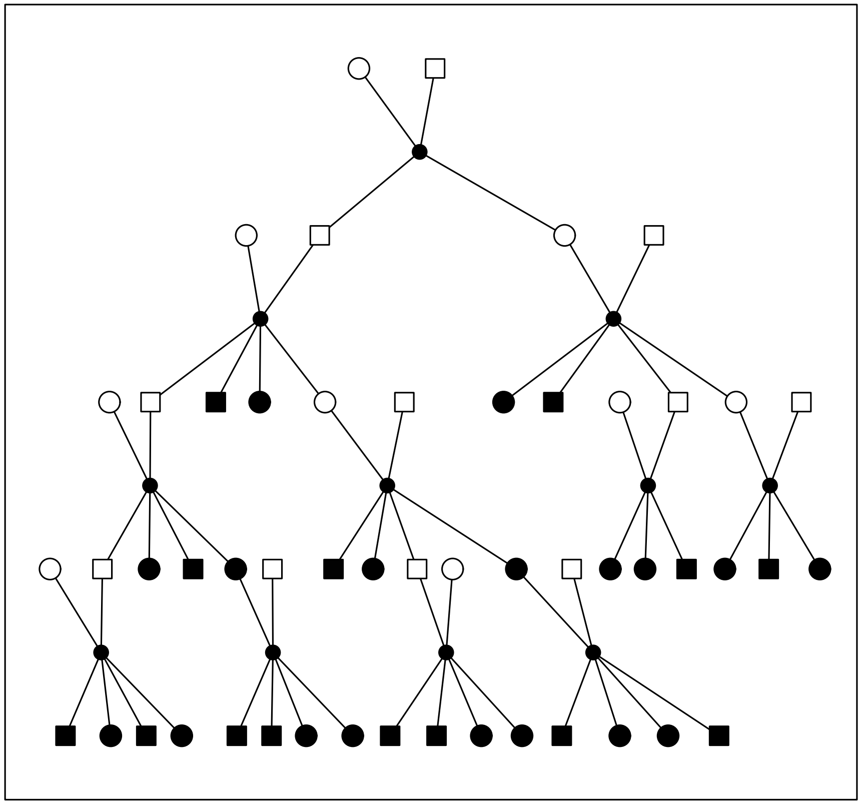
Uncertainty in Inheritance and the Detection of Genetic Linkage



Elizabeth Thompson
University of Washington

For Fields Institute, Toronto
Lecture 2, April 4, 2006

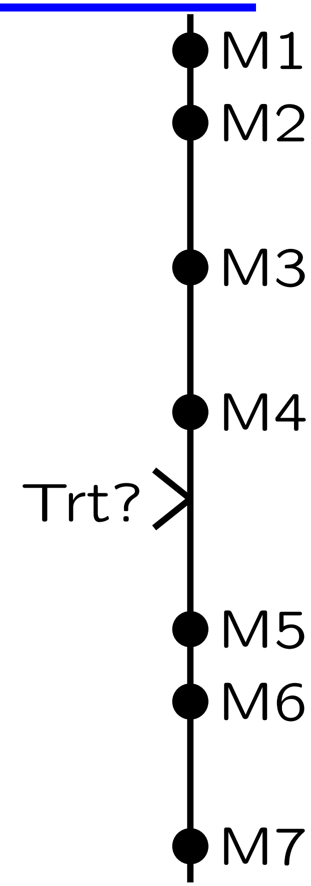
Linkage analysis with pedigree data

GIVEN:

- A set of pedigrees, and some trait of interest.
- A set of DNA markers, with known genetic model (genetic map, and allele frequencies).
- Data on trait(s) and at markers, for some subset of the individuals.

QUESTION:

- Does any DNA on the chromosome of the markers affect the trait? H_0 : No.
- If so, what is the likely location of this DNA, relative to markers.



Linkage detection and linkage estimation

- Two broad questions:
  Tests for detection of linkage (many possible statistics)
  Estimating locations using log-likelihood ratios (lod scores)

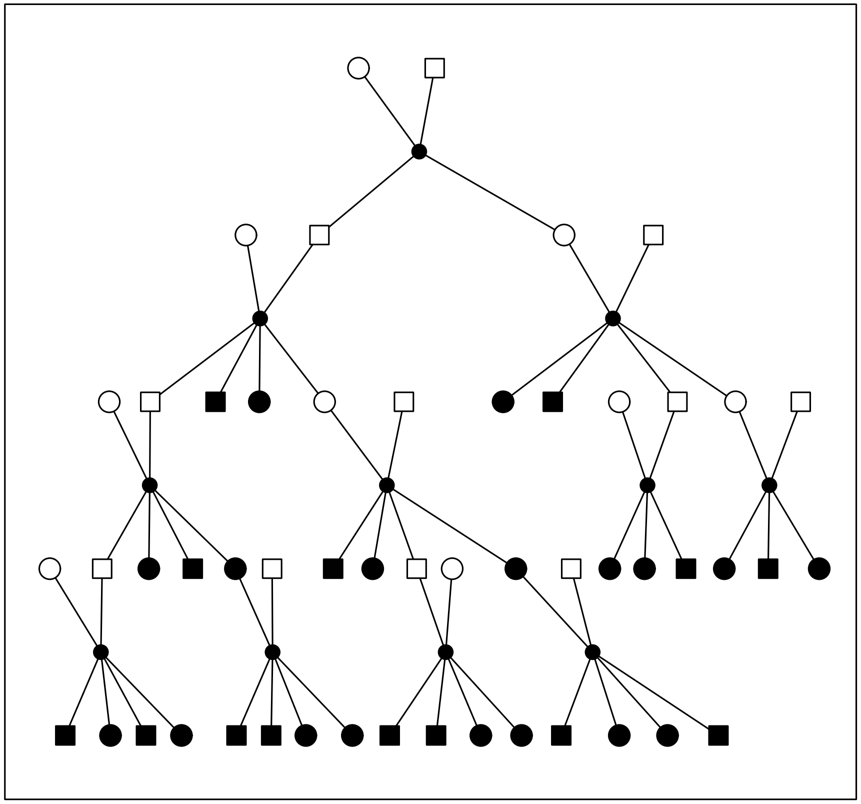
The lod score can be used both for estimation and testing, subject to assumption of a trait model.

- Tests have well-known unresolved issues:
  Assessing statistical significance of a lod score.
  Correcting for testing multiple linked locations (max lod score).

Particularly when applied to extended pedigrees.

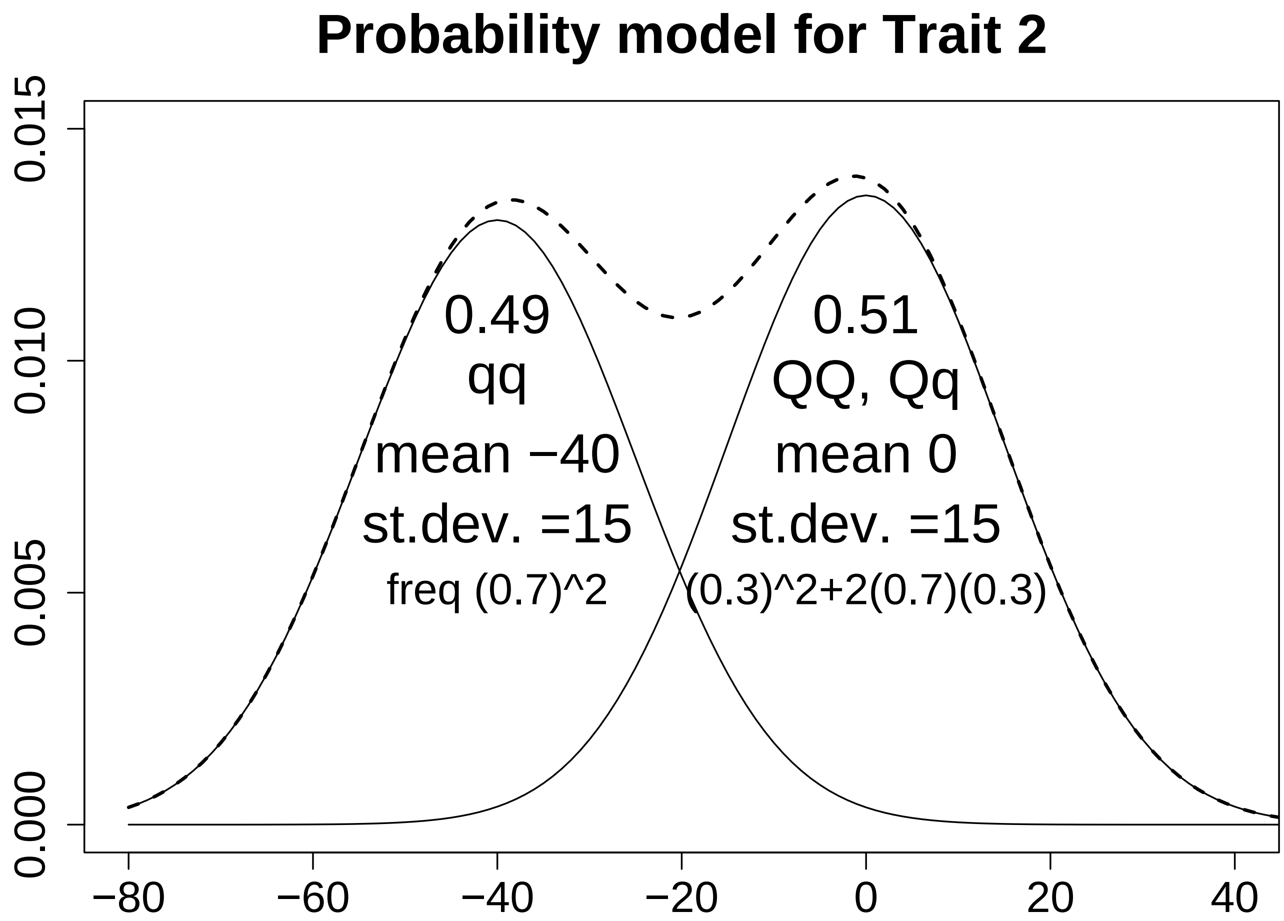
- Goal is to address both these, and also
  Assessing the uncertainty in this inference
  that derives from uncertainty in inheritance of DNA
  (not from map/model misspecification etc.)

Simulated Ped3x52 data used as example



- 3 copies of pedigree:
  each 52 individuals
- On each copy, 32 (shaded) individuals observed for 12 markers, and several quantitative traits.
- Markers spaced evenly at 10cM ($\approx 10^7$bp). Each has 4 alleles, freqs 0.4, 0.3, 0.2, 0.1.
- Locus for Trt2 is midway between M10 and M11.

Quantitative Trt2 simulation model

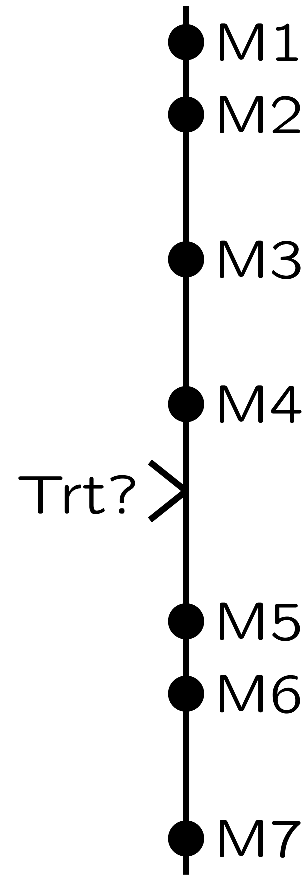


Lod Scores under a given trait model

- The statistic normally used for both testing and estimation when a trait model for trait data Z is assumed is the lod score.
- Z = trait data, $\mathbf{Y}$ = marker data (all markers).
- All parameters of model for Z and $\mathbf{Y}$ assumed known, apart from trait locus position γ.
- Definition: at hypothesized trait locus position γ.

$$\begin{aligned} \mathrm{lod}(\gamma) &= \log_{10}(\mathrm{P}_{\gamma}(Z, \mathbf{Y}) / \mathrm{P}_0(Z, \mathbf{Y})) \\ &= \log_{10}(\mathrm{P}_{\gamma}(Z \mid \mathbf{Y}) / \mathrm{P}(Z)) \end{aligned}$$

where subscript 0 denotes
H_0: independence of Z and $\mathbf{Y}$.



The latent variables of genome inheritance

•**MENDEL's FIRST LAW (1866)**: At meiosis, at each location in the genome, each parent individual segregates a randomly chosen one of its two copies independently to each offspring.

- Specify inheritance by $S_{i,j} = 0$ or 1, $i = 1, ..., m$; $j = 1, ..., l$ as in meiosis i at position j the maternal or paternal DNA (respectively) of the parent is transmitted to the offspring.

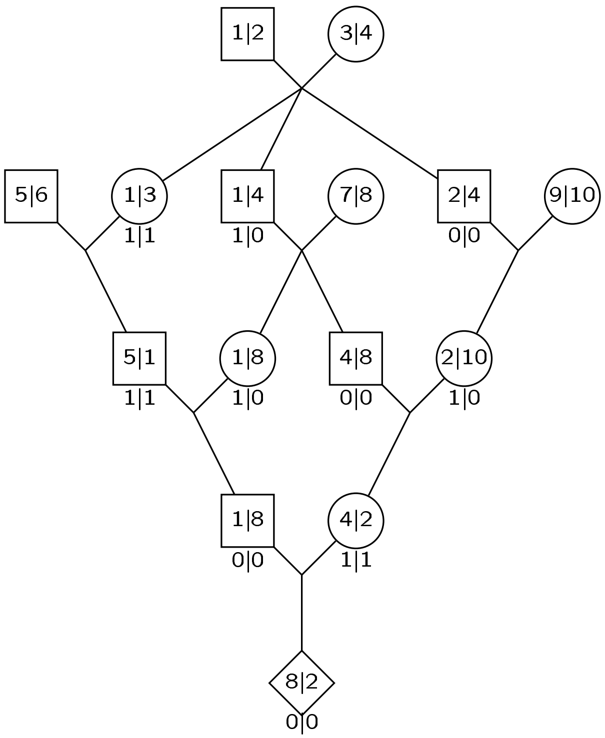
- Mendel's First Law: $\mathrm{P}(S_{i,j} = 0) = \mathrm{P}(S_{i,j} = 1) = 1/2$
  Meioses i are independent: i.e. $S_{i,\bullet} = \{S_{i,j}; j = 1, ..., l\}$.

- At location j, $j = 1, \ldots, l$, $S_{\bullet,j} = \{S_{i,j}; i = 1, \ldots, m\}$, determine the founder origin of the DNA present in each individual, at that location.

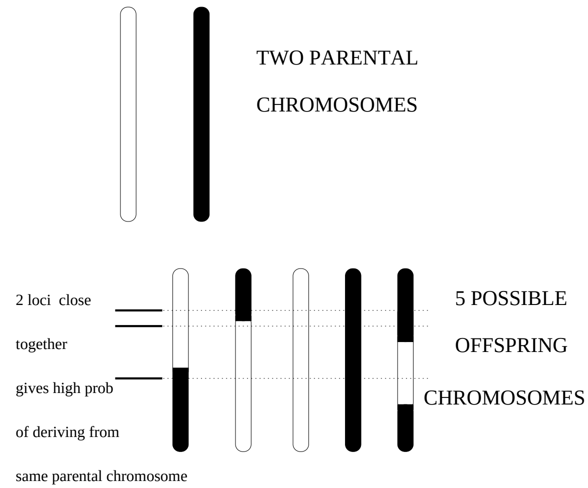
- Dependence in $S_{i,j}$ over j, determined by spacing of locations along the chromosome: close locations $\Rightarrow$ high correlation.

The inheritance of genome: at a locus and over loci

At a locus j:



$S_{\bullet,j}$ specifies inheritance at j



At loci j, j', $\text{P}(S_{i,j} = S_{i,j'})$ decreases as $d(j, j')$ increases.

Tests for linkage look for association in inheritance at specified locations and inheritance of trait phenotypes.

The complete-data case: "observed" $\mathbf{S}$

- Suppose marker data $\mathbf{Y}$ determine $\mathbf{S}$ at marker locations. (In reality, never happens.)

- At hypothesized trait locus position γ, the lod score becomes:

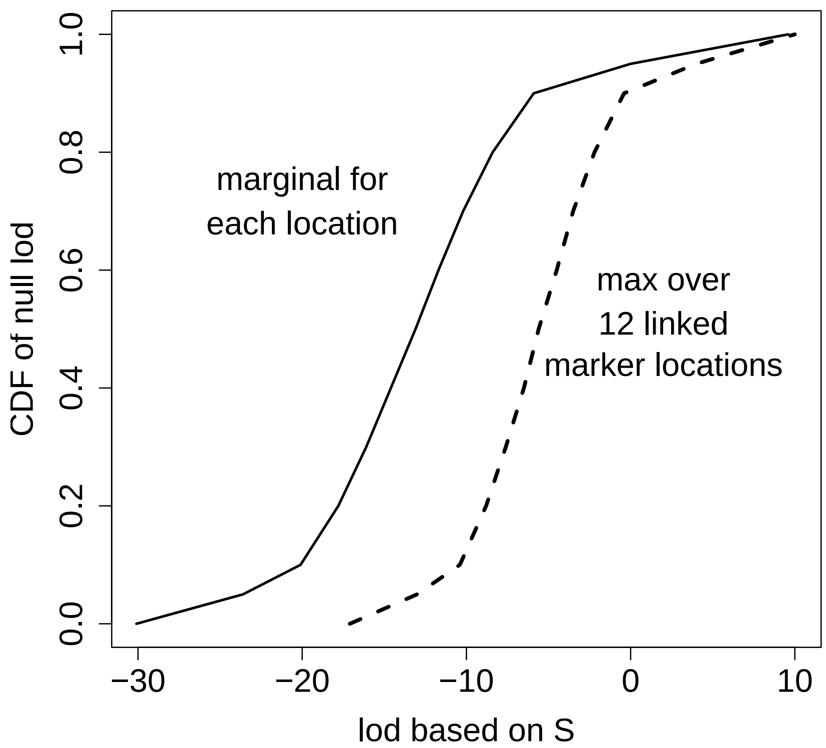
$$\text{lod}(\gamma) \;=\; \log_{10}(\mathsf{P}_\gamma(Z \mid \mathbf{S})/\mathsf{P}(Z))$$

- First, this can be computed, for any γ.

- Second, at marker location j, this lod score depends only on $S_{\bullet,j}$: let $t(S_{\bullet,j})$ be the lod score at marker j location. (Condition on Z, so suppress Z in notation.)

- Third, we can use $t(S_{\bullet,j})$ as a test statistic to test for linkage to marker location j.

Case of observed $S_{\bullet,j}$ at locations $j = 1, ..., 12$



- We can determine a P-value:
- If we observe $t(S_{\bullet,j}) = t_{obs}$:

$$\begin{aligned} p &= \pi(t_{obs}) \\ &= \mathrm{P}_0(t(S_{\bullet,j}) \geq t_{obs}), \end{aligned}$$

where $S_{\bullet,j} \sim \mathrm{P}_0$.

- Simulation of $\mathbf{S}$ under P_0 is trivial.

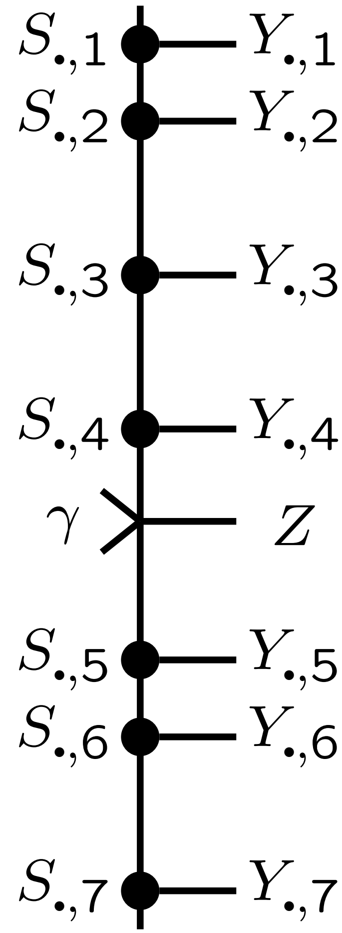
- Omnibus test using maximum lod score:
  Use $t^*(\mathbf{S}) = \max_j(t(S_{\bullet,j}))$.

Back to reality: $\mathbf{S}$ are latent variables

- We observe marker data

$$\mathbf{Y} = \{Y_{\bullet,j}, j = 1, ..., l\}.$$

- The marker data at locus j depends only on the inheritance pattern $S_{\bullet,j}$ at locus j.
- Conditional on $\mathbf{S}$, Z is independent of $\mathbf{Y}$.
- Assuming no genetic interference, the inheritance patterns $S_{\bullet,j}$ are Markov over j.
- This hidden Markov (HMM) structure permits some exact computations, and/or Monte Carlo (MCMC) approaches, for imputing $\mathbf{S}$ conditional on $\mathbf{Y}$



Back to reality: the lod score

- We observe only marker genotypes $\mathbf{Y}$ of some individuals.

- The lod score is

$$\mathrm{lod}(\gamma) = \log_{10}(\mathrm{P}_\gamma(Z \mid \mathbf{Y})/\mathrm{P}(Z))$$

- For multiple markers, on extended pedigrees, $\mathrm{P}_\gamma(Z \mid \mathbf{Y})$ cannot even be computed.

- However, conditional on $\mathbf{S}$, Z is independent of $\mathbf{Y}$. So

$$\begin{aligned} \mathrm{P}_\gamma(Z \mid \mathbf{Y}) &= \sum_{\mathbf{S}} \mathrm{P}_\gamma(Z \mid \mathbf{S})\mathrm{P}(\mathbf{S} \mid \mathbf{Y}) \\ &= \mathrm{E}(\mathrm{P}_\gamma(Z \mid \mathbf{S}) \mid \mathbf{Y}) \end{aligned}$$

Monte Carlo Estimation of the lod score

- On small pedigrees:
  We can compute $\mathsf{P}(S_{\bullet,j} \mid \mathbf{Y})$ or
  we can i.i.d. sample $\mathbf{S}$ from $\mathsf{P}(\mathbf{S} \mid \mathbf{Y})$.

- On large pedigrees, we cannot compute exactly, but
  we can MCMC sample $\mathbf{S} = \{S_{i,j}\}$ from $\mathsf{P}(\mathbf{S} \mid \mathbf{Y})$.

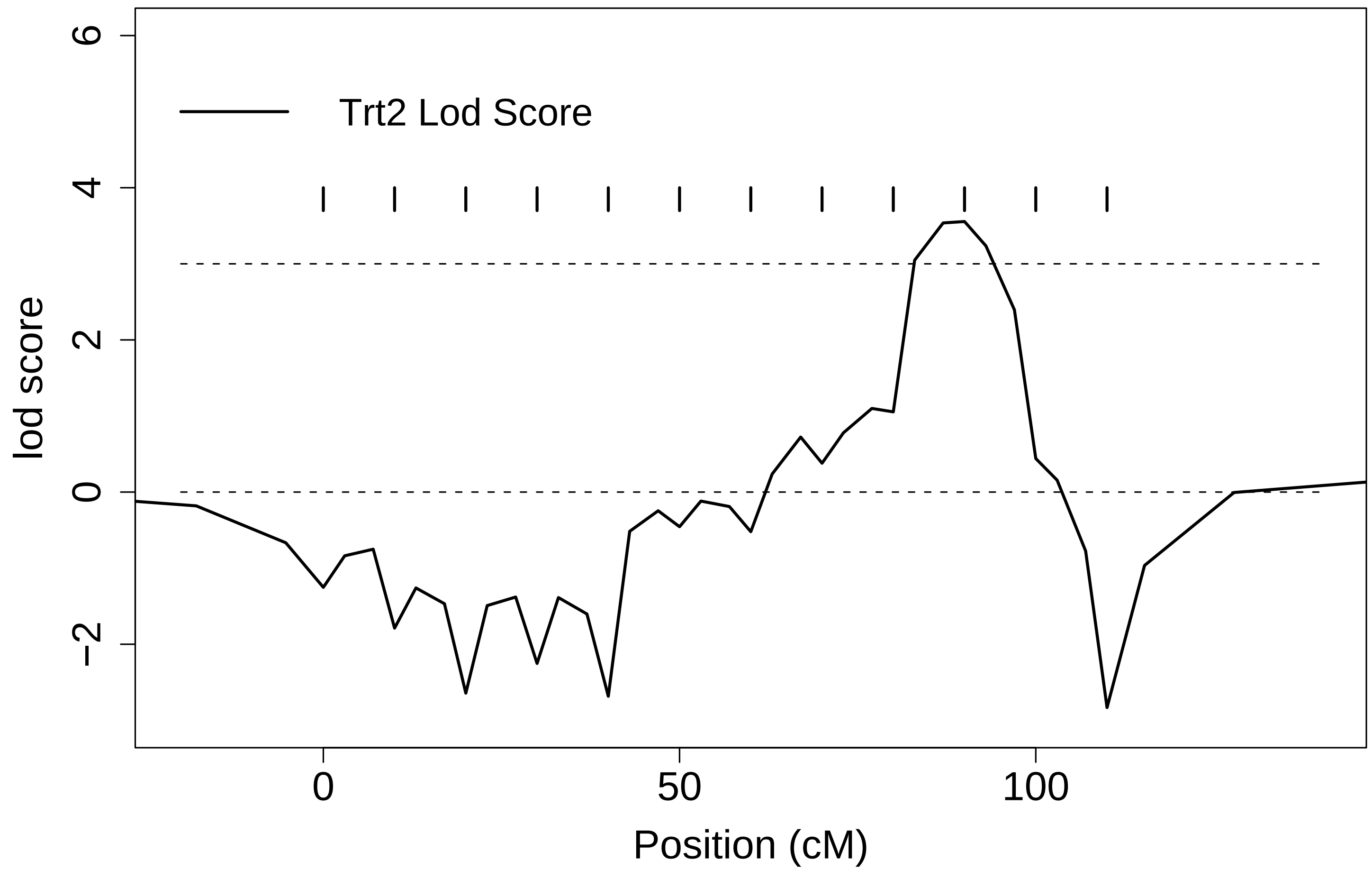
- Consider set of n realizations $\mathbf{S}^{(\ell)}$ from $\mathsf{P}(\mathbf{S} \mid \mathbf{Y})$:

$\mathsf{P}_\gamma(Z \mid \mathbf{Y}) = \mathsf{E}(\mathsf{P}_\gamma(\mathbf{Z} \mid \mathbf{S}) \mid \mathbf{Y})$, can be estimated by

$$n^{-1} \sum_{\ell=1}^{n} \mathsf{P}_\gamma(Z \mid \mathbf{S}^{(\ell)}).$$

- Hence the full lod score curve (over γ) can be estimated from one set of (MCMC) realizations from $\mathsf{P}(\mathbf{S} \mid \mathbf{Y})$.

Lod score for location γ of Trt2



This reaches the value 3!! What does this mean??

Assessing significance: the classical approach

- What is the significance of a lod score of 3?
  What is the uncertainty, due to uncertainty in $\mathbf{S}$?
  How do we adjust for multiple testing;
  that is, for using the maximum lod score?

- Given some statistic $W(\mathbf{Y})$ (here the lod score), only some form of simulation will provide the p-value for a test based on the values of $W(\mathbf{Y})$.

(Again, condition on Z: omit Z from $W()$.)

- That is, repeat the entire process for datasets $\mathbf{Y}^{(k)}$ resimulated under the null hypothesis of no trait linkage.

- If $k = 1, ..., N$, N large,

$p = (N+1)^{-1}(1 + \sum_{k=1}^{N} I(W(\mathbf{Y}^{(k)}) \geq W(\mathbf{Y})))$.

Disadvantages of the standard approach

- Computationally very intensive: N large (~ 500?).
  —MCMC for each resimulated $\mathbf{Y}^{(k)}$.

- Parameters (allele freqs) for resimulation of marker data $\mathbf{Y}^{(k)}$?? Even harder if resimulate trait data Z – trait model? ascertainment??

- MCMC gives an estimate the distribution of $t(\mathbf{S})$ given $\mathbf{Y}$: here $t(\mathbf{S})$ is the complete-data lod score (at γ or max). What a waste of information to use the MCMC only to sum over $\mathbf{S}$ to estimate $W(\mathbf{Y})$ (the lod score, or max lod score).

- We know (almost) nothing about the distribution of $W(\mathbf{Y})$, but (almost) everything about the distribution of $t(\mathbf{S})$ given $\mathbf{Y}$.

- Information that $\mathbf{Y}$ provides about $t(\mathbf{S})$ is confounded with the evidence $t(\mathbf{S})$ provides about H_0.

A Fuzzy P-Value

- Definition (Geyer & Meeden, 2005): A r.v. with the distrib. of $(Q|\mathbf{Y})$, where Q is U(0,1) (unconditionally) under H_0.

Then E(P($Q \leq \alpha|\mathbf{Y}$)) $= \alpha$ where E() is over $\mathbf{Y}$ under H_0.
i.e. under H_0, the fuzzy p-value has a U(0,1) distribution.

- Let $\pi(\mathbf{S}) =$ P$(t(\mathbf{S}_0) > t(\mathbf{S})|\mathbf{S}) \sim U(0,1)$ under H_0.
So E(P($\pi(\mathbf{S}) \leq \alpha)|\ \mathbf{Y}) = \alpha$ where E() is over $\mathbf{Y}$ under H_0.
A r.v. with the distribution of $\pi(\mathbf{S})$ given $\mathbf{Y}$ is a fuzzy p-value.

- Now $\pi(\mathbf{S}) =$ P$(t(\mathbf{S}_0) > t(\mathbf{S})|\mathbf{S})\ =\ $ P$(t(\mathbf{S}_0) > t(\mathbf{S})|\mathbf{S},\mathbf{Y})$.
So let $\mathbf{S}_0^{(h)}, h = 1, ..., m\ \sim$ P$_0$, and $\mathbf{S}^{(\ell)},\ \ell = 1, ..., n, \sim$ P$(\cdot \mid \mathbf{Y})$:
Then $\eta(\mathbf{S}^{(\ell)},\ \mathbf{Y}) =$ P$(\mathbf{t}(\mathbf{S_0}) > \mathbf{t}(\mathbf{S}^{(\ell)})|\mathbf{S}^{(\ell)},\ \mathbf{Y}), \qquad \ell = 1, ..., n$
estimated by $m^{-1} \sum_{h=1}^{m} I(t(\mathbf{S}_0^{(h)}) > t(\mathbf{S}^{(\ell)}))$,
gives n realizations from the fuzzy p-value dsn.

- Discreteness can be dealt with exactly (C. J. Geyer).

Fuzzy p-values for lod scores

- Use the lod score were $\mathbf{S}$ observable

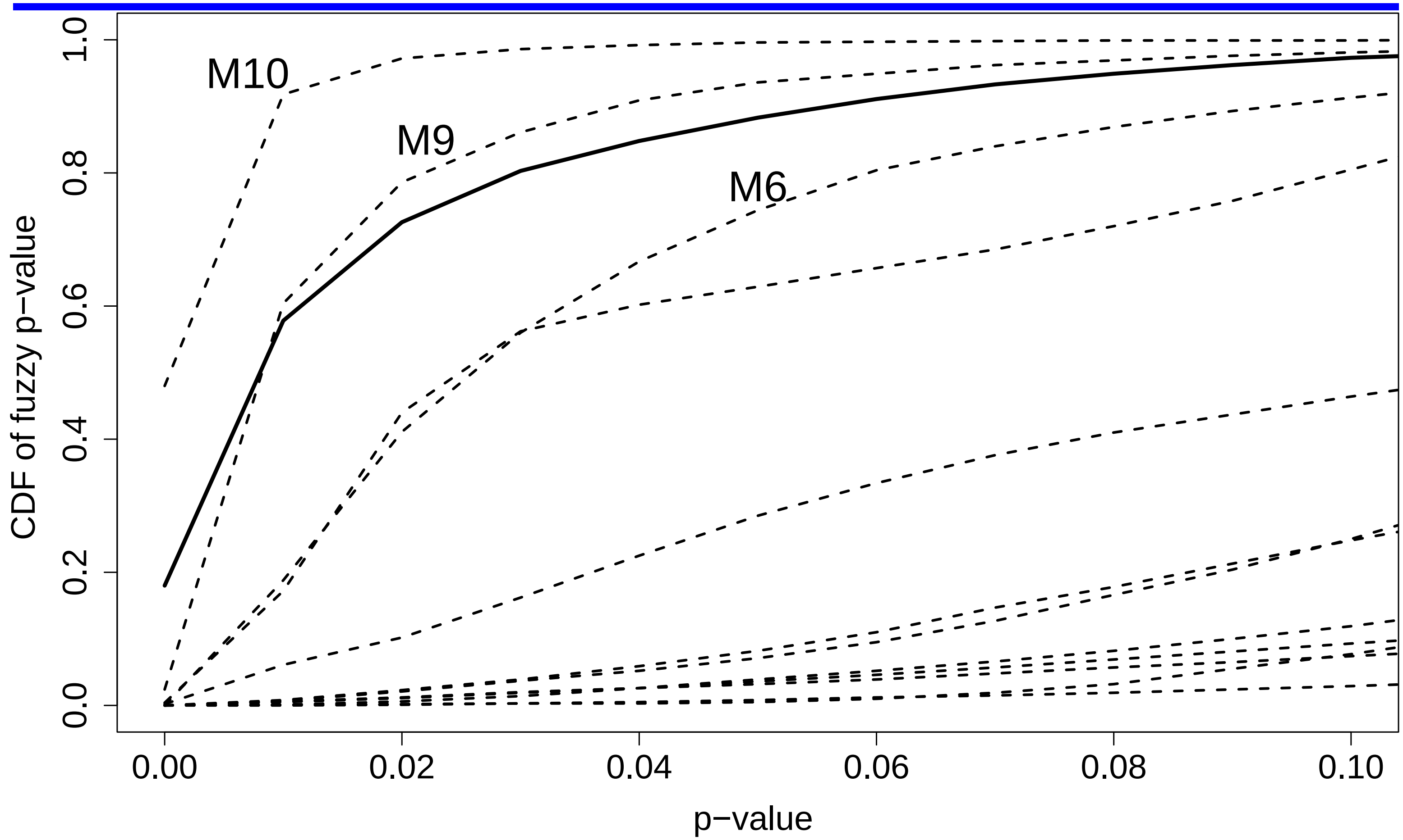
$$t_\gamma(\mathbf{S}) \;=\; \log_{10}\left(\mathrm{P}_\gamma(Z \mid \mathbf{S}) / \mathrm{P}(Z)\right)$$

for each location γ, and compute the fuzzy p-value both point-wise and adjusted for multiple testing (max over markers).

- We already have the (MCMC) realizations from $\mathrm{P}(\mathbf{S} \mid \mathbf{Y})$.
  We already compute $t_\gamma(\mathbf{S})$ (or $\mathrm{P}_\gamma(Z \mid \mathbf{S})$)
  in computing the MCMC estimate of the lod score!!

- The fuzzy p-value CDF measures both strength of evidence, and uncertainty, putting the uncertainty onto the p-value scale.

Linkage detection from lod scores at markers



Strong evidence for linkage at marker 10: $\mathrm{P}(\pi(\mathbf{S}) \leq 0.05 \mid \mathbf{Y}) = 0.98$.

Less strong when adjusted: $\mathrm{P}(\pi^*(\mathbf{S}) \leq 0.05 \mid \mathbf{Y}) \approx 0.85$.

Advantages of the fuzzy p-value

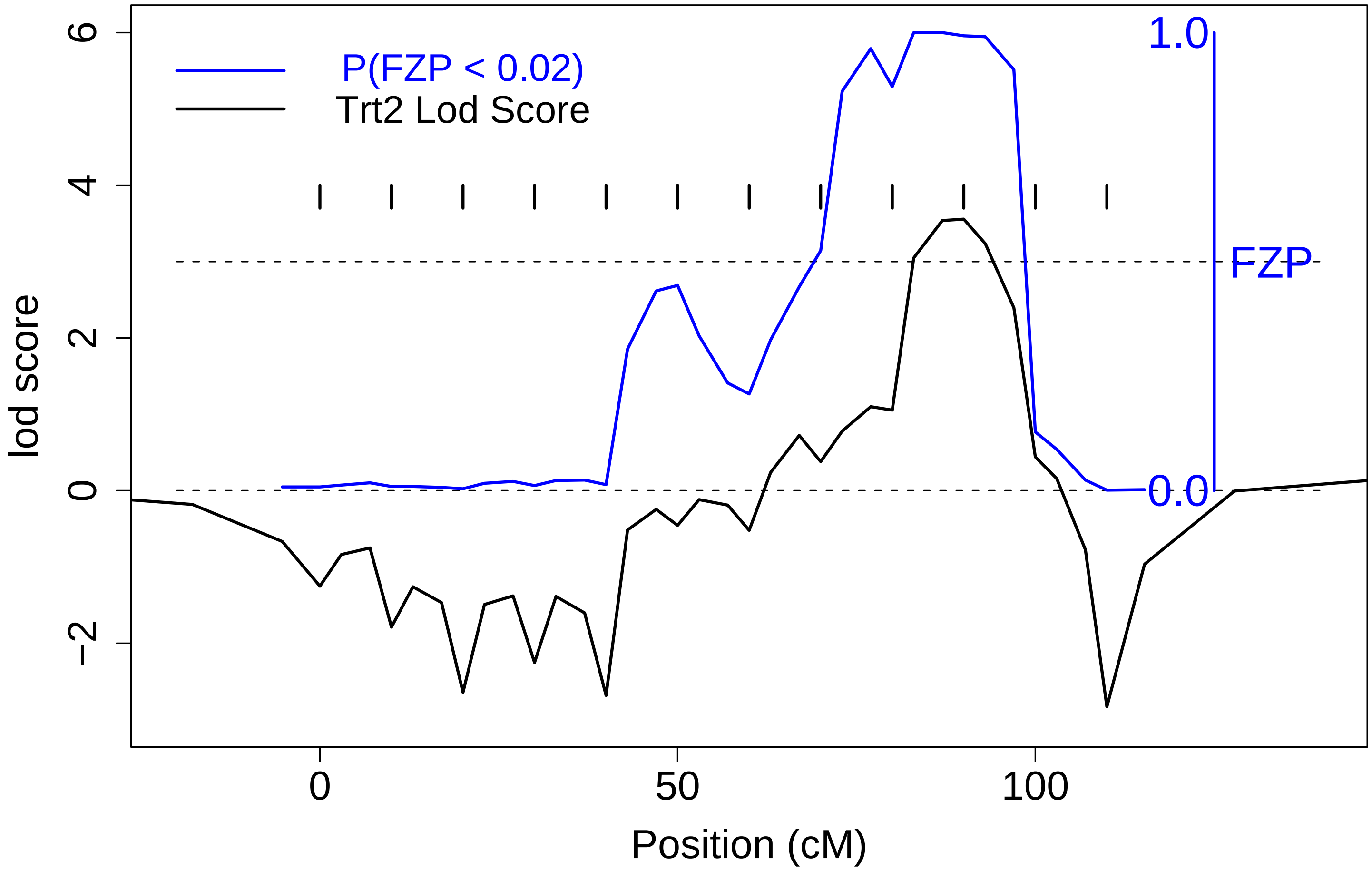
- Can be easily estimated from two Monte Carlo samples (one unconditional, and one conditional on $\mathbf{Y}$).

- Does not require resimulation of data $\mathbf{Y}$ (or Z), which is both a computational and a statistical (robustness) advantage.

- Provides a valid p-value, including any correction desired for testing at multiple linked markers.

- Separates the uncertainty about $t(\mathbf{S})$ from the evidence in $t(\mathbf{S})$.

Pointwise lod-based fuzzy p-values for Trt2



This is not a 98% fuzzy confidence set.

Fuzzy confidence intervals, after inferring linkage

- To construct a confidence interval for γ we need a test of
  H_γ :   trait location is γ, for each γ.

(Note, under H_γ, Z and $\mathbf{S}$ at markers are not independent.)

- Given $\mathbf{S}$ at markers, reject H_γ if
  $t_\gamma(\mathbf{S}) = -\log(\mathsf{P}_\gamma(Z|\mathbf{S}) / \sup_{\gamma^*} \mathsf{P}_{\gamma^*}(Z|\mathbf{S}))$ too large.

- Now, as before, we realize $\mathbf{S}$ both conditional only on Z (easy) and also given the marker data $\mathbf{Y}$ and Z, under H_γ.

- The latter can be done using MCMC to sample conditionally on $\mathbf{Y}$ and importance sampling reweighting to condition on Z.

- In principle, this works — the program runs.
  Details of performance remain to be worked out.

CONCLUSION

• It is latent inheritance patterns $\mathbf{S}$ that provides evidence for genetic hypotheses such as linkage, but marker data $\mathbf{Y}$ are a very imperfect reflection of $\mathbf{S}$.

• Basing linkage tests and estimates on lod scores computed from data $\mathbf{Y}$ is very computationally intensive, requires detailed marker model, and raises unsolved multiple testing issues.

• Evidence in $\mathbf{S}$ is confounded with uncertainty about $\mathbf{S}$.

• Fuzzy p-values address these issues, putting uncertainty in $\mathbf{S}$ directly on evidence scale.

• Fuzzy p-values can be applied to any test statistic. However, using the lod score has the advantage that, in principle, estimation (i.e. confidence intervals) can also be addressed.