The Lander-Green Algorithm in Practice

Biostatistics 666

Last Lecture: Lander-Green Algorithm

$$L = \sum_{I_1} \cdots \sum_{I_m} P(I_1) \prod_{i=2}^{m} P(I_i \mid I_{i-1}) \prod_{i=1}^{m} P(G_i \mid I_i)$$

- More general definition for I, the "IBD vector"
- Probability of genotypes given “IBD vector”
- Transition probabilities for the “IBD vectors”

Lander-Green Recipe

- 1. List all meiosis in the pedigree
  - There should be *2n* meiosis for *n* non-founders
- 2. List all possible IBD patterns
  - Total of 2^{2n} possible patterns by setting each meiosis to one of two possible outcomes
- 3. At each marker location, score P(G|I)
  - Evaluate using founder allele graph

Lander-Green Recipe

- 4. Build transition matrix for moving along chromosome

$$T^{\otimes n+1} = \begin{bmatrix} (1-\theta)T^{\otimes n} & \theta T^{\otimes n} \\ \theta T^{\otimes n} & (1-\theta)T^{\otimes n} \end{bmatrix}$$

  - Patterned matrix, built from matrices for individual meiosis

Lander-Green Recipe

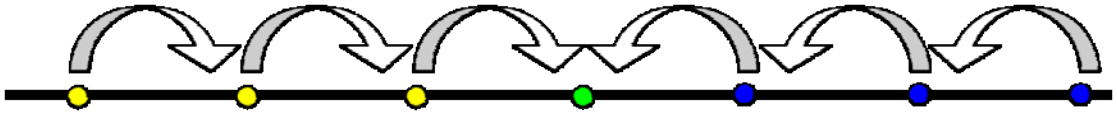
- 5. Run Markov chain
  - Start at first marker, *m*=1
    - Build a vector listing $P(G_{\text{first marker}}|I)$ for each I
  - Move along chromosome
    - Multiply vector by transition matrix
  - Combine with information at the next marker
    - Multiply each component of the vector by $P(G_{\text{current marker}}|I)$
  - Repeat previous two steps until done

Pictorial Representation

- Forward recurrence

- Backward recurrence

- At an arbitrary location



Today:
Lander-Green Algorithm in practice

- Common applications of the algorithm
  - Non-parametric linkage analysis
  - Parametric linkage analysis
  - Information content calculation (time permitting)

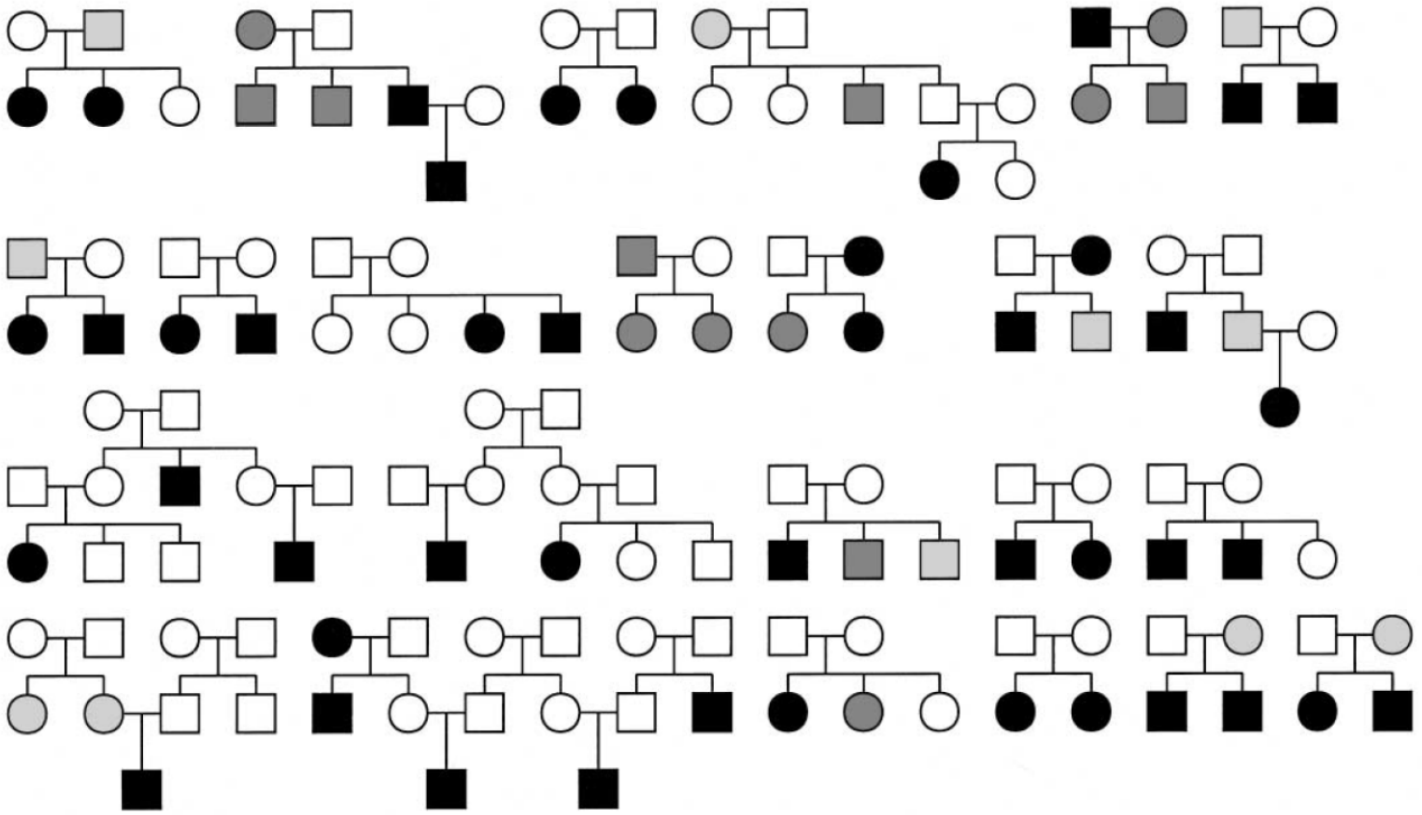
Uses of the Lander Green Algorithm

- Non-parametric linkage analysis
- Parametric linkage analysis
- Information content calculation

Nonparametric Linkage Analysis

- *Model-free*
- Does not require specification of a trait model
- Test for evidence of excess IBD sharing among affected individuals

Nonparametric Linkage Analysis Typical Dataset



Non-parametric Analysis for Arbitrary Pedigrees

- Must rank general IBD configurations
  - Low scores correspond to no linkage
  - High scores correspond to linkage
- Multiple possible orderings are possible
  - Especially for large pedigrees
- Under linkage, probability for vectors with high scores should increase

Nonparametric Linkage Statistic

- Statistic *S(I)* which ranks IBD vectors
- Then, following Whittemore and Halpern (1995)

$$S(G) = \sum_I S(I) P(I \mid G)$$

$$\mu = \sum_G S(G) P(G)$$

$$\sigma^2 = \sum_G \left[S(G) - \mu \right]^2 P(G)$$

$$Z = \frac{S(G) - \mu}{\sigma} \sim N(0,1)$$

Nonparametric Linkage Statistic

- Original definition not useful for multipoint data…
- Kruglyak et al (1996) proposed:

$$S(G) = \sum_I S(I)P(I \mid G)$$

$$\mu = \sum_I S(I)P(I)$$

$$\sigma^2 = \sum_I \left[S(I) - \mu\right]^2 P(I)$$

$$Z = \frac{S(G) - \mu}{\sigma} \sim N(0,1)$$

The Pairs Statistic

- Sum of IBD sharing for all affected pairs

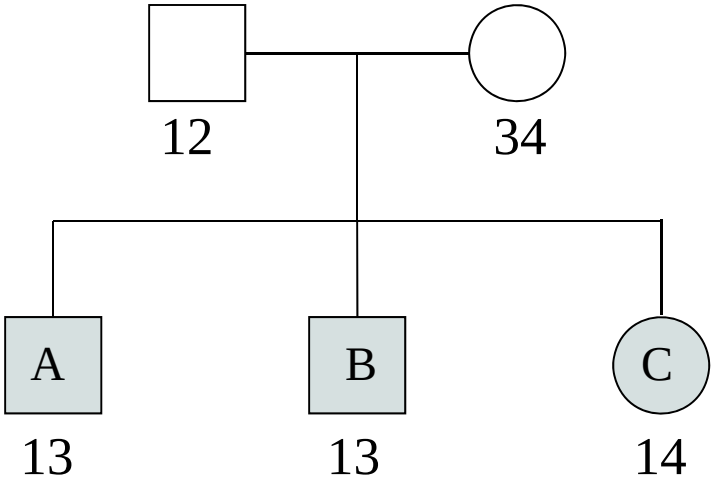
$$S_{pairs}(I) = \sum_{(a,b)\in(\text{affected pairs})} IBD(a,b \mid I)$$

$$\mu = \sum_{I} S_{pairs}(I) P_{uniform}(I)$$

$$\sigma^2 = \sum_{I} \left(S_{pairs}(I) - \mu\right)^2 P_{uniform}(I)$$

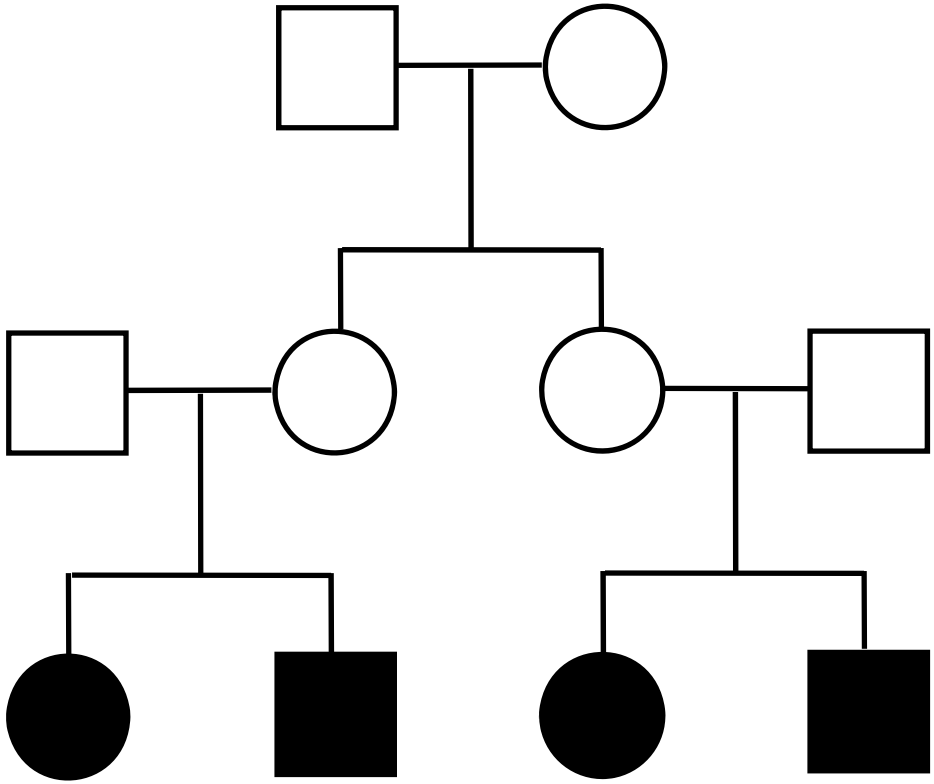
The S_{pairs} Statistic

- Total allele sharing among affected relatives

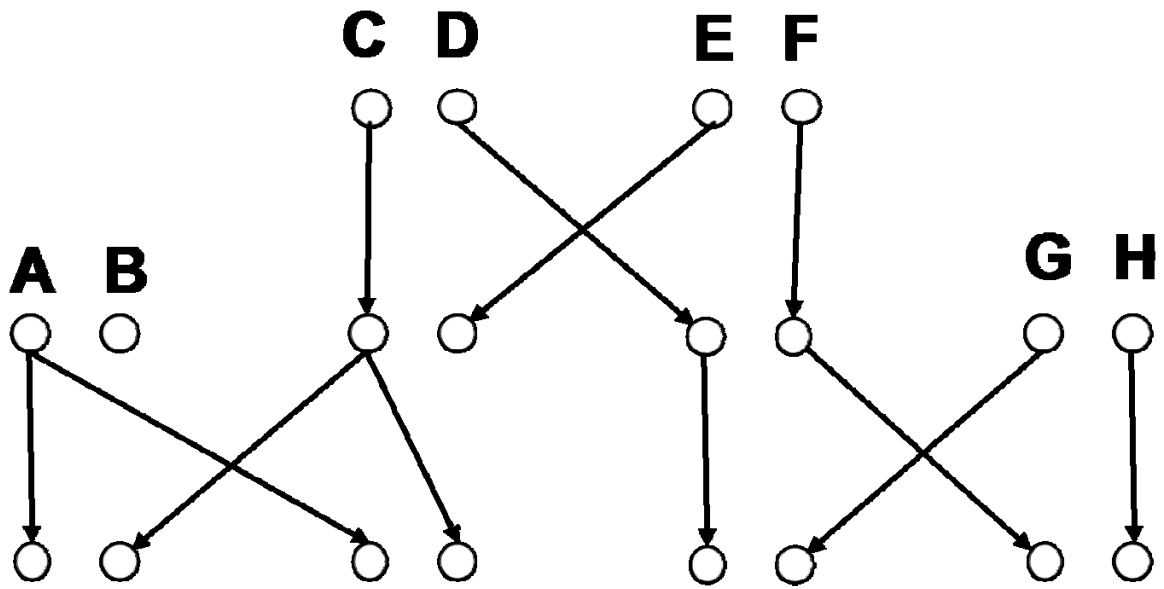


Sibpair:	A-B		A-C		B-C	
S_{Pairs} =	2	+	1	+	1	= 4

Example:
Pedigree with 4 affected individuals



What is $S_{pairs}(I)$ for this Descent Graph?



The NPL Score

- Non-parametric linkage score

$$Z(I) = \left(S_{pairs} - \mu\right) / \sigma$$

$$Z_{NPL} = \sum_{I} Z(I) P(I \mid G)$$

- Variance will always be ≤ 1 so using standard normal as reference gives conservative test.

Accurately Measuring NPL Evidence for Linkage

- For a single marker…

$$\sigma^2 = \sum_G [S(G) - \mu]^2 P(G)$$

- Estimating variance of statistic over all possible genotype configurations is not practical for multipoint analysis

- One possibility is to evaluate the empirical variance of the statistic over families in the sample…

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 is proportional to S(I)
- "Generalization" of the MLS method

Kong and Cox Method

$$P(I \mid \delta) = P(I)\left(1 + \delta \frac{S(I) - \mu}{\sigma}\right)$$

$$L(\delta) = \prod_{families} \sum_{I} P(G \mid I) P(I \mid \delta)$$

$$LOD = \log_{10} \frac{L(\hat{\delta})}{L(\delta = 0)}$$

Note: Alternative NPL Statistics

- Any arbitrary statistic can be used
- Vectors with high scores must be more common when linkage exists
- Statistics have been defined that
  - Focus on the most common allele among affecteds
  - Count number of founder alleles among affecteds
  - Evaluate linkage for quantitative traits

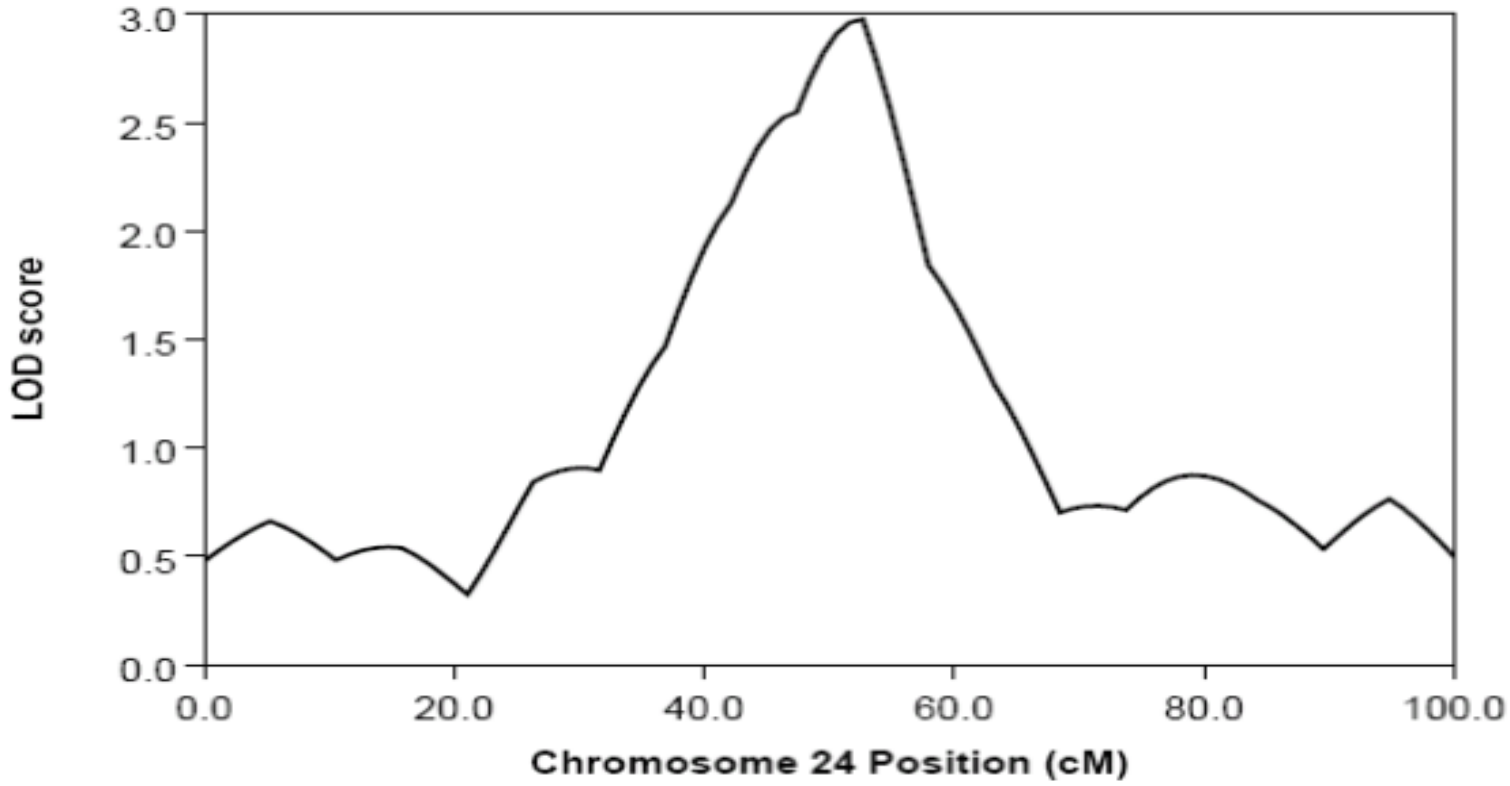
Many Alternative NPL Statistics!

TABLE I. Example 1: Outbred Sib Pair and First Cousin

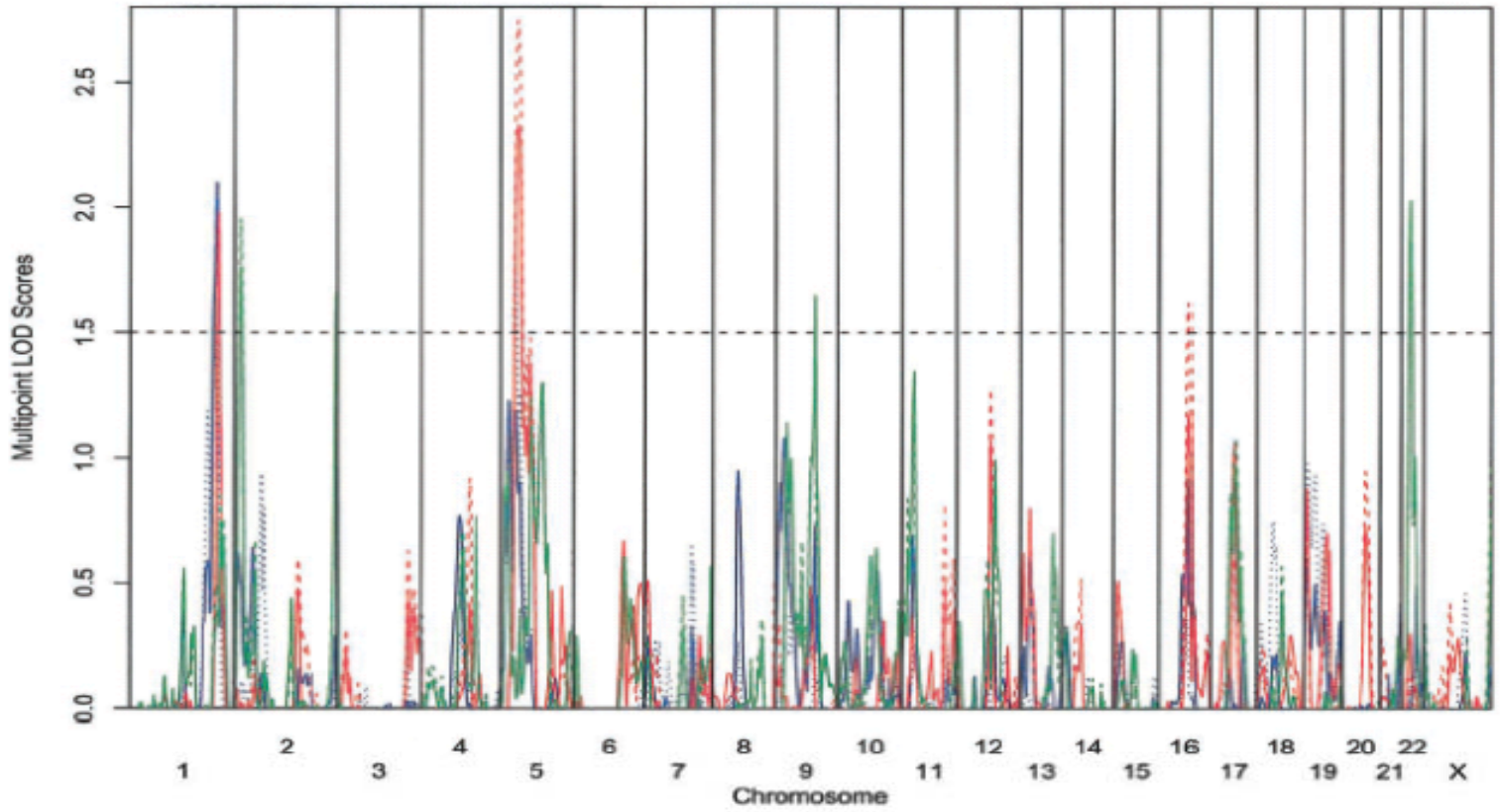
Configuration (sib, sib, cousin)	Null prob.	$S_{pairs} - \mu_0$	$S_{all} - \mu_0$	$S_{-\#alleles} - \mu_0$	$S_{everyone} - \mu_0$	$S_{-\#geno} - \mu_0$	$S_{fewest} - \mu_0$
c_1 1 2 3 4 5 6	.125	–1.5	–.41	–1.375	–.125	–.25	–.0625
c_2 1 2 3 4 1 5	.125	–.5	–.16	–.375	–.125	–.25	–.0625
c_3 1 2 1 3 4 5	.3125	–.5	–.16	–.375	–.125	–.25	–.0625
c_4 1 2 1 3 2 4	.125	.5	.09	.625	–.125	–.25	–.0625
c_5 1 2 1 2 3 4	.1875	.5	.09	.625	–.125	.75	–.0625
c_6 1 2 1 3 1 4	.0625	1.5	.59	.625	.875	–.25	–.0625
c_7 1 2 1 2 2 3	.0625	2.5	.84	1.625	.875	.75	.9375

McPeek (1999) Genetic Epidemiology 16:225–249

Typical Plot for NPL Along Chromosome



Age Related Macular Degeneration: Example of Non-Parametric Scan



All macular degeneration cases; **Geographic Atrophy;** **Neovascular Disease**

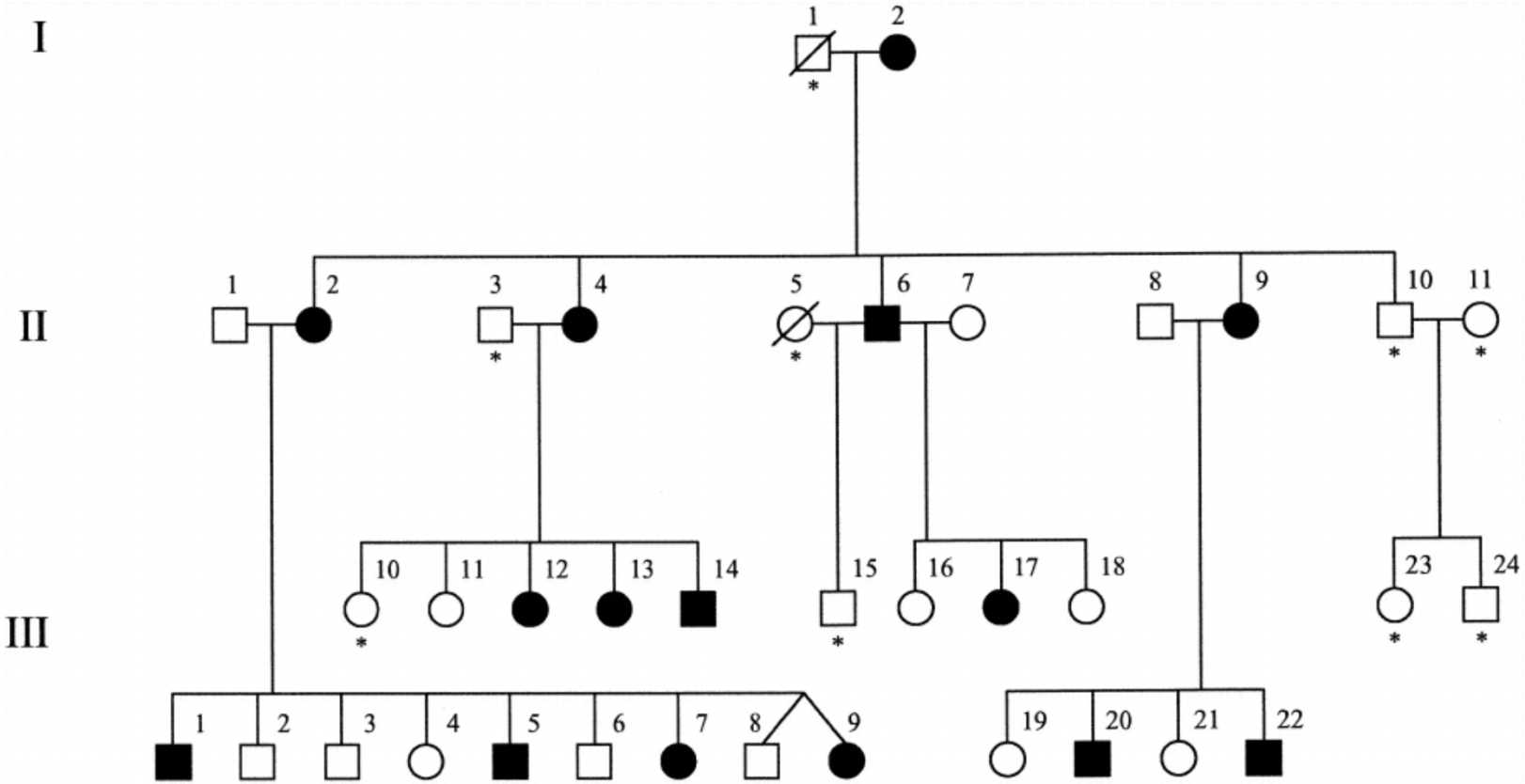
Parametric Linkage Analysis

- *X* phenotype data (affected/normal)
- *I* inheritance vector (meiosis outcomes)

- Calculate *P(X|I)* based on…

- Trait locus allele frequencies
  - *p* and *q*
- Penetrances for each genotype
  - f_{11}, f_{12}, f_{22}

Parametric Linkage Analysis: Typical Interesting Pedigree



Parametric Linkage Analysis

$$P(X \mid I) = \sum_{a_1} \cdots \sum_{a_{2f}} \prod_{i} P(a_i) \prod_{j} P(X_j \mid \mathbf{a}, I)$$

- Sum over all allele states for each founder
  - Due to incomplete penetrance
- Once $P(X|I)$ is available, the trait "plugs into" the calculation as if it was a marker locus
- $P(X|I)$ will typically be large for only a few I

Likelihood Ratio Test

- Evaluate evidence for linkage as…

$$LR(I) = \frac{P(X \mid I_{observed})}{\sum_{i \in I^*} P(X \mid i) P_{uniform}(i)}$$

- Is set of meiotic outcomes suggested by marker data likely given the trait model?

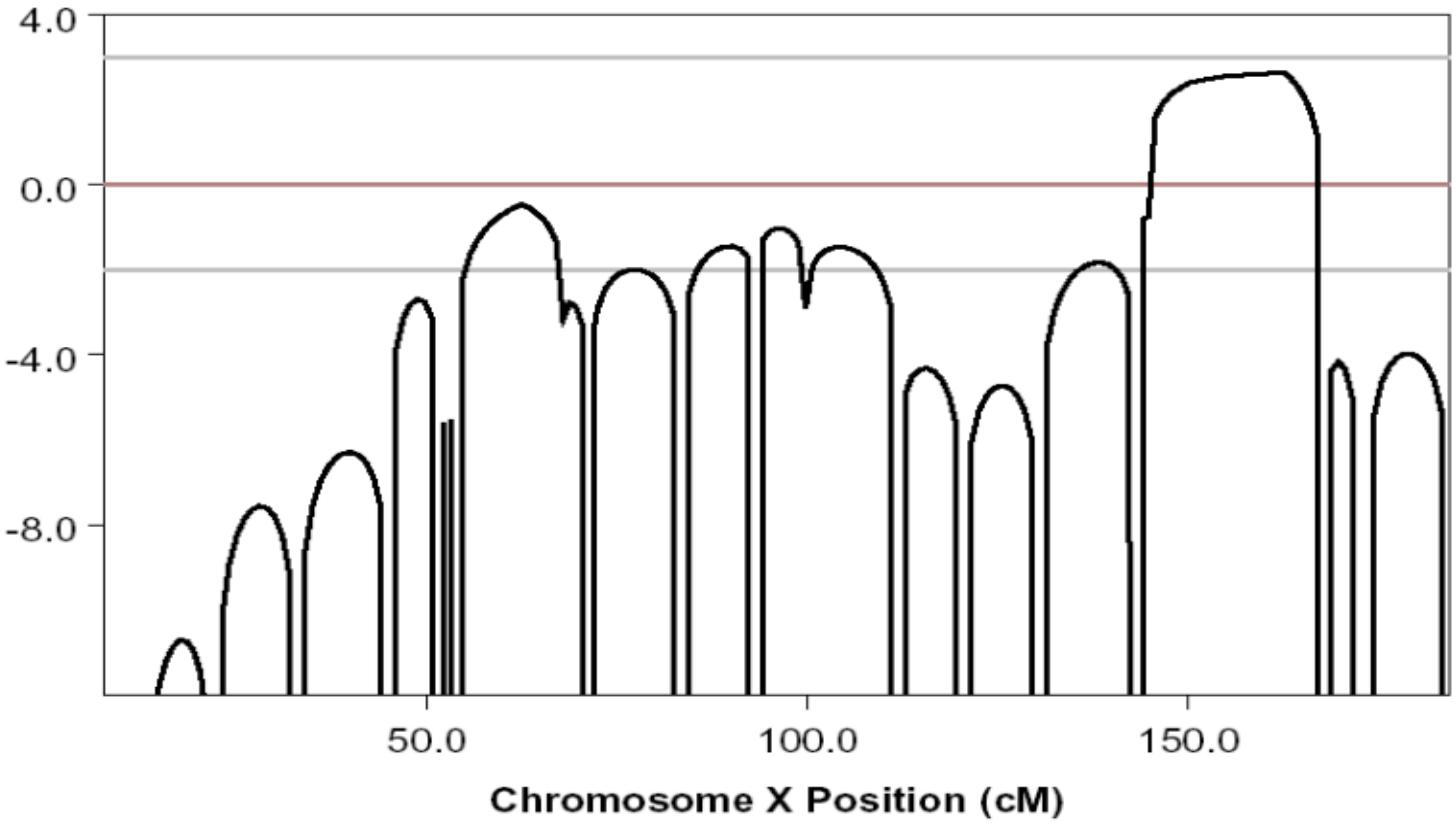
Allowing for uncertainty…

- Weighted sum over possible meiotic outcomes…

$$LR = \sum_{i \in I^*} LR(i) P(i \mid G)$$

$$= \frac{\sum_{i \in I^*} P(X \mid i) P(i \mid G)}{\sum_{i \in I^*} P(X \mid i) P_{uniform}(i)}$$

Parametric LOD Score Plot: X-Linked Cone Rod Dystrophy Example



Genotype Data Informativeness

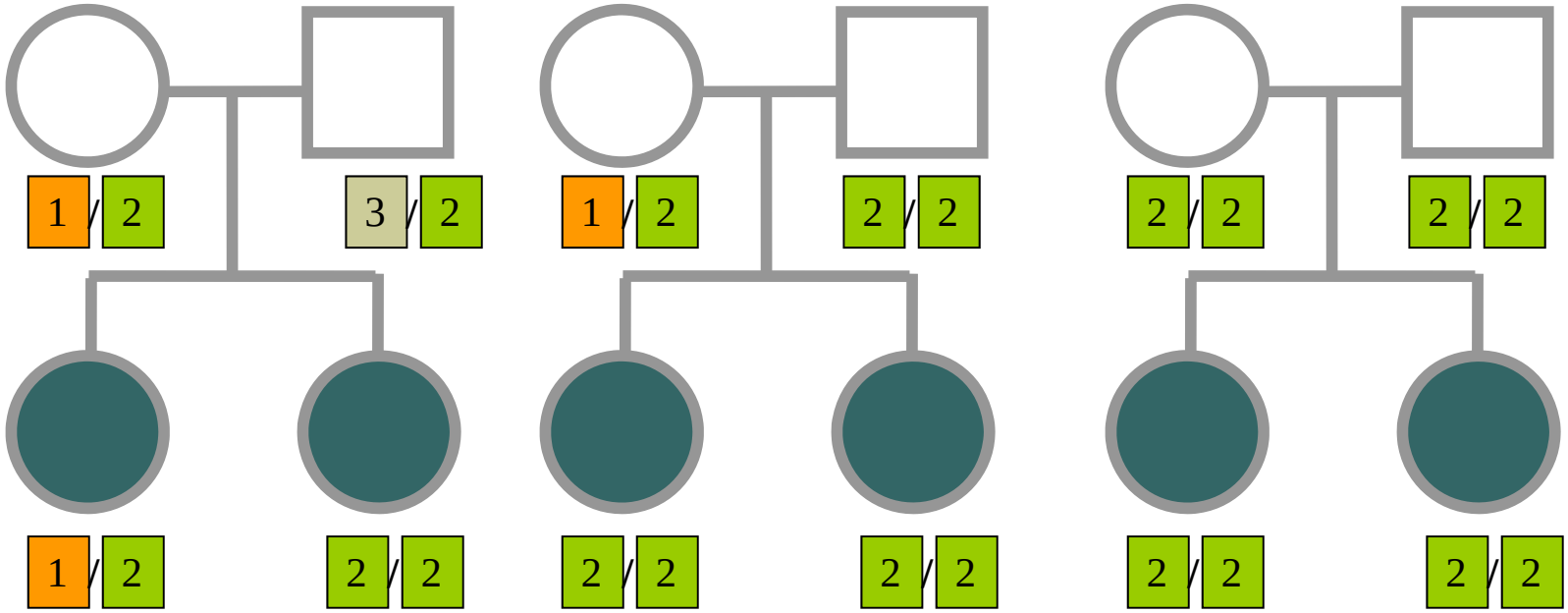
- Based on the Shannon entropy measure:

$$E = -\sum P_i \log_2 P_i$$

$$I = 1 - \frac{E}{E_0}$$

- Ranges between 0 and 1.
- Randomness in distribution of conditional probabilities.

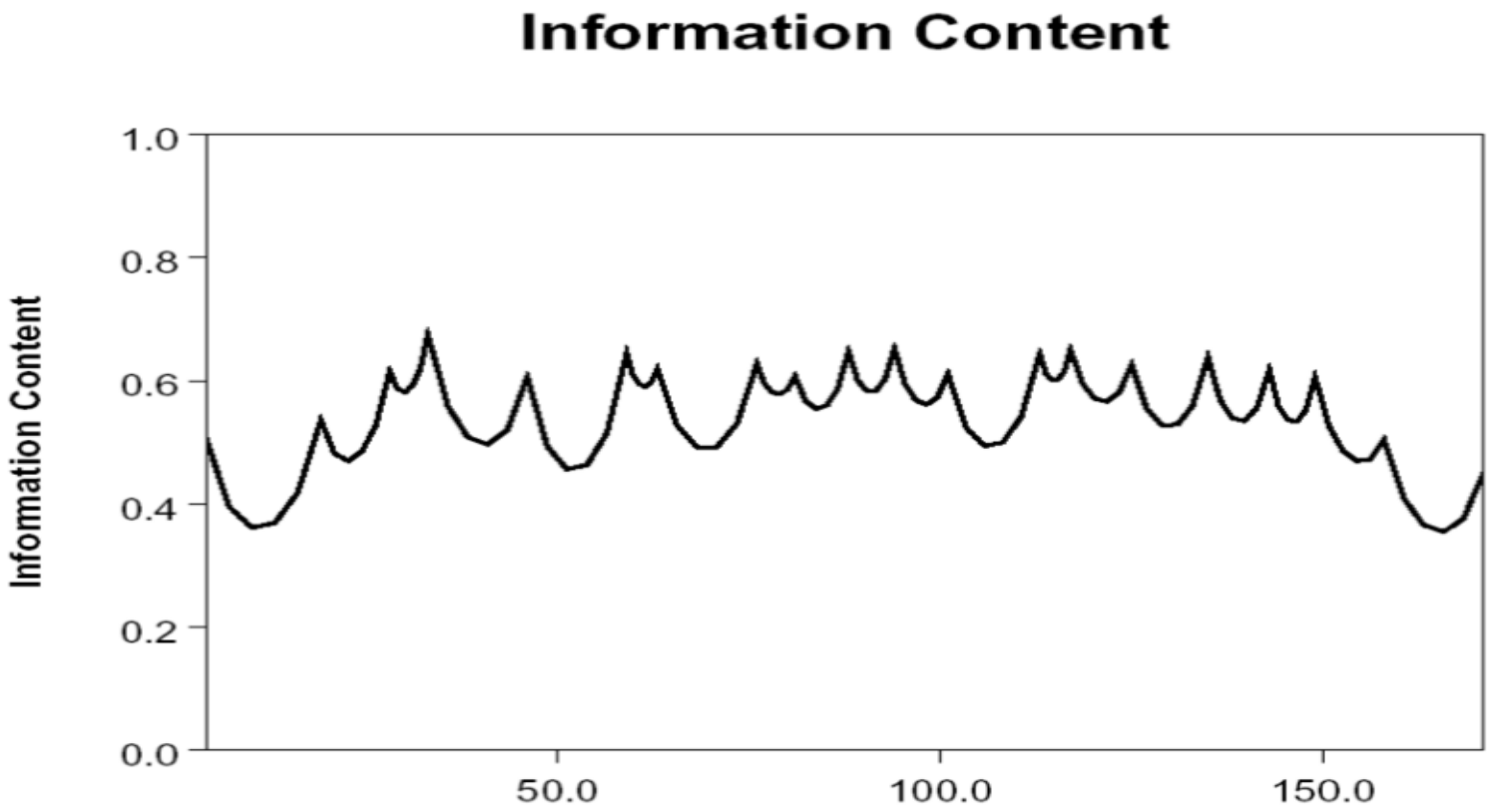
Some Exemplar Entropies



Information = 1 **Information = 0.5** **Information = 0**

(with 4 inheritance vectors)

Example of Multipoint Information Content



More on Information Content...

- The theoretical maximum is 1.0
  - All probability concentrated on one inheritance vector
- The practical maximum is lower
  - It will depend on which individuals are genotyped
- Useful in a comparative manner
  - Identifies regions where study conclusions are less certain

Today

- Non-parametric linkage analysis
- Parametric linkage analysis
- Information content

Reference

- Kruglyak, Daly, Reeve-Daly, Lander (1996) *Am J Hum Genet* **58:**1347-63

- Whittemore and Halpern (1994) *Biometrics* **50:**109-117