Errors in Genetic Data

Gonçalo Abecasis

Errors in Genetic Data

- Pedigree Errors
- Genotyping Errors
- Phenotyping Errors

Common Errors in Pedigrees

- Genetic studies require correct relationships
  - Specify expected pattern of sharing under null
- … But rely on self-reporting
- Common errors
  - Sibs are really half-sibs, half-sibs are really sibs, unrelated individuals are related

I never make mistakes, but…

- CSGA (1997) A genome-wide search for asthma susceptibility loci in ethnically diverse populations. *Nat Genet* **15:**389-92
- ~15 families with wrong relationships
- No significant evidence for linkage

- Error checking is essential!

Results

Our analysis of the pedigree structures by means of the genotypes generated as part of the genome scan highlighted that, in each of the ethnic groups, there were individuals identified as males that were likely to be females (and vice versa), half siblings labeled as full siblings, and pedigree members that showed no relationship to their supposed pedigree. Given that not all of the parents were available for study, it was difficult to distinguish between parental errors and blood- or DNA-sample mixups. In summary, 24.4% of the families contained pedigree errors and 2.8% of the families contained errors in which an individual appeared to be unrelated to the rest of the members of the pedigree and were possibly blood-sample mixups. The percentages were consistent across all ethnic groups. In total, 212 individuals were removed from the pedigrees to eliminate these errors.

Genomewide Search for Type 2 Diabetes Susceptibility Genes in Four American Populations

Margaret Gelder Ehm,[1] Maha C. Karnoub,[1] Hakan Sakul,[2,*] Kirby Gottschalk,[1] Donald C. Holt,[1] James L. Weber,[3] David Vaske,[3,†] David Briley,[1] Linda Briley,[1] Jan Kopf,[1] Patrick McMillen,[1] Quan Nguyen,[1] Melanie Reisman,[1] Eric H. Lai,[1] Geoff Joslyn,[2,‡] Nancy S. Shepherd,[1] Callum Bell,[2,§] Michael J. Wagner,[1] Daniel K. Burns,[1] and the American Diabetes Association GENNID Study Group[1]

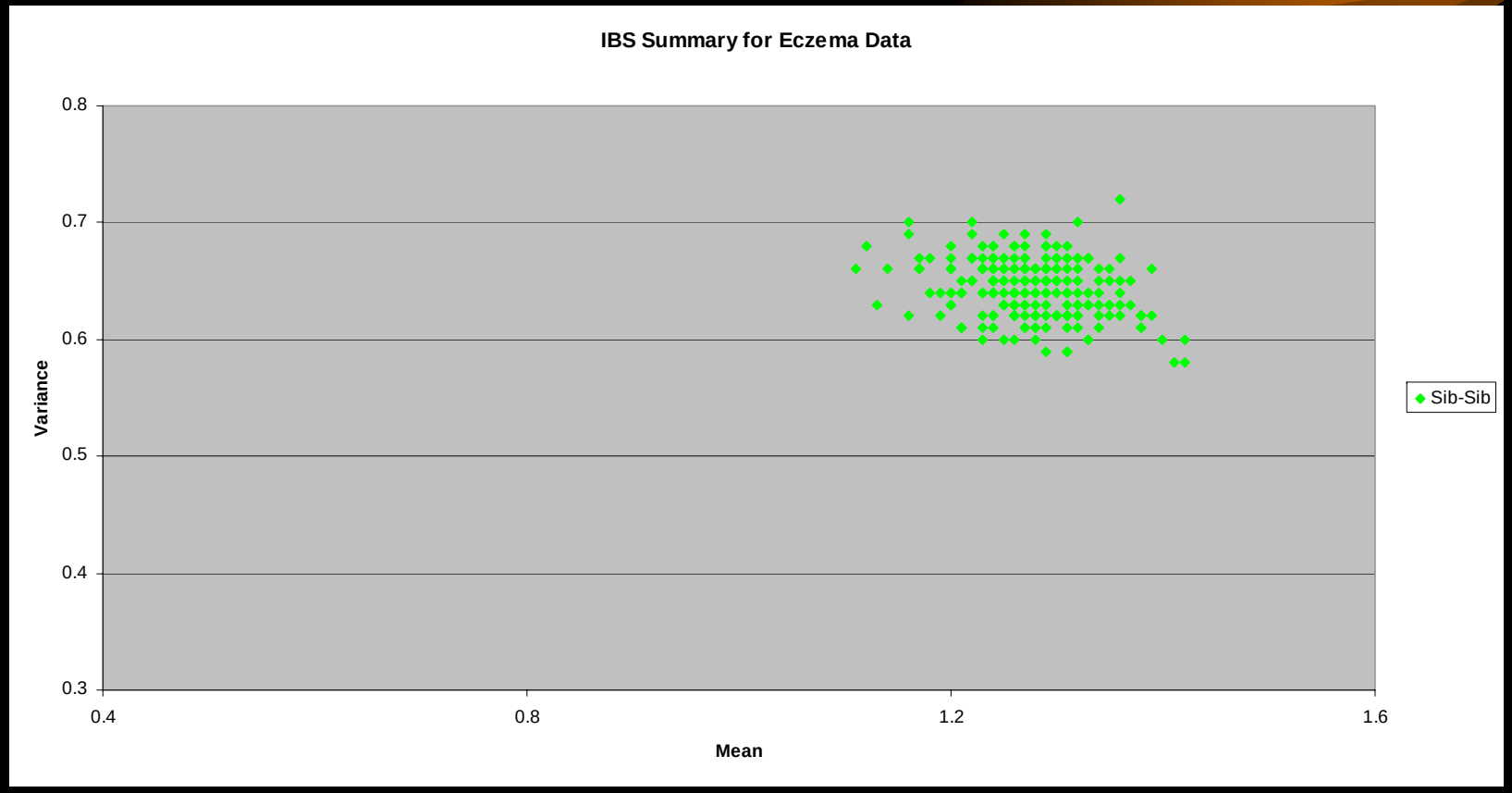
Relationship Checks

- Overall patterns of sharing
  - Depend on relationship
    - Siblings share more than half-siblings
    - Siblings share the same as parent-offspring pairs
      - On average!
      - But greater variability
    - Unrelated individuals share less than any relatives
- Can be estimated from genome-wide data
- Some errors are easily detected
  - Illegitimate offspring

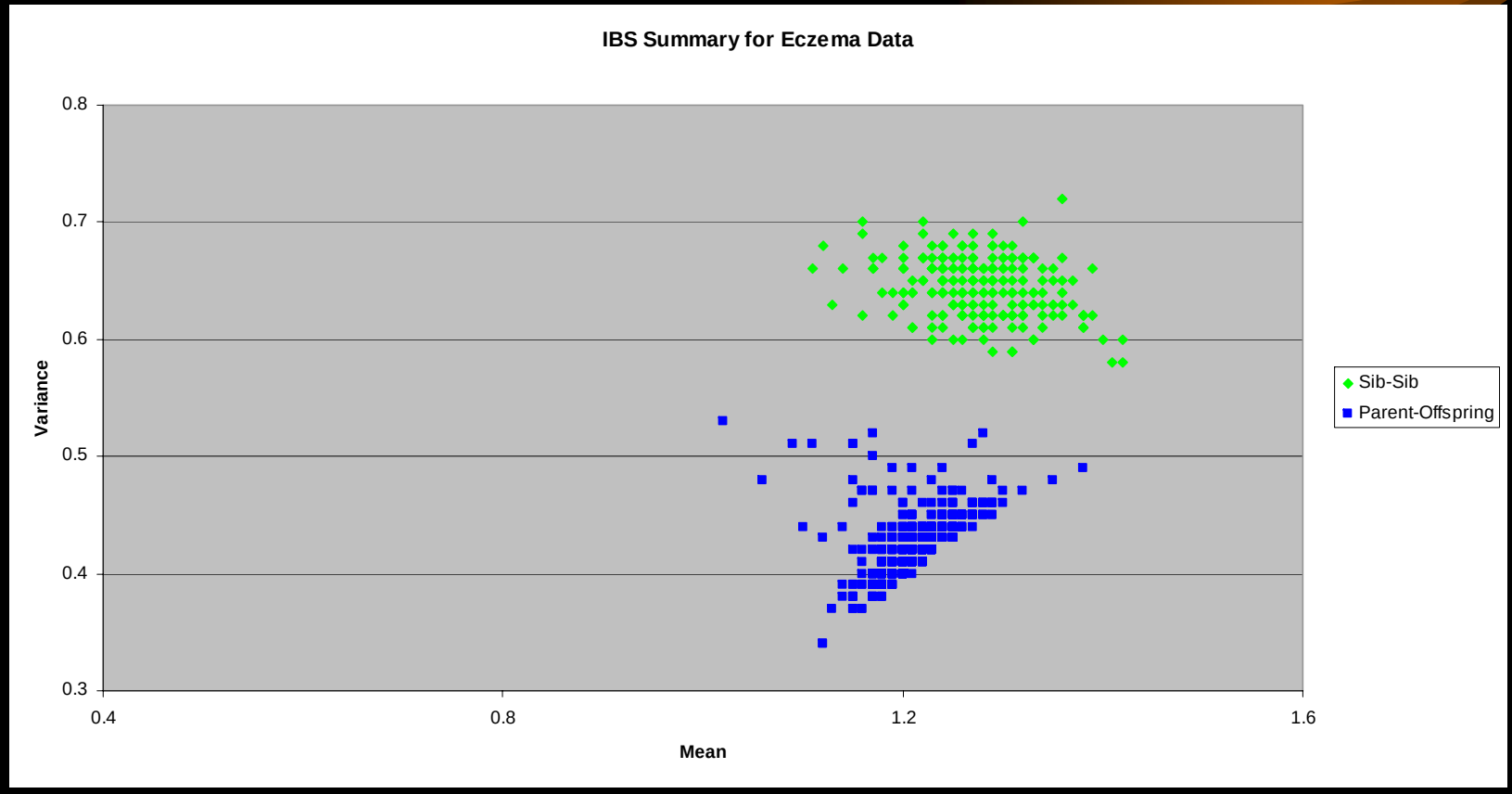
Identity-by-state

- Alleles shared by pair of individuals
  - Due to chance
    - Depends on marker informativeness
  - Shared chromosome
    - Depends on relatedness
- Define two statistics
  - Average sharing across markers
  - Variability of sharing between markers

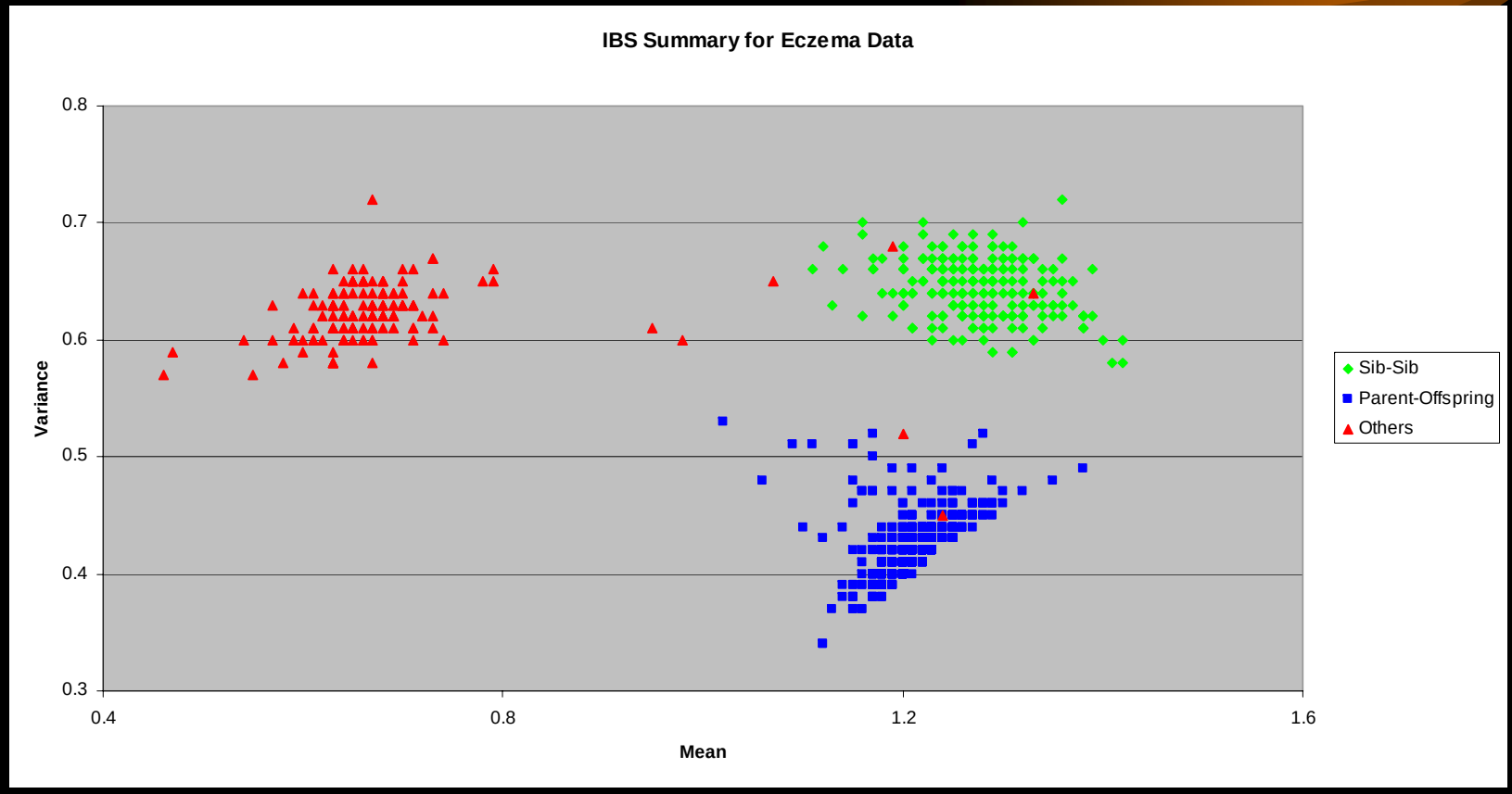
Actual Genome Scan (Sibs)



Parent-Offspring



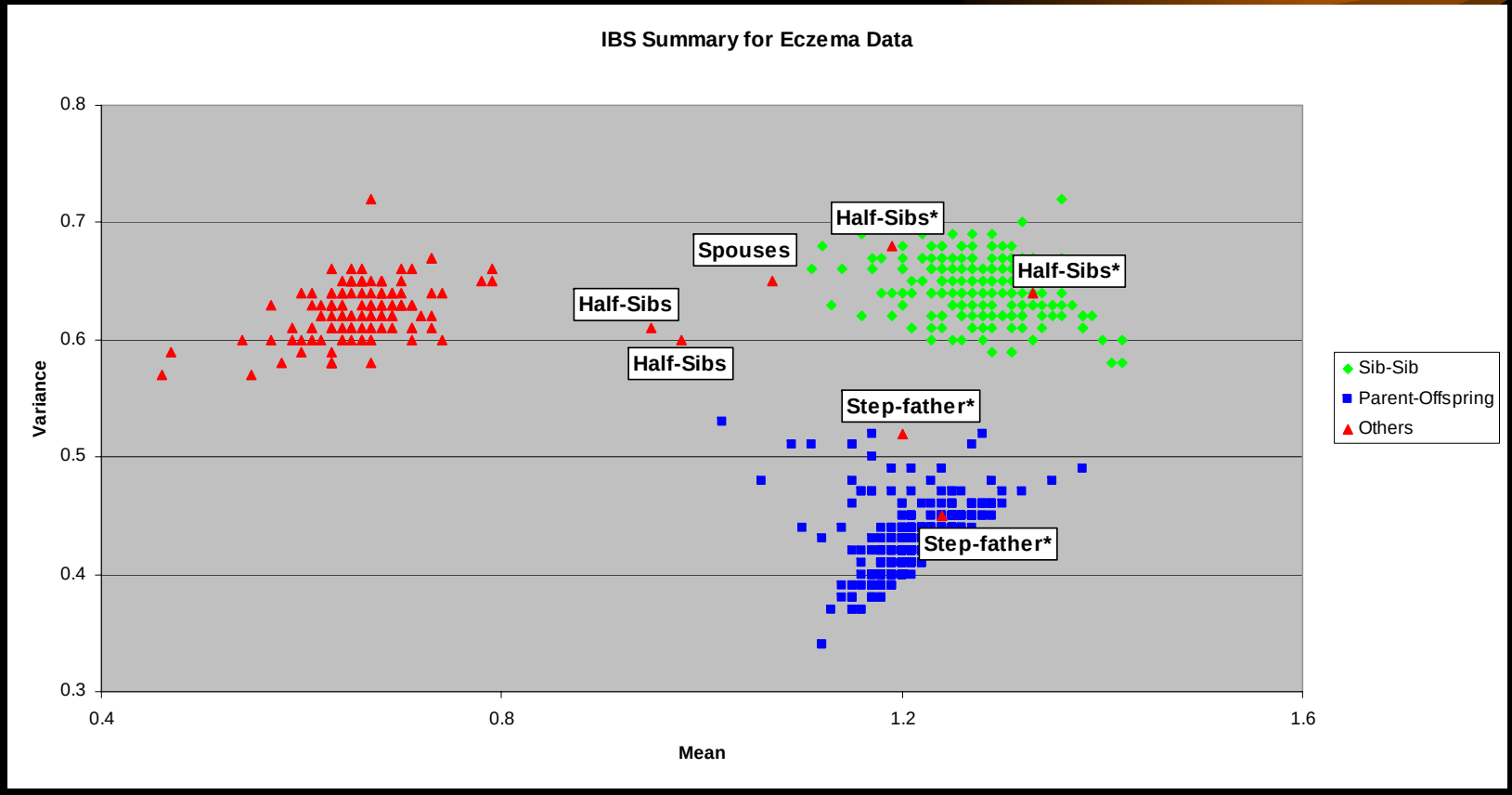
Other-Relatives



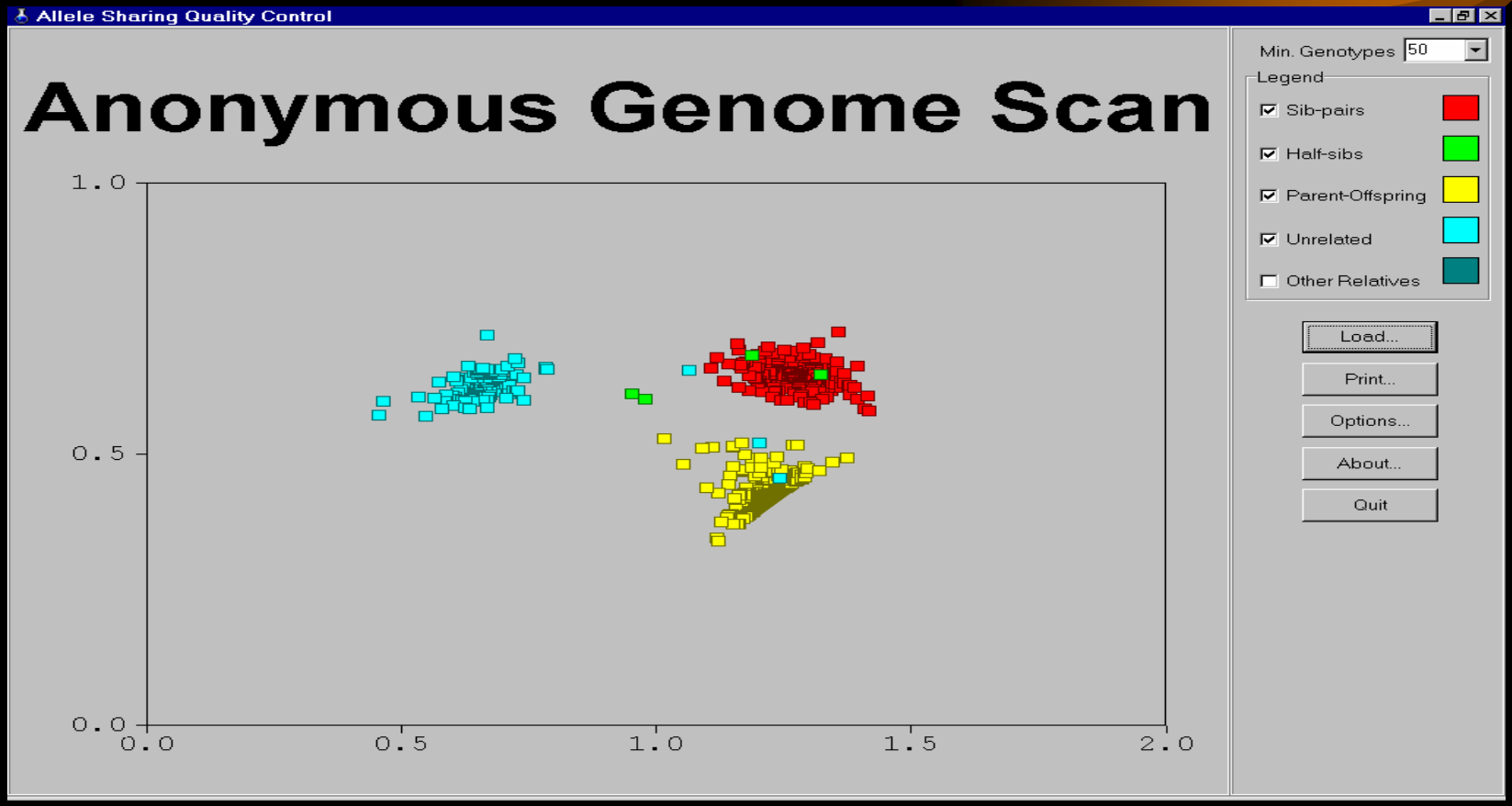
Unique Patterns of Sharing

Relation	Markers	Mean	St. Dev.
Half-Sib	311	0.95	0.61
Half-Sib	343	0.98	0.60
Spouses	320	1.07	0.65
Half-Sib	324	1.19	0.68
Step-Parent	335	1.20	0.52
Step-Parent	288	1.24	0.45
Half-Sib	289	1.33	0.64

Problems


IBS Summary for Eczema Data
Variance
Mean
0.8
0.7
0.6
0.5
0.4
0.3
0.4
0.8
1.2
1.6
Half-Sibs*
Spouses
Half-Sibs*
Half-Sibs
Half-Sibs
Step-father*
Step-father*
Sib-Sib
Parent-Offspring
Others

GRR Example


Allele Sharing Quality Control
Anonymous Genome Scan
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
Min. Genotypes
50
Legend
Sib-pairs
Half-sibs
Parent-Offspring
Unrelated
Other Relatives
Load...
Print...
Options...
About...
Quit

Alternative Approaches

- Maximum likelihood
- Calculate probability of observed data for each relationship, and select relationship that makes observed data most likely

Maximum Likelihood References

- Boehnke and Cox (1997), *AJHG* **61:**423-429
- Broman and Weber (1998), *AJHG* 63:1563-4
- McPeek and Sun (2000), *AJHG* 66:1076-94
- Epstein et al. (2000), *AJHG* 67:1219-31

Errors in Genotyping

- Increasing focus on SNPs
  - Very abundant
  - Easy to automate (only 2 alleles to score)
    - Plenty of scope for mistakes!
- Even 1% is expensive
  - ~10-50% loss of power for linkage
  - ~5-20% loss of power for association

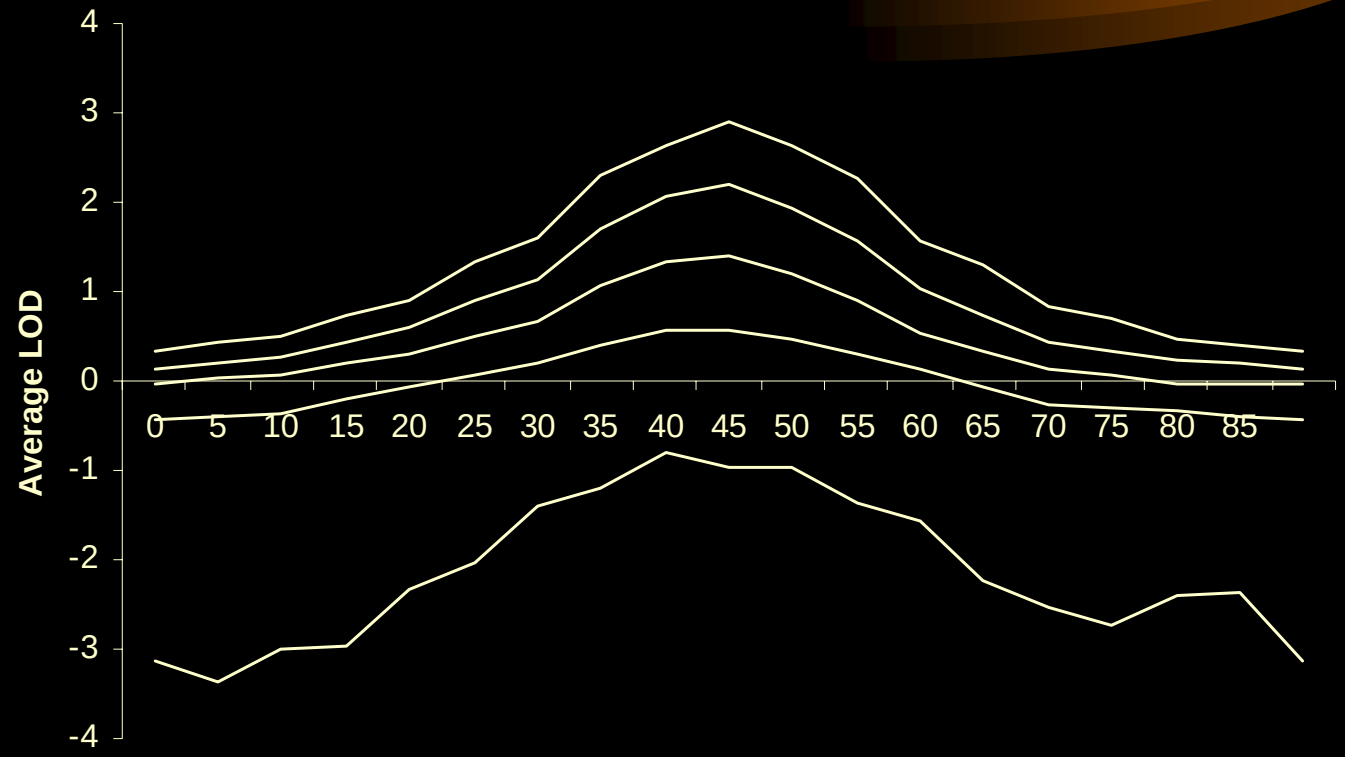
Genotyping Error

- 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



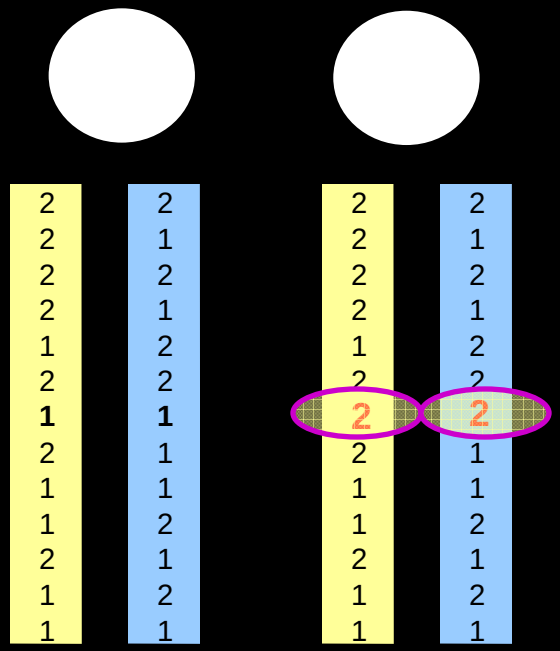
Successive lines for 0, ½, 1, 2 and 5% error.

SNP Errors Are Hard to Find

- Consider the following trio
  - Mother 1 / 2
  - Father 1 / 2
  - Child 1 / 2
- Any single genotype can be changed and the trio still looks valid
- Consistency checks detect <30% of SNP genotyping errors

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



Checking for Recombination

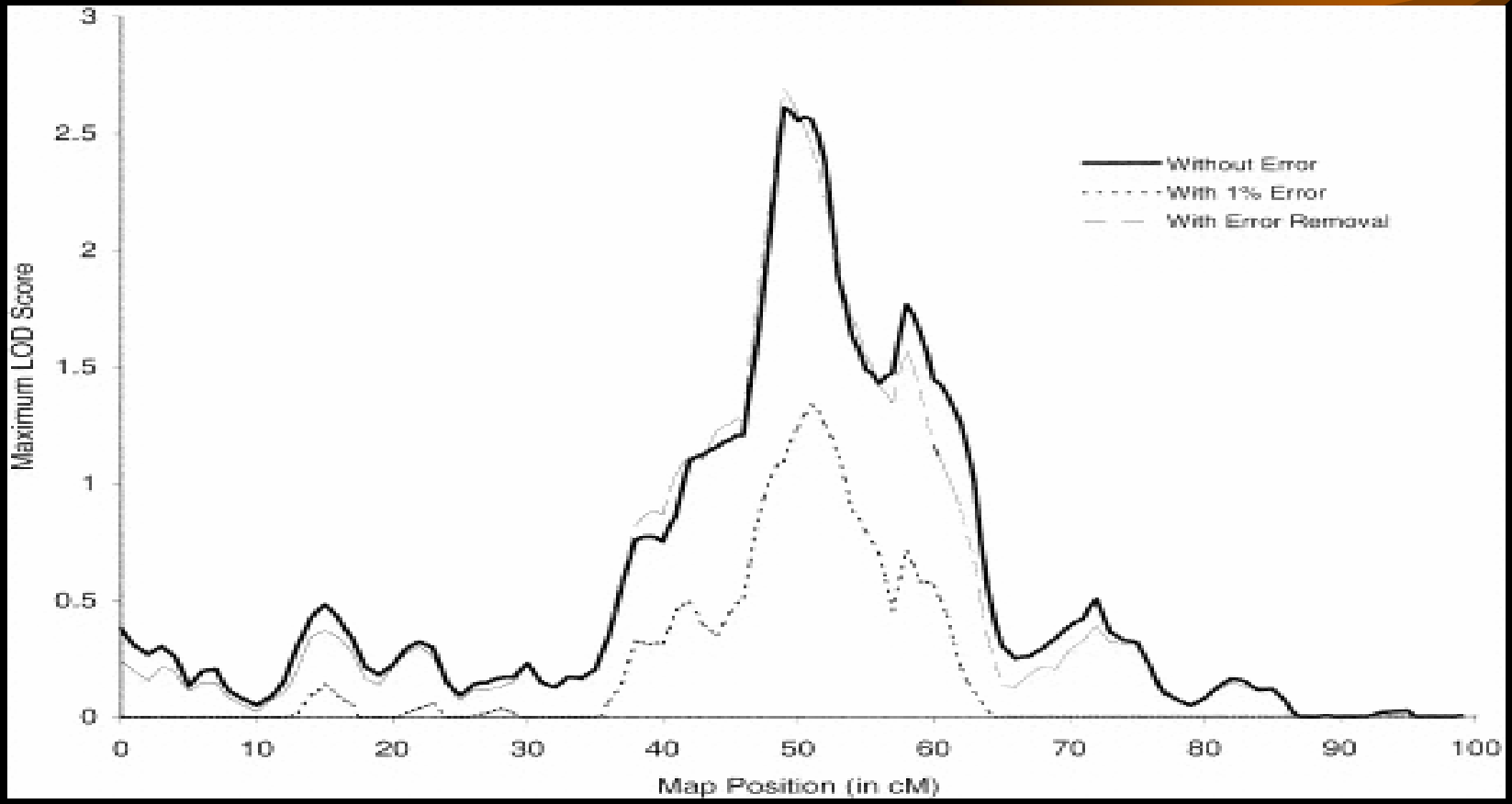
- Between closely linked markers
  - Recombination fraction < 0.01 (~ 1 Mb)
- Double recombinants almost never occur

- Requirements
  - Problem chromosome must be observed in at least two individuals
  - More effective for larger families

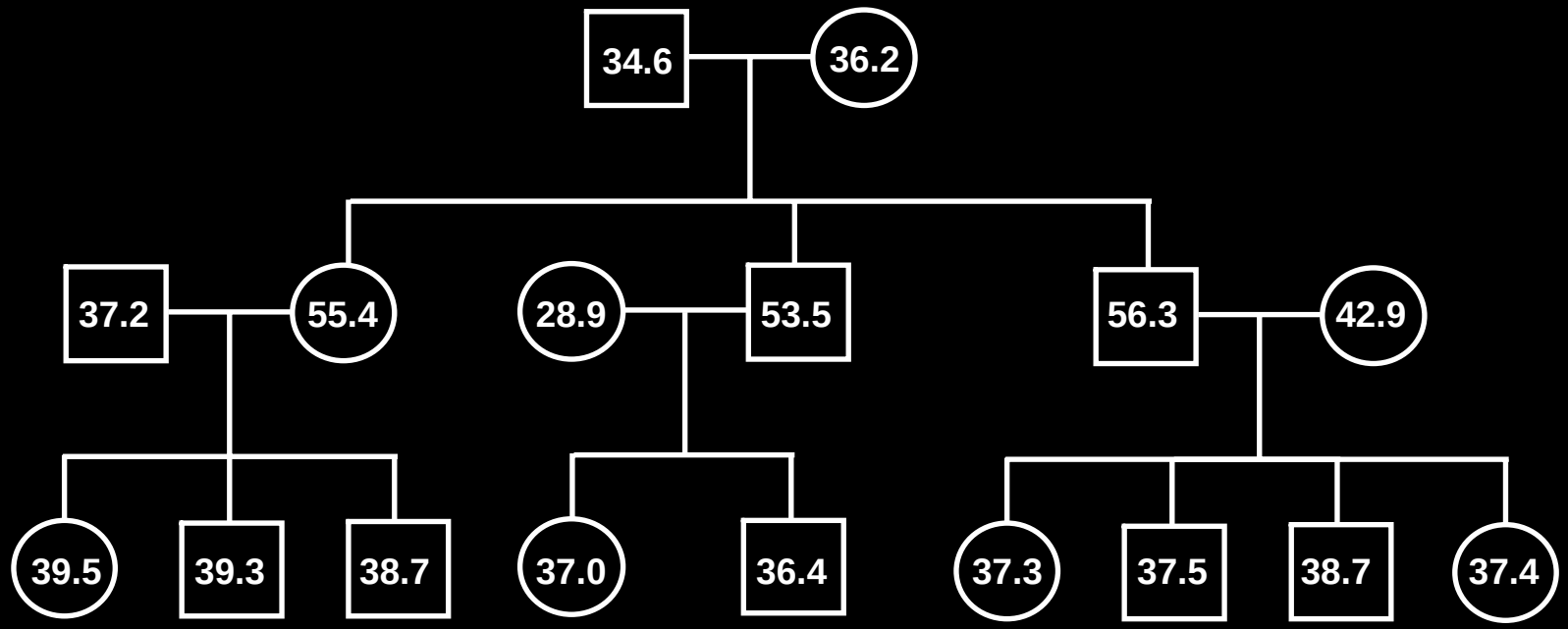
Sensitivity Analysis

- 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 $r_{linked}/r_{unlinked}$
  - $r_{\text{linked}} = L(\mathbf{G} \setminus g|\theta) / L(\mathbf{G}|\theta)$
  - $r_{unlinked} = L(\mathbf{G} \setminus g|\theta = ½) / L(\mathbf{G}|\theta = ½)$

Best Case Outcome…

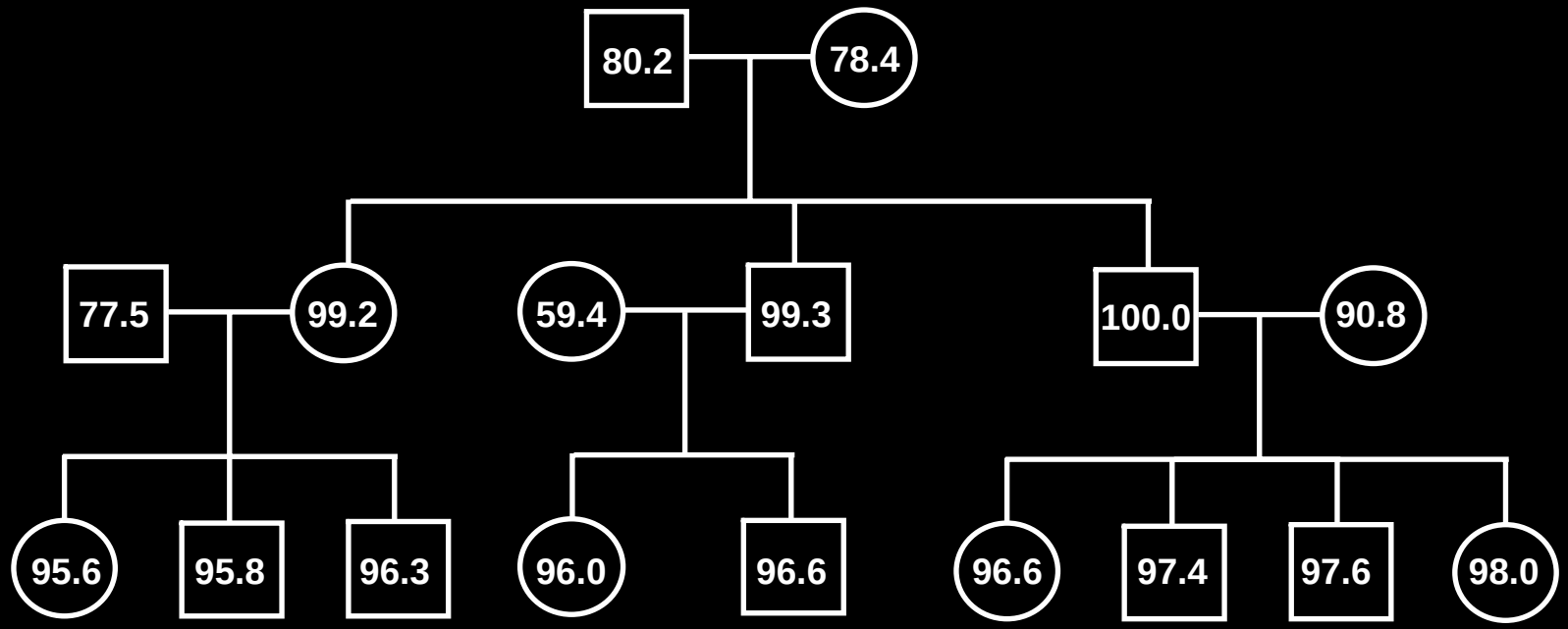


Mendelian Errors Detected (SNP)



% 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

Computational Problem

- Extend standard multipoint linkage analyses framework (Kruglyak et al, 1996) to allow efficient modeling of genotyping errors.

- Requires calculation of observed data for each possible inheritance vector.
  - Iteration over all founder alleles
  - Iteration over all possible inheritance vectors

A simple error model

- With probability *(1 – e)*
  - True and observed genotypes identical
- With probability *e*
  - Observed genotyped drawn at random from population
- More biological error models exist, but simple models such as this appear to do well in practice

Computational Problem, Previous Attempts

- Sieberts et al. (2001) carried out calculations for trios of individuals
  - Assumed no more than one error per individual
- Analyzed 3 individuals for 312 markers
  - 7.42 seconds without error model
  - 15.25 minutes with error model

Computational Problem, Merlin 2005

- 1000 sibpairs, 100 markers, 8 alleles

- 3 seconds without error model
- 5 seconds with error model
- 4.15 minutes to estimate error rates

Computational Problem, Merlin 2005

- 1000 sib-trios, 312 markers, 8 alleles

- 16 seconds without error model
- 38 seconds with error model
- ~44 minutes to estimate error rates

Brief Simulations

- 1000 sibpairs, 20 markers, 4 alleles, $\Theta = 0.05$
- Average LOD scores, 100 simulations
- Data with no effect
  - No error 0.01 (0.26)
  - Error, not modelled -1.77 (1.00)
  - Error, modelled -0.02 (0.24)
- Sibling recurrence risk = 1.5
  - No error 10.48 (2.77)
  - Error, not modelled 3.16 (1.48)
  - Error, modelled 9.02 (2.48)
  - Error, cleaned data 4.09 (1.65)

Observations for Real Data

- CIDR genome scan
  - Per allele error model fits best
  - Error rate of 0.0013 per allele
  - Likelihood ratio of 676 over 370 markers
- Marshfield genome scan
  - Per allele error model fits best
  - Error rate of 0.0036 per allele
  - Likelihood ratio of 863 over 780 markers

Error Modeling Options

--flag Uses sensitivity analysis to identify problem genotypes

--fit Estimate an error rate using all available data

--perAllele, --perGenotype

Allow user to fix error rate

Merlin Example

- Analyze data in:
  - asp.dat, asp.ped and asp.map
  - error.dat, error.ped, and error.map
- First, analyse without accounting for error
  - Use –pair or –npl for a nonparametric analysis

Removing Errors

- Use the –error option to flag problematic genotypes
- Run pedwipe to remove these from the data
- Rerun analysis without problem genotypes

Modeling Errors

- Repeat analysis with –fit and –pairs
- Compare your results …
- Convenient flags:
  - --grid, --pdf, --markerNames, …