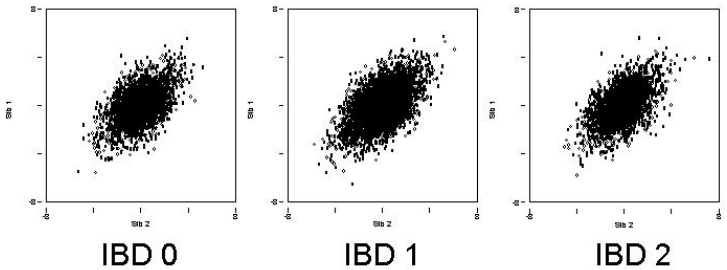
Affected Sibling Pair Analysis

Quantitative Trait Analysis

Linkage



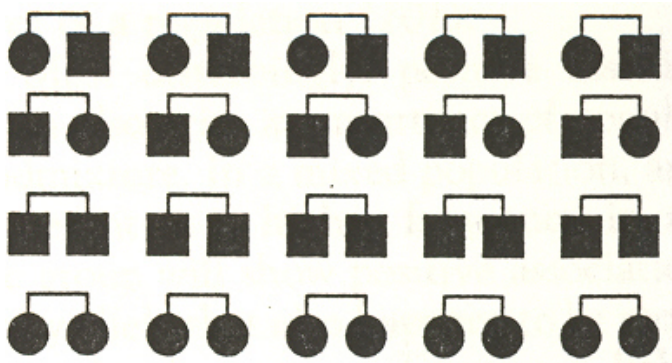
No Linkage



- Individuals who share particular regions IBD are more similar than those that don't …
- … but most linkage studies rely on affected sibling pairs, where all individuals have the same phenotype!

Allele Sharing Analysis

- Traditional analysis method for discrete traits
- Looks for regions where siblings are more similar than expected by chance
- No specific disease model assumed



Historical References

- Penrose (1953) suggested comparing IBD distributions for affected siblings.
  - Possible for highly informative markers (eg. HLA)
- Risch (1990) described effective methods for evaluating the evidence for linkage in affected sibling pair data.
- Soon after, large-scale microsatellite genotyping became possible and geneticists attempted to tackle more complex diseases…

Simple Case

- If IBD could be observed
- Each pair of individuals scored as
  - IBD=0
  - IBD=1
  - IBD=2
- Test whether sharing distribution is compatible with 1:2:1 proportions of sharing IBD 0, 1 and 2.

Sib Pair Likelihood (Fully Informative Data)

Under the null hypothesis :

$$L = (1/4)^{n_{IBD0}} (1/2)^{n_{IBD1}} (1/4)^{n_{IBD2}}$$

Under the alternative hypothesis

$$L = (z_0)^{n_{IBD0}} (z_1)^{n_{IBD1}} (z_2)^{n_{IBD2}}$$

$$LOD = \log_{10} \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = 1/4, z_1 = 1/2, z_2 = 1/4)}$$

The MLS Method

- Introduced by Risch (1990, 1992)
  - *Am J Hum Genet* **46:**242-253
- Uses IBD estimates from partially informative data
  - Uses partially informative data efficiently
- The MLS method is still one of the best methods for analysis pair data
- I will skip details here …

Non-parametric Analysis for Arbitrary Pedigrees

- Must rank general IBD configurations which include sets of more than 2 affected individuals
  - Low ranks correspond to no linkage
  - High ranks correspond to linkage
- Multiple possible orderings are possible
  - Especially for large pedigrees
- In interesting regions, IBD configurations with higher rank are more common

Non-Parametric Linkage Scores

- Introduced by Whittemore and Halpern (1994)
- The two most commonly used ones are:
  - *Pairs* statistic
    - Total number of alleles shared IBD between pairs of affected individuals in a pedigree
  - *All* statistic
    - Favors sharing of a single allele by a large number of affected individuals.

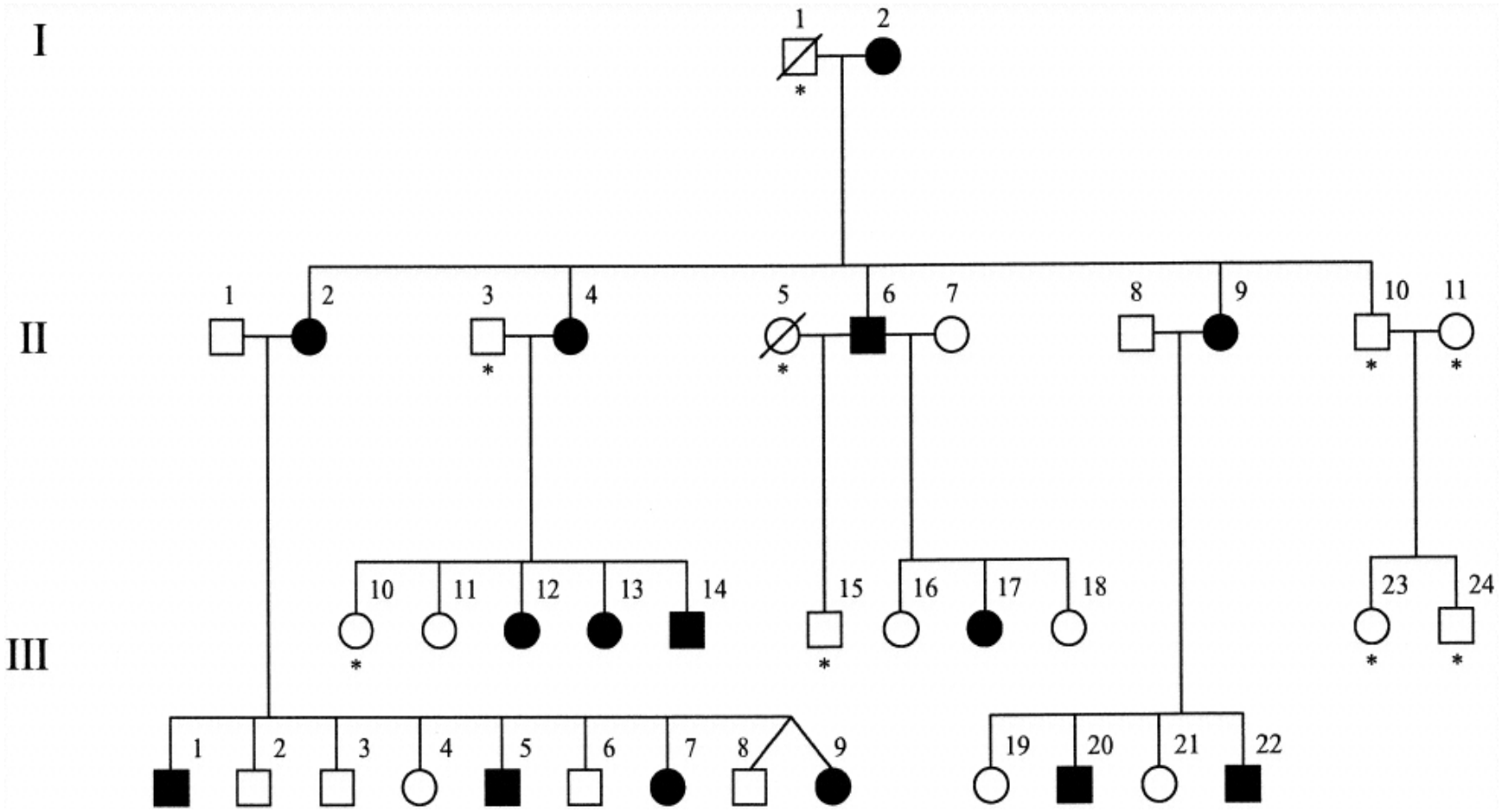
Kong and Cox Method

- A probability distribution for IBD states
  - Under the null and alternative
- Null
  - All IBD states are equally likely
- Alternative
  - Increase (or decrease) in probability of each state is modeled as a function of sharing scores
- "Generalization" of the MLS method

Parametric Linkage Analysis

- Alternative to non-parametric methods
  - Usually ideal for Mendelian disorders
- Requires a model for the disease
  - Frequency of disease allele(s)
  - Penetrance for each genotype
- Typically employed for single gene disorders and Mendelian forms of complex disorders

Typical Interesting Pedigree



Checking for Genotyping Error

Genotyping Error

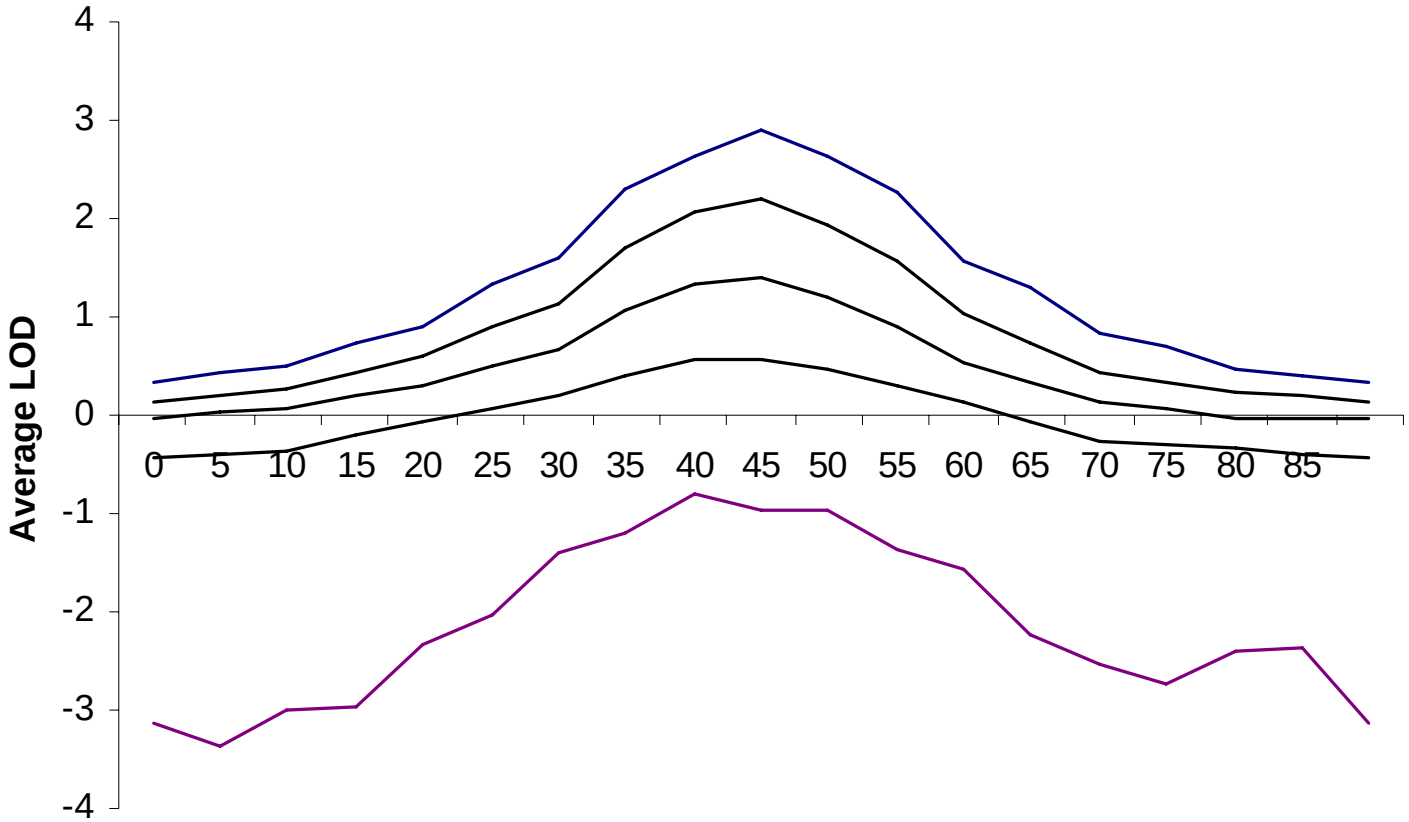
- Genotyping errors can dramatically reduce power for linkage analysis (Douglas et al, 2000; Abecasis et al, 2001)

- Explicit modeling of genotyping errors in linkage and other pedigree analyses is computationally expensive (Sobel et al, 2002)

Intuition: Why errors matter …

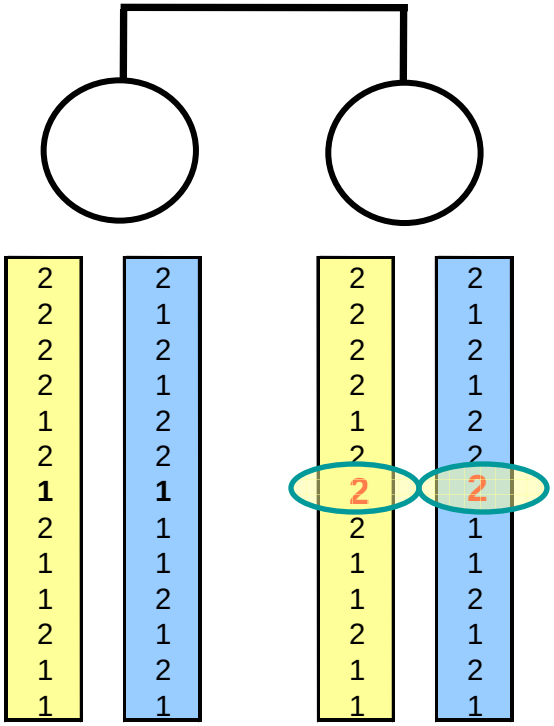
- Consider ASP sample, marker with *n* alleles
- Pick one allele at random to change
  - If it is shared (about 50% chance)
    - Sharing will likely be reduced
  - If it is not shared (about 50% chance)
    - Sharing will increase with probability about *1 / n*
- Errors propagate along chromosome

Effect on Error in ASP Sample



Error Detection

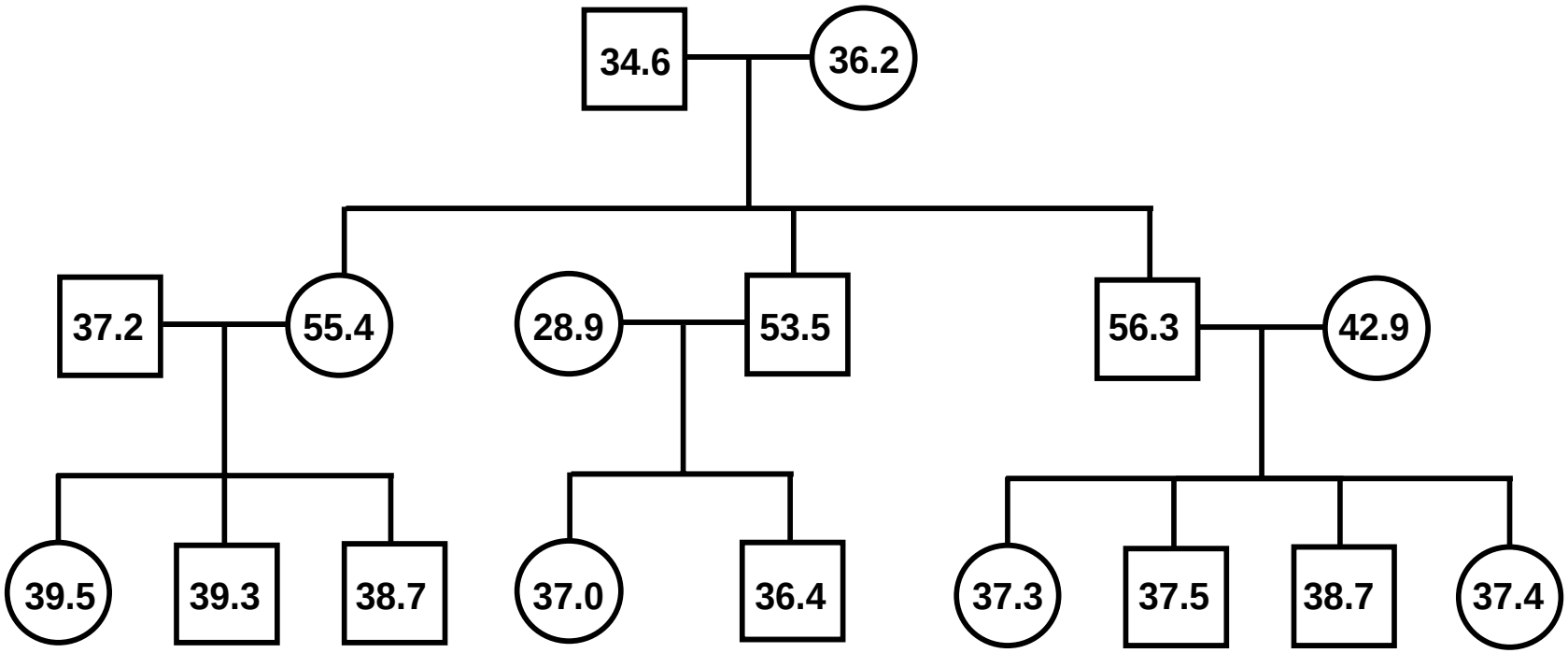
- Genotype errors can change inferences about gene flow
  - May introduce additional recombinants
- Likelihood sensitivity analysis
  - How much impact does each genotype have on likelihood of overall data



Sensitivity Analysis

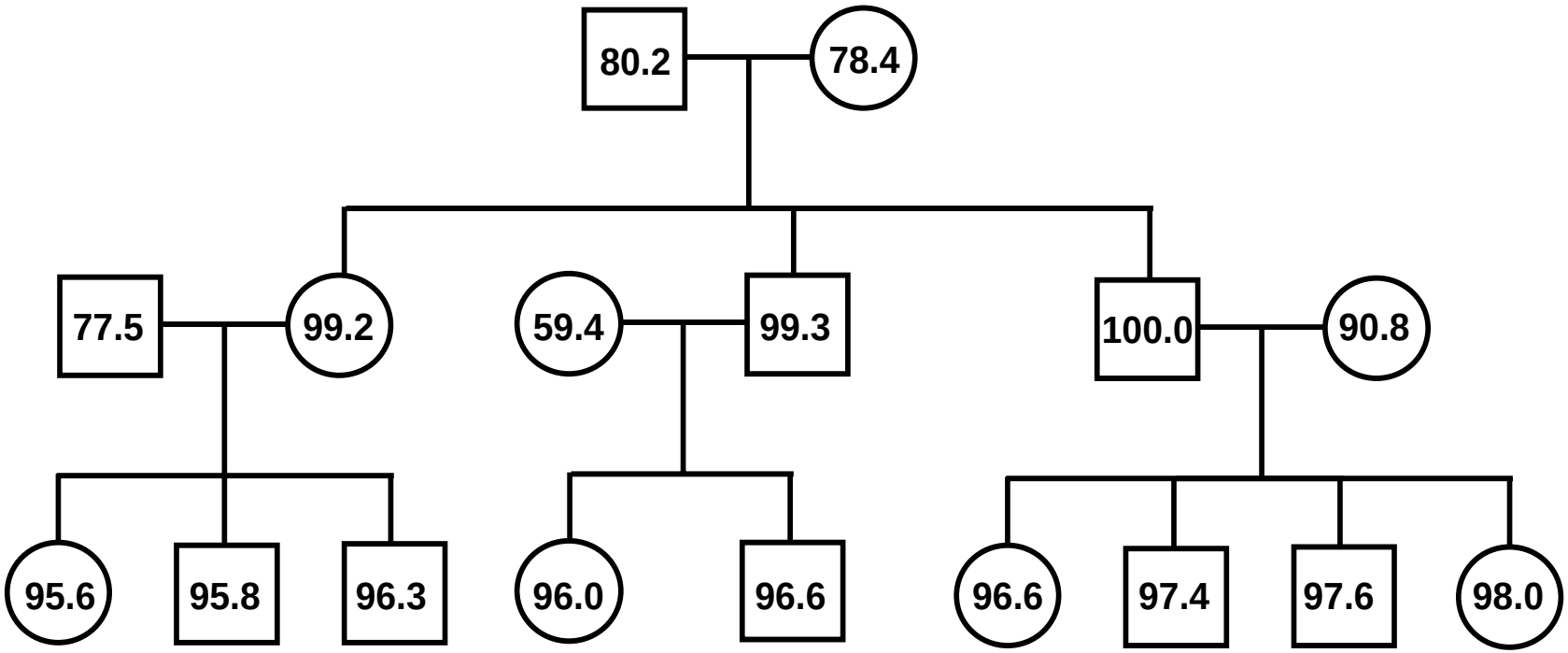
- First, calculate two likelihoods:
  - $L(\mathbf{G}|\theta)$, using actual recombination fractions
  - $L(\mathbf{G}|\theta = ½)$, assuming markers are unlinked
- Then, remove each genotype and:
  - $L(\mathbf{G} \setminus g|\theta)$
  - $L(\mathbf{G} \setminus g|\theta = ½)$
- Examine the ratio $r_{linked}/r_{unlinked}$
  - $r_{\text{linked}} = L(\mathbf{G} \setminus g|\theta) / L(\mathbf{G}|\theta)$
  - $r_{unlinked} = L(\mathbf{G} \setminus g|\theta = ½) / L(\mathbf{G}|\theta = ½)$

Mendelian Errors Detected (SNP)



% of Errors Detected in 1000 Simulations

Overall Errors Detected (SNP)



Error Detection

	Mendelian Errors	Unlikely Genotypes	Overall Detection Rate
No Genotyped Parents			
2 siblings	0.00	0.16	0.16
3 siblings	.00	.38	0.38
4 siblings	.00	.61	0.61
5 siblings	.00	.77	0.77
One Genotyped Parent			
2 siblings	0.13	0.34	0.47
3 siblings	.13	.58	0.71
4 siblings	.12	.72	0.84
5 siblings	.12	.78	0.91

Simulation: 21 SNP markers, spaced 1 cM

Markers That Are not Independent

SNPs

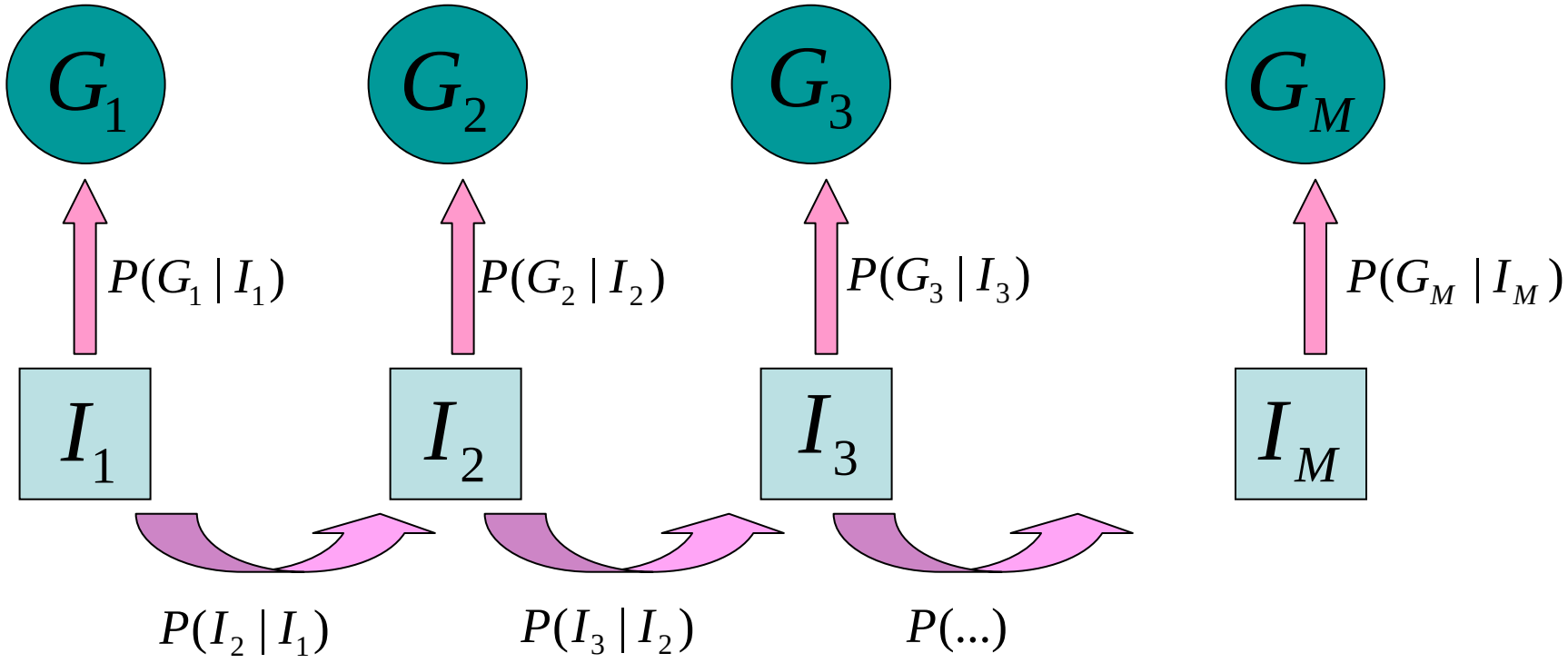
- Abundant diallelic genetic markers
- Amenable to automated genotyping
  - Fast, cheap genotyping with low error rates
- Rapidly replacing microsatellites in many linkage studies

The Problem

- Linkage analysis methods assume that markers are in linkage equilibrium
  - Violation of this assumption can produce large biases

- This assumption affects ...
  - Parametric and nonparametric linkage
  - Variance components analysis
  - Haplotype estimation

Standard Hidden Markov Model

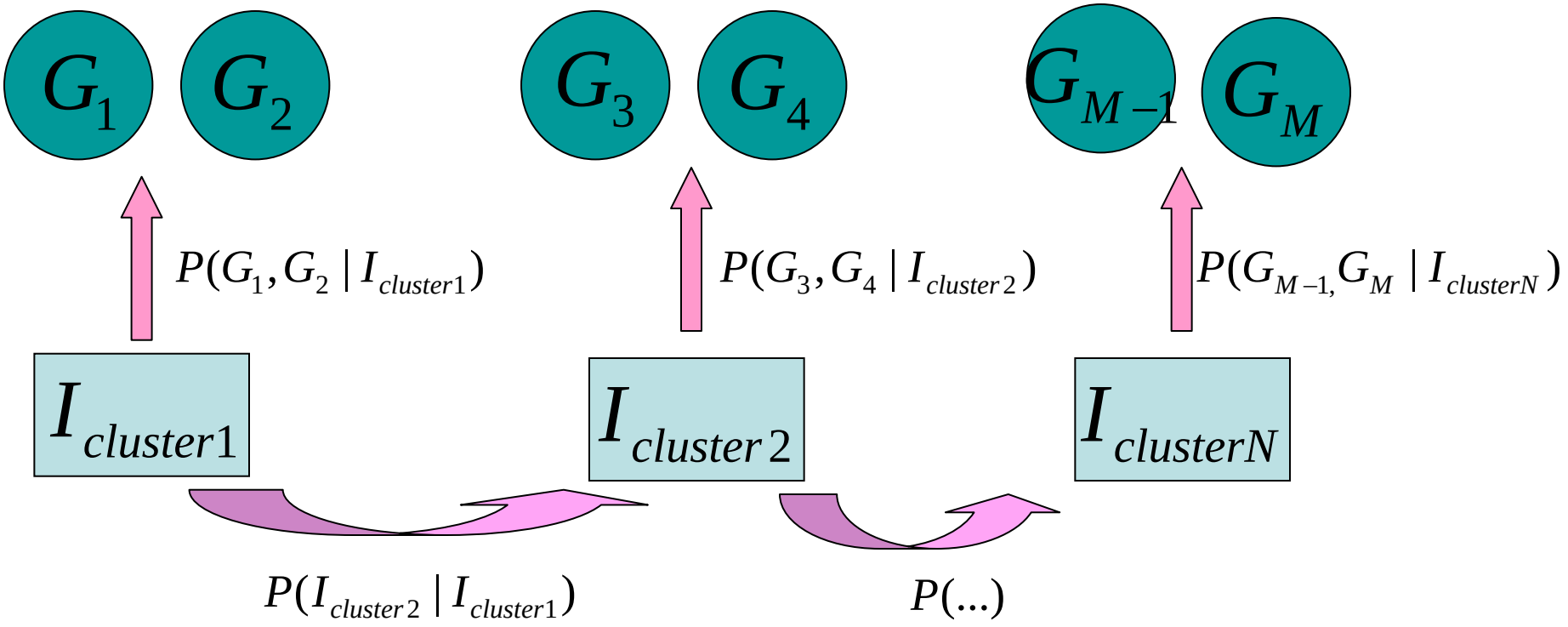


Observed Genotypes Are Connected Only Through IBD States …

Our Approach

- Cluster groups of SNPs in LD
  - Assume no recombination within clusters
  - Estimate haplotype frequencies
  - Sum over possible haplotypes for each founder
- Two pass computation …
  - Group inheritance vectors that produce identical sets of founder haplotypes
  - Calculate probability of each distinct set

Hidden Markov Model



Example With Clusters of Two Markers …

Practically …

$$P(G_1 \ldots G_C \mid f_1 \ldots f_h, v) = \sum_{H_1=1}^{h} \ldots \sum_{H_{2f}=1}^{h} \Pr(G_1 \ldots G_C \mid H_1 \ldots H_{2f}, v) \Pr(H_1 \ldots H_{2f} \mid f_1 \ldots f_h)$$

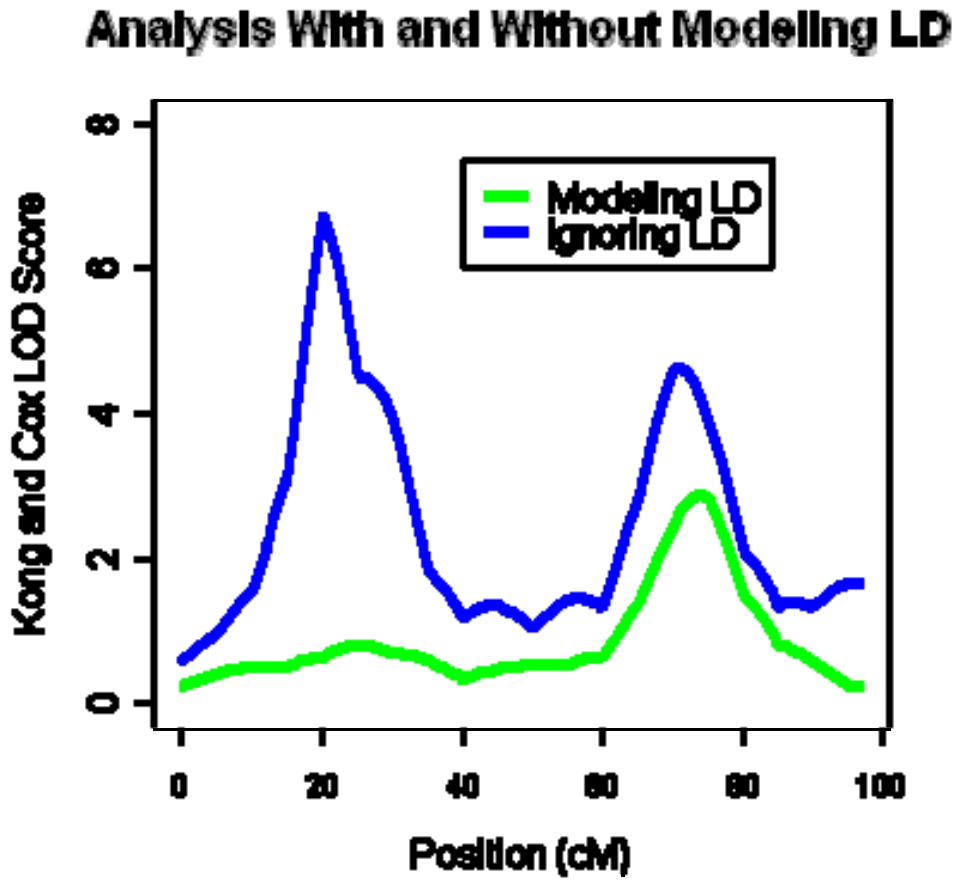
$$= \sum_{H_1=1}^{h} \ldots \sum_{H_{2f}=1}^{h} \Pr(G_1 \ldots G_C \mid H_1 \ldots H_{2f}, v) \prod_{i=1}^{2f} \Pr(H_i \mid f_1 \ldots f_h)$$

- Probability of observed genotypes $G_1 \ldots G_C$
  - Conditional on haplotype frequencies f_1 .. f_h
  - Conditional on a specific inheritance vector v
- Calculated by iterating over founder haplotypes

Computationally …

- Avoid iteration over h^{2f} founder haplotypes
  - List possible haplotype sets for each cluster
  - List is product of allele graphs for each marker
- Group inheritance vectors with identical lists
  - First, generate lists for each vector
  - Second, find equivalence groups
  - Finally, evaluate nested sum once per group

Example of What Could Happen…



Simulations …

- 2000 genotyped individuals per dataset
  - 0, 1, 2 genotyped parents per sibship
  - 2, 3, 4 genotyped affected siblings

- Clusters of 3 markers, centered 3 cM apart
  - Used Hapmap to generate haplotype frequencies
    - Clusters of 3 SNPs in 100kb windows
    - Windows are 3 Mb apart along chromosome 13
    - All SNPs had minor allele frequency > 5%
  - Simulations assumed 1 cM / Mb

Average LOD Scores (Null Hypothesis)

Analysis Strategy	Average LOD		
	Ignore LD	Model LD	Independent SNPs
No parents genotyped			
… 2 sibs per family	2.111	-0.016	-0.015
… 3 sibs per family	3.202	-0.010	-0.013
… 4 sibs per family	2.442	-0.022	-0.015
One parent genotyped			
… 2 sibs per family	0.603	-0.004	-0.003
… 3 sibs per family	0.703	-0.002	-0.004
… 4 sibs per family	0.471	-0.012	-0.010
Two parents genotyped			
… 2 sibs per family	-0.006	-0.006	-0.006
… 3 sibs per family	0.008	0.008	0.005
… 4 sibs per family	-0.014	-0.014	-0.012

5% Significance Thresholds (based on peak LODs under null)

Analysis Strategy	Significance Threshold: Ignore LD	Significance Threshold: Model LD	Significance Threshold: Independent SNPs
No parents genotyped			
… 2 sibs per family	11.37	1.33	1.26
… 3 sibs per family	15.80	1.34	1.28
… 4 sibs per family	13.46	1.27	1.17
One parent genotyped			
… 2 sibs per family	4.97	1.43	1.35
… 3 sibs per family	5.48	1.38	1.27
… 4 sibs per family	4.32	1.42	1.35
Two parents genotyped			
… 2 sibs per family	1.58	1.58	1.40
… 3 sibs per family	1.55	1.54	1.43
… 4 sibs per family	1.44	1.44	1.30

Empirical Power

Analysis Strategy	Power (Model 2)		
	Ignore LD	Model LD	Independent SNPs
No parents genotyped			
… 2 sibs per family	0.188	0.289	0.276
… 3 sibs per family	0.336	0.617	0.530
… 4 sibs per family	0.538	0.920	0.871
One parent genotyped			
… 2 sibs per family	0.163	0.207	0.184
… 3 sibs per family	0.384	0.535	0.493
… 4 sibs per family	0.697	0.852	0.811
Two parents genotyped			
… 2 sibs per family	0.153	0.155	0.171
… 3 sibs per family	0.424	0.428	0.438
… 4 sibs per family	0.800	0.800	0.794

Disease Model, $p = 0.10$, $f_{11} = 0.01$, $f_{12} = 0.02$, $f_{22} = 0.04$

Conclusions from Simulations

- Modeling linkage disequilibrium crucial
  - Especially when parental genotypes missing
- Ignoring linkage disequilibrium
  - Inflates LOD scores
  - Both small and large sibships are affected
  - Loses ability to discriminate true linkage