Pseudo-Bayes MCMC for the estimation of multipoint linkage likelihoods

Elizabeth Thompson

University of Washington

Research supported in part by NIH grant GM-46255.

Parts of this work are joint with Dr. Andrew George.

Thanks for use of data to Drs. Bird, Schellenberg, Wijsman.

The genetic mapping problem

•Given: L genetic markers at known locations λ_i in the genome, and known allele frequencies $\mathbf{q_i}$, $i = 1, ..., L$. $\Lambda_M = \{\lambda_i, \mathbf{q_i}\}$.

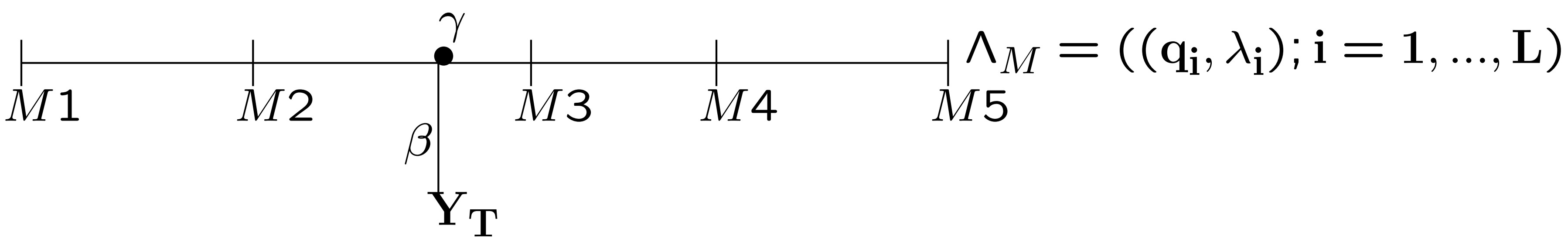
•Given: a trait, and a presumed trait model, parametrized by β, specifying how trait is determined by underlying genes.

•Given: data on the trait phenotypes and marker genotypes for some of the members of some number of pedigree structures.

•Estimate: the location γ of a locus affecting the trait, in some region of the genome.

•Approach: compute a likelihood and hence a location lod score.

What and why the LOCATION LOD score

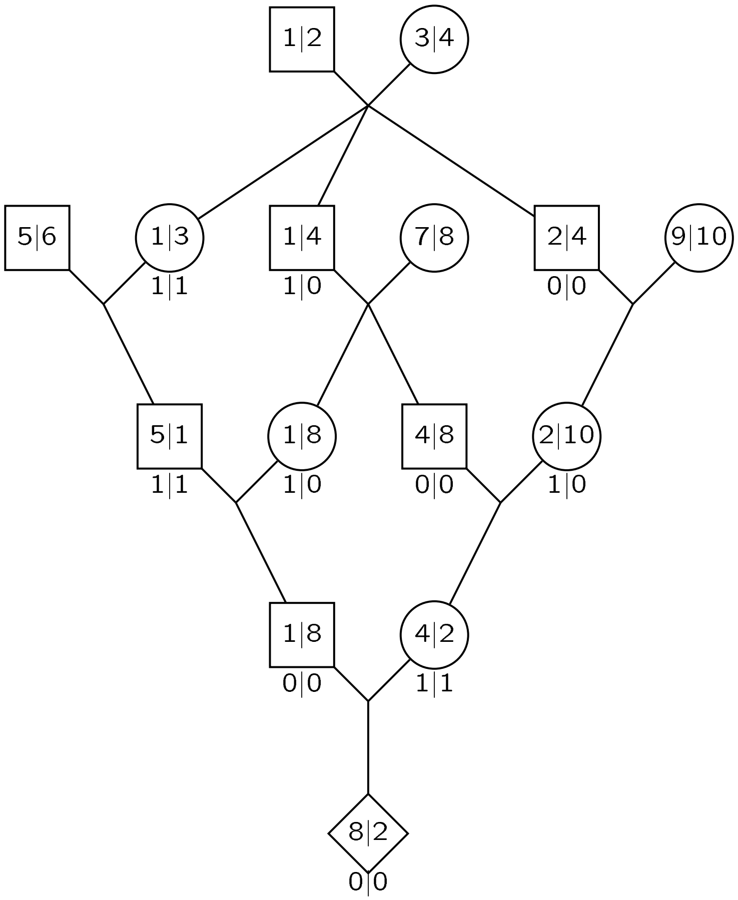


Parameter $\xi = (\beta, \gamma, \Lambda_M)$. Data $\mathbf{Y} = (\mathbf{Y_M}, \mathbf{Y_T})$

$$\text{lod}(\gamma) \;=\; \log_{10}\left(\frac{\Pr(\mathbf{Y}; \mathbf{\Lambda_M}, \beta, \gamma)}{\Pr(\mathbf{Y}; \mathbf{\Lambda_M}, \beta, \gamma = \infty)}\right)$$

Exact computation is infeasible

The Inheritance of genes and genome



Label the two haploid genomes of every founder: Founder genome labels (FGL).

Inheritance of FGL:

$$S_{i,j} \;=\; 0 \;\text{ or }\; 1$$

as in meiosis i at locus j the maternal or paternal gene (respectively) of the parent is transmitted to the offspring.

Basics of genetics: for statisticians

- Meioses i are independent: $S_{i,\bullet}$ are independent, a priori.
- Mendel's First Law: $\text{Pr}(S_{i,j} = 0) \;=\; \text{Pr}(S_{i,j} = 1) \;=\; 1/2$
- Recombination: $\text{Pr}(S_{i,j-1} \neq S_{i,j}) = \rho_{j-1}$ ($\forall\ i$ for convenience)

$$\text{Pr}(S_{\bullet,j} \mid S_{\bullet,j-1}) \;=\; \rho_{j-1}^{R_{j-1}}(1-\rho_{j-1})^{m-R_{j-1}}$$

where $R_{j-1} = (\#i : S_{i,j} \neq S_{i,j-1})$

- No genetic interference: $\text{Pr}(S_{i,j}|\mathbf{S}_{-(i,j)}) = \text{Pr}(S_{i,j}|S_{i,j-1}, S_{i,j+1})$

$$\text{Pr}(\mathbf{S}) \;=\; P(S_{\bullet,1})\prod_{2}^{L}\text{Pr}(S_{\bullet,j} \mid S_{\bullet,j-1})$$

Sampling and computation

The likelihood is

$$L(\xi) \;=\; P_\xi(\mathbf{Y}) \;=\; \sum_{\mathbf{S}} P_\xi(\mathbf{S},\mathbf{Y}\;) \;=\; \sum_{\mathbf{S}} \mathbf{P}_\xi(\mathbf{Y} \mid \mathbf{S})\; \mathbf{P}_\xi(\mathbf{S})$$

$$P_\xi(\mathbf{S},\mathbf{Y}\;) \;=\; \mathrm{Pr}(S_{\bullet,1}) \prod_{j=2}^{L} \mathrm{Pr}(S_{\bullet,j} \mid S_{\bullet,j-1}) \prod_{j=1}^{L} \mathrm{Pr}(Y_{\bullet,j} \mid S_{\bullet,j})$$

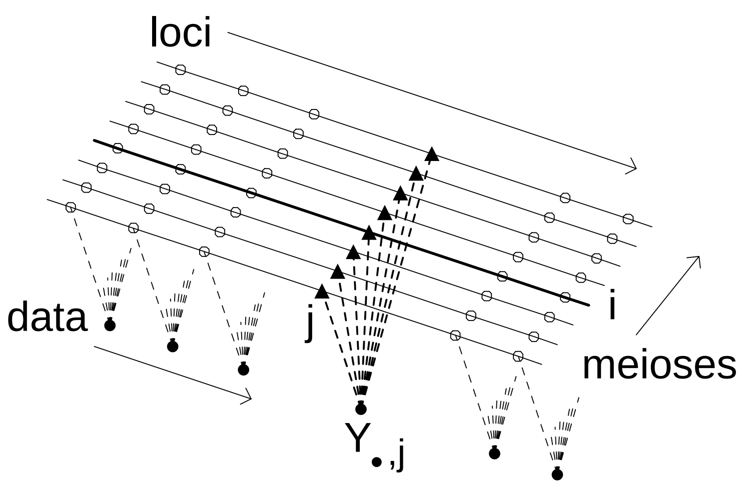
On small pedigrees, or for few loci, we can compute $\mathrm{Pr}(\mathbf{Y})$

Then we can compute $\mathrm{Pr}(S_{\bullet,j} \mid \mathbf{Y})$, for each j.

On larger pedigrees, we cannot compute, but

we can SAMPLE $\mathbf{S} = \{S_{i,j}\}$ from $\mathrm{Pr}(\mathbf{S} \mid \mathbf{Y})$. (joint **S**)

Block-Gibbs MCMC Samplers



L-sampler: resample $S_{\bullet,j}$ given $\mathbf{Y}$ and $S_{\bullet,j'}, j \neq j'$

M-sampler: resample $\{S_{i,\bullet}; i \in I^*\}$ given $\mathbf{Y}$ and $\{S_{i',\bullet}; i' \notin I^*\}$

LM-sampler: Heath (1997), Thompson & Heath (1999)

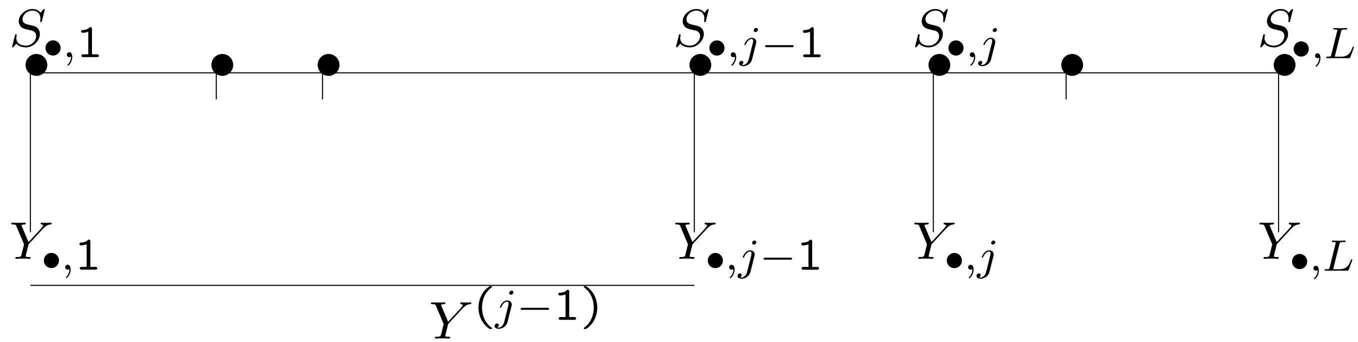
Previous estimators of the lod score

Lange-Sobel (1991)

$$
\begin{aligned}
L(\beta,\gamma,\Lambda_M) &= P_{\beta,\gamma,\Lambda_M}(\mathbf{Y_M},\mathbf{Y_T}) \\
&\propto P_{\beta,\gamma,\Lambda_M}(\mathbf{Y_T} \mid \mathbf{Y_M}) \\
&= \sum_{\mathbf{S}_M} P_{\beta,\gamma}(\mathbf{Y_T} \mid \mathbf{S_M})\mathbf{P}_{\mathbf{\Lambda_M}}(\mathbf{S_M} \mid \mathbf{Y_M}) \\
&= \mathsf{E}_{\Lambda_M}(P_{\beta,\gamma}(\mathbf{Y_T} \mid \mathbf{S_M}) \mid \mathbf{Y_M}).
\end{aligned}
$$

Advantages; sample only S_M and compute over S_T
(but for each γ) – a Rao-Blackwellized estimate.
Disadvantages: (1) sample only given $\mathbf{Y_M}$,
(2)sampling is MCMC.

Sequential imputation



Irwin, Kong et al. (1994)

$$
\begin{aligned}
P^*(S_{\bullet,j}) &= P_{\xi_0}(S_{\bullet,j} \mid S^{*(j-1)}, Y^{(j)}) &= P_{\xi_0}(S_{\bullet,j} \mid S^*_{\bullet,j-1}, Y_{\bullet,j}) \\
w_j &= P_{\xi_0}(Y_{\bullet,j} \mid Y^{(j-1)}, S^{*(j-1)}) &= P_{\xi_0}(Y_{\bullet,j} \mid S^*_{\bullet,j-1}) \\
P^*(\mathbf{S}^*) &= \frac{P_{\xi_0}(\mathbf{S}^*, \mathbf{Y})}{\prod_{j=1}^{L} w_j} \text{ so } L(\xi_0) = \sum_{\mathbf{S}} P_{\xi_0}(\mathbf{S}, \mathbf{Y}) &= \mathsf{E}_{\mathbf{P}^*}\left(\prod_{\mathbf{j=1}}^{\mathbf{L}} \mathbf{w_j}\right)
\end{aligned}
$$

Adv: i.i.d sampling. Disadv: P^* may be far from $P_{\xi_0}(\mathbf{S}|\mathbf{Y})$

Likelihood ratio estimation

Thompson, Guo (1991)

$$\frac{L(\xi)}{L(\xi_0)} = \frac{P_\xi(\mathbf{Y})}{P_{\xi_0}(\mathbf{Y})} = \mathsf{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y},\mathbf{S})}{P_{\xi_0}(\mathbf{Y},\mathbf{S})} \mid \mathbf{Y}\right)$$

$$\frac{L(\beta,\ \gamma_1,\ \Lambda_M)}{L(\beta,\ \gamma_0,\ \Lambda_M)} = \mathsf{E}_{\xi_0}\left(\frac{P_{\xi_1}(\mathbf{Y_T},\mathbf{Y_M},\mathbf{S_T},\mathbf{S_M})}{P_{\xi_0}(\mathbf{Y_T},\mathbf{Y_M},\mathbf{S_T},\mathbf{S_M})} \mid \mathbf{Y_T},\mathbf{Y_M}\right)$$

$$= \mathsf{E}_{\xi_0}\left(\frac{P_{\gamma_1}(\mathbf{S}_T \mid \mathbf{S}_M)}{P_{\gamma_0}(\mathbf{S}_T \mid \mathbf{S}_M)} \mid \mathbf{Y_T},\mathbf{Y_M}\right)$$

for two hypothesized trait locus positions γ_1 and γ_0.

Advantage: Actual estimate is simple: fast and accurate for local LR

Disadvantage: Need good MCMC. Works well only for $\gamma_1 \approx \gamma_0$: combining local LR estimates is hard. We want $L(\gamma)/L(\gamma = \infty)$.

Monte Carlo likelihood/posterior estimates

- **Lange-Sobel (1991)** : MCMC likelihood estimator
  MCMC sampling of $\mathbf{S}_M$ given $\mathbf{Y_M}$.

- **Irwin, Kong et al. (1994)** : Sequential imputation
  likelihood estimator – i.i.d. sample: importance sampling.

- **Thompson, Guo (1991)** : local likelihood ratio
  estimation– MCMC sampling of $(\mathbf{S}_M, S_T \mid \mathbf{Y_M}, \mathbf{Y_T})$

- **Heath (1997)** and others : Fully Bayesian MCMC
  approaches– sample γ, β, $\mathbf{q_i}$ etc. etc.

Problems of a fully Bayesian approach

A Bayesian approach (e.g. Loki:Heath), puts priors on (β, γ) and samples from $\pi_{\Lambda_M}(\beta, \gamma, \mathbf{S} \mid \mathbf{Y})$.

Four problems (from a likelihood perspective):

- (i) β is mixed up in the estimate. lod score should not be based on integrated likelihood. (Note β typically multidimensional.)

- (ii) γ is continuous (typically binned), but likelihood is pointwise function of γ

- (iii) sampling low-prob areas is hard (e.g. unlinked?!)

- (iv) Moving between equal probability areas can be hard (e.g. unlinked?!)

From Bayes back to lods

- (i) First we fix $\theta = (\Lambda_M, \beta)$. ($\xi = (\theta, \gamma)$)

- (ii) For single parameter γ

$$\pi_\theta(\gamma|\mathbf{Y}) \;\propto\; P_\theta(\mathbf{Y};\gamma)\,\pi(\gamma) \quad \text{so} \quad \mathbf{L}(\gamma) \;\propto\; \pi_\theta(\gamma|\mathbf{Y})/\pi(\gamma)$$

- (iii) discretize γ – to get $L(\gamma)$ at discrete points

- (iv) ALSO $\pi(\gamma)$ is arbitrary – choose it to improve estimate – it is a pseudo-prior

- (v) Choose it so that the posterioris approximately uniform

How to sample γ and $\mathbf{S}$ from posterior

- For $(\mathbf{S}_M, S_T)$, use LM-sampler (block Gibbs) as before

- For γ use M-H proposal γ^* based only on $\mathbf{S}_M$ (not S_T)
Update S_T given $(\gamma^*, \mathbf{S}_M)$ for new γ^*: joint update of $(\gamma, \mathbf{S}_T)$.

- Sequential imputation start-up and restarts.

- Preliminary run provides $\pi(\gamma)$ such that posterior $\approx$ uniform.

- And we use Rao-Blackwellized estimators.

Rao-Blackwellized Estimators from pseudo-Bayes

Suppose we have realizations $(\gamma^{(n)}, \mathbf{S}^{(n)})$ from the posterior given $\mathbf{Y} = (\mathbf{Y_M}, \mathbf{Y_T})$.

$$\text{Crude estimator}: \quad \widehat{L(\gamma)}_1 = N^{-1} \sum_{\tau=1}^{N} I(\gamma^{(\tau)} = \gamma)/\pi(\gamma)$$

$$\text{Better estimator}: \quad \widehat{L(\gamma)}_2 = N^{-1} \sum_{\tau=1}^{N} h(\mathbf{S}_M^{(\tau)}, \gamma)$$

where

$$h(\mathbf{S}_M, \gamma) = E_{\pi_\theta}\left(\frac{I(\gamma)}{\pi(\gamma)} \middle| \mathbf{S}_M, \mathbf{Y} \right)$$

Crude estimator is function of realized $\gamma^{(\tau)}$.
RB-estimator is function of realized $\mathbf{S}_M$.

Now compute this!

$$h(\mathbf{S}_M, \gamma) \;=\; E_{\pi_\theta}\left(\frac{I(\gamma)}{\pi(\gamma)}\middle|\, \mathbf{S}_M, \mathbf{Y}\right) \;=\; \frac{P_\theta(\gamma, \mid \mathbf{S}_M, \mathbf{Y}_M, Y_T)}{\pi(\gamma)}$$

$$= \frac{P_\theta(Y_T \mid \mathbf{S}_M, \mathbf{Y}_M, \gamma) P_\theta(\mathbf{S}_M, \mathbf{Y_M}) \pi(\gamma)}{\pi(\gamma) \sum_{\gamma^*} P_\theta(Y_T \mid \mathbf{S}_M, \mathbf{Y}_M, \gamma^*) P_\theta(\mathbf{S}_M, \mathbf{Y}_M) \pi(\gamma^*)}$$

$$= \frac{P_\theta(Y_T \mid \mathbf{S}_M, \gamma)}{\sum_{\gamma^*} P_\theta(Y_T \mid \mathbf{S}_M, \gamma^*) \pi(\gamma^*)}$$

At given $\mathbf{S}_M$ compute for each γ.

Compare this to the Lange estimate!

—similar integration over S_T given realized $\mathbf{S}_M$.

—different in that sampling is of $(\mathbf{S}_M, \gamma)$ given $(\mathbf{Y_M}, \mathbf{Y_T})$ at given β.

Early-onset Alzheimer's diesease in the VG group

- Relatively late onset
  - many unobserved pedigree members
  - younger members uninformative
- Not all VG EOAD pedigrees segregate PS2 on Chr 1.
- There are affected individuals not carrying PS2.
- There are unaffected individuals carrying PS2
  - including older individuals.
- Many characteristics of a complex trait.

Pedigree data summary

Family data			AD data				Marker data
Pedigree	Size	Gen	Aff	Unaff	Unobs	Onset	No.obsvd
HB	50	6	13	28	9	60.6	27
HD	41	5	14	17	10	52.2	14
R	53	4	17	30	6	50.8	31
KS	53	5	11	36	6	65.5	27
WFL	21	3	6	14	1	63.8	15
W	6	2	4	2	0	59.8	4

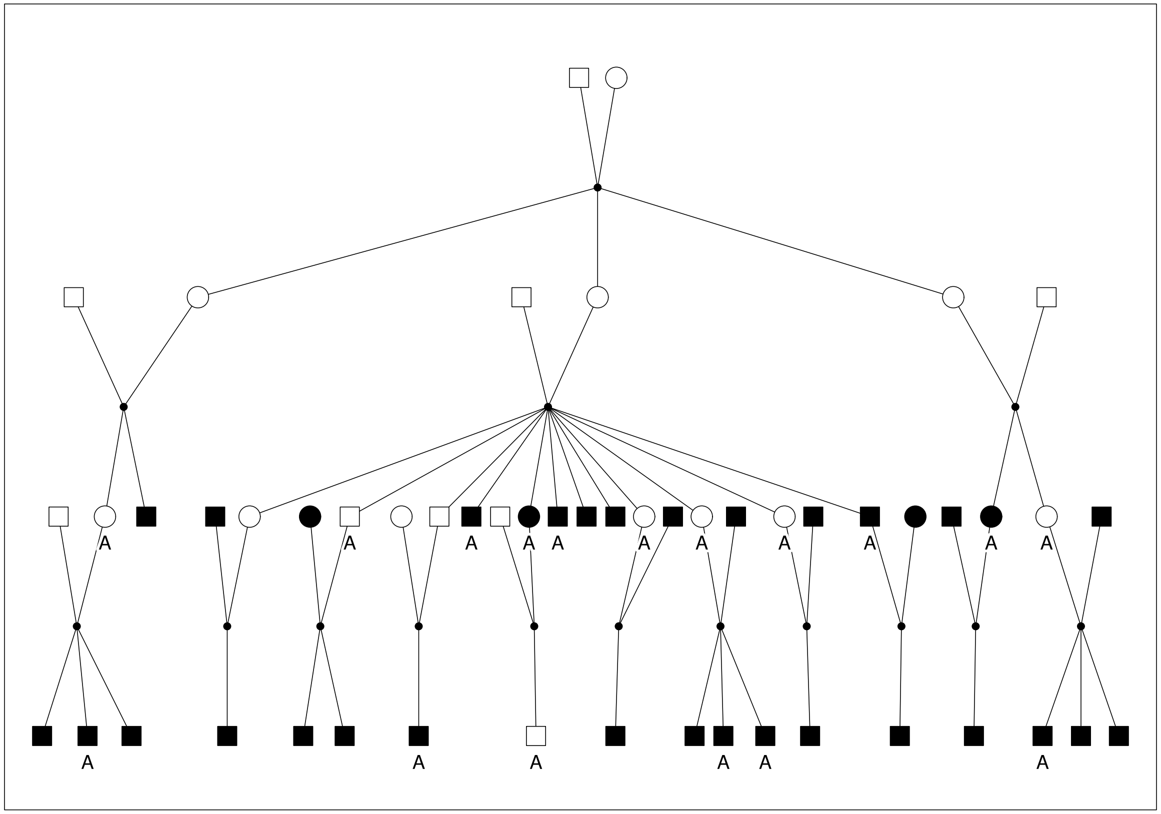
Acknowledge: Drs. Bird, Schellenberg, & Wijsman.

Marker data summary

Index	Marker	Map Position (cM)	Number of Alleles
1	D1S306	0.00	12
2	D1S249	5.48	15
3	D1S245	12.64	10
4	D1S237	17.64	13
5	D1S229	22.56	8
6*	D1S479	27.17	11
7	D1S446	36.95	13
8	D1S235	39.47	9
9	D1S180	52.34	11
10	D1S102	60.51	6

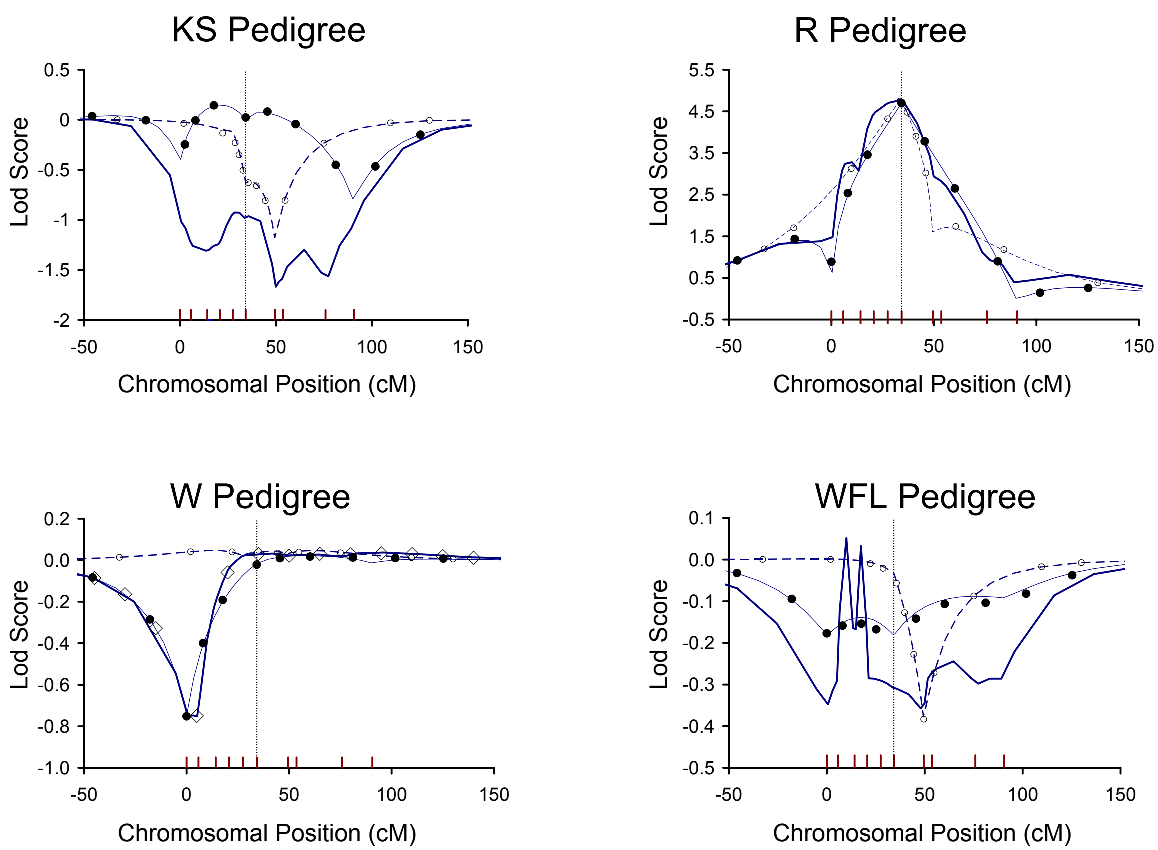
Example pedigree: approximate

SIMPED: disease status and marker availability



Gender, trait, and marker info are altered for confidentiality.

Does it work 1?– lod score estimates



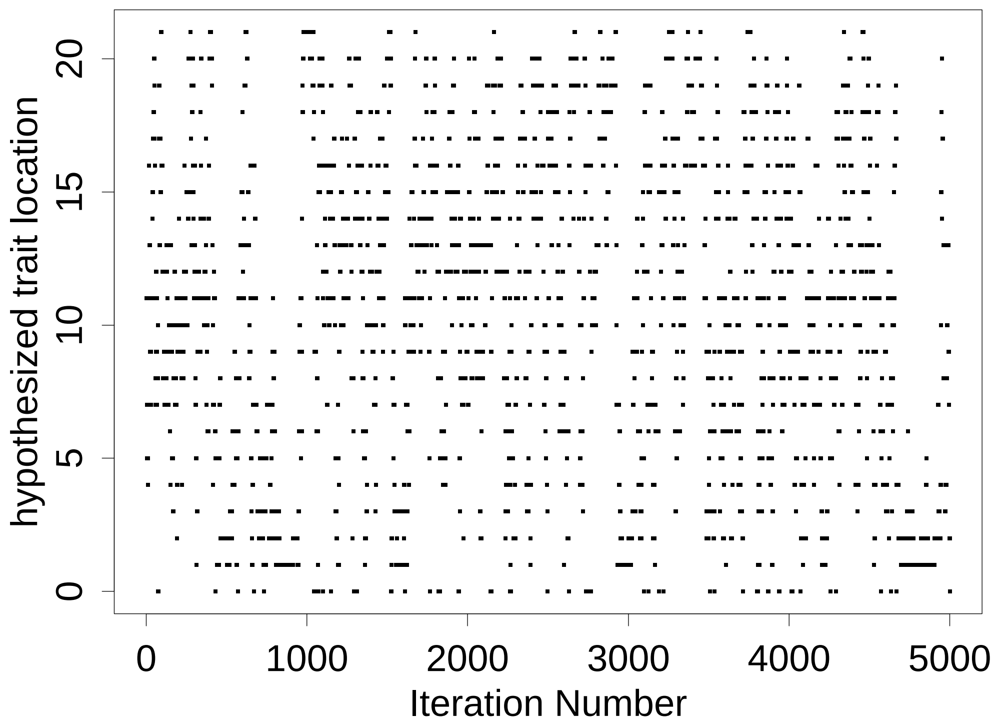
Does it work 2?– Run-time comparisons

Pedigree	MS-L Bayes length	MS-L Bayes time	MS-L VSSE time	MS-T Bayes length	MS-T Bayes time	MS-T VSSE time
KS	10:20	12.8	292.9	8:20	15.0	1156.8
R	1.5:3	2.7	62.0	3:7	4.9	41.0
W	0.2:0.4	0.1	0.1	0.2:0.5	0.1	0.1
WFL	2:4	0.8	0.6	2:5	1.0	0.3

Pedigree	MS-A length	MS-A time
KS	50:100	90.5
R	35:70	56.5
W	1:2	0.4
WFL	3:5	1.9

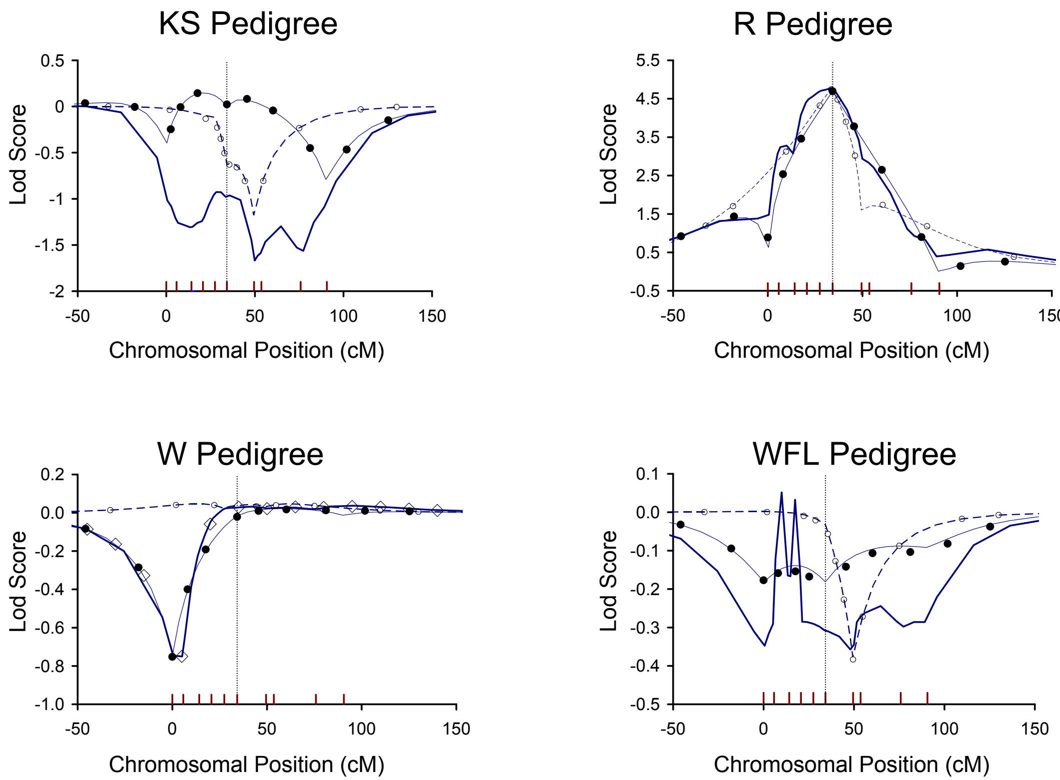
3-marker exact comp. by VITESSE
CPU times in minutes
run-lengths in 1000 MCMC scans;
(preliminary:final)

Does it work 3? – mixing



Plot of realized γ over random block of 5000 scans: 0=unlinked. Actually, this plot is from earlier analyses on same pedigrees.

Do we need 10 markers?



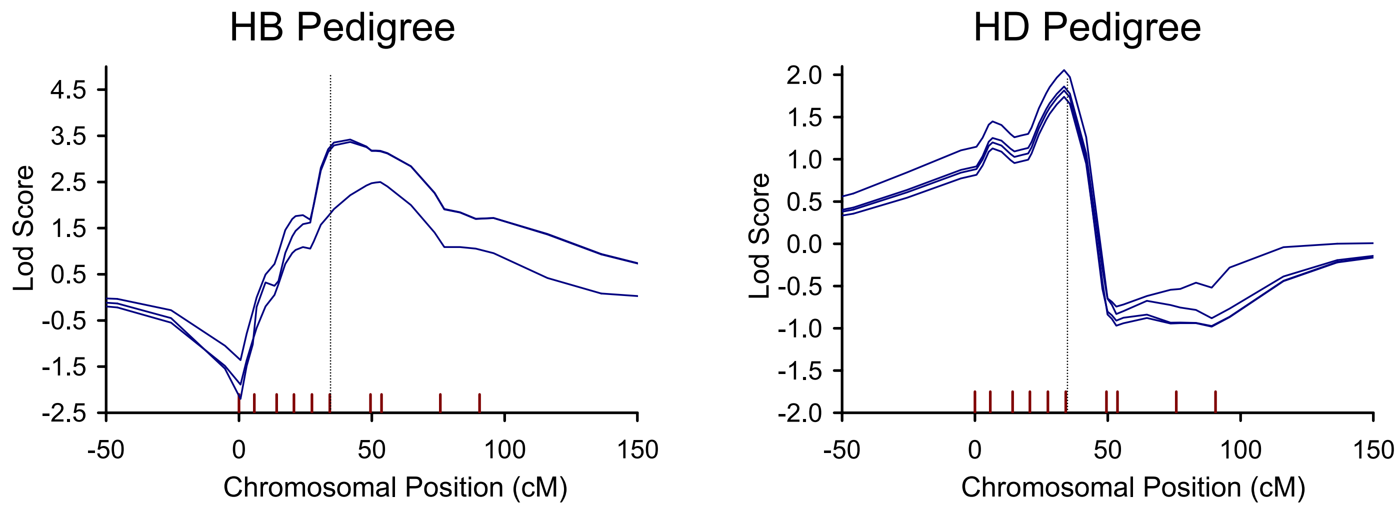
Linked cases: Localization is better. lod scores are higher.
Unlinked case: Rejection of linkage is possible.

Run-time comparisons: complex pedigrees

Marker pair	HB pedigree			HD pedigree		
	Bayes		FSTLNK	Bayes		FSTLNK
	length	time	time	length	time	time
MP-L1	20:40	31.2	257.8	8:18	6.7	201.6
MP-L2	8:16	12.1	174.2	20:40	18.5	75.7
MP-T1	60:180	96.4	362.1	300:600	172.3	158.5
MP-T2	30:90	63.9	859.6	50:100	47.9	122.3

Exact computations: only 2 markers, only by FASTLINK
MCMC estimates: more challenging, but still ok

Complex pedigrees remain a challenge



HD is OK, but for HB which runs are correct?

• If at $\mathbf{S}$ and propose an $\mathbf{S}^\dagger$, Metropolis-Hastings ratio is based on $P(\mathbf{S},\mathbf{Y})/P(\mathbf{S}^\dagger,\mathbf{Y})$.

• This suggests weight to be given to a run restricted to some part of a space of $\mathbf{S}$ should be based on average $P(\mathbf{S},\mathbf{Y})$.

• This is not so easy, but we can easily estimate mean log $P(\mathbf{S},\mathbf{Y})$:
Estimate of ECDLL $=$ exp(log $P(\mathbf{S},\mathbf{Y}) \mid \mathbf{Y})$.

• In example, ECLLD is 2 units higher for HB runs with higher max lod and in the correct position.
That is, the part of the space is 100 times more probable.

CONCLUSION

• Sampling of inheritance patterns given genetic data remains a challenging MCMC problem for multiple markers, missing data, extended pedigrees ...

• Likelihood and lod score estimators can be based on realized inheritance, but need good estimators as well as good samplers

• With both, real-time MCMC estimation of lod scores is both feasible and practical, and even when exact computation is feasible MCMC can be quicker.

• lod scores based on multiple markers provide additional information on gene localization: improved estimation is important for localizing the genes of complex traits.