

An extension to the HapCluster algorithm for mapping interacting genes

Thomas Mailund

March 4, 2007

I describe an extension to the HapCluster algorithm from Waldron et al.² that allows for mapping two unlinked genes for interaction effects in affecting disease risks. This idea is essentially just using an idea I discussed with Ed Waldron about using two clusters to find two mutations, but forcing them to explore two different genes, thus preventing them getting stuck in the same place.

Introduction

The HapCluster algorithm searches for the location of a variant affecting disease status by trying to identify flexible clusters of case haplotypes in the vicinity of a mutant allele. Rather than explicitly modelling the genealogy of the sample, the algorithm uses a locus-dependent distance measure between haplotype-sequences to identify a cluster of sequences close to the “ancestor sequence” where the mutation occurred. The “ancestor sequence” and distance threshold are nuisance parameters integrated out in the MCMC. Sequences within a certain distance of the ancestor sequence are considered to carry the risk affecting allele. Letting A_{ij} denote the number of affected individuals with genotype ij , U_{ij} the number of unaffected individuals with genotype ij and θ_{ij} the disease risk for individuals with genotype ij —where i and j essentially means that the haplotype is within the cluster, a mutant M , or outside the cluster, a wild-type, W , meaning $ij = \{MM, MW, WW\}$ —the likelihood of the model parameters are

$$L(\theta_{MM}, \theta_{MW}, \theta_{WW}, \Theta) = \prod_{ij} \theta_{ij}^{A_{ij}} (1 - \theta_{ij})^{U_{ij}} \quad (1)$$

where Θ denotes remaining parameters such as cluster size, ancestral sequence, etc. The risk parameters, θ_{ij} are assumed independent and integrated out as

$$L(\Theta) = \prod_{ij} \int_0^1 \pi(\theta_{ij}) \theta_{ij}^{A_{ij}} (1 - \theta_{ij})^{U_{ij}} d\theta_{ij} \quad (2)$$

and for the risk prior $\pi(\theta_{ij})$ the uninformative beta density $\pi(\theta_{ij}) = 1$ is used, reducing this to

$$L(\Theta) = \prod_{ij} B(A_{ij} + 1, U_{ij} + 1) \quad (3)$$

where $B(\alpha, \beta)$ denotes the beta function. For further details on the remaining MCMC parameters, moves and proposal densities, see Waldron et al.².

An extension to interacting unlinked genes

Extending this model to two unlinked genes can be done simply by introducing an extra cluster to the MCMC, so there now is a cluster per gene and such that each individual now has two genotypes derived from these clusters as before, one per gene. That is, for the first gene, each individual will have one of the genotypes MM , MW or WW and the same for the second gene. We can then denote by $A_{ij,kl}$ the number of affected individuals with genotype ij in the first gene and genotype kl in the second gene. As for the single gene case, we assign a risk per genotype $\theta_{ij,kl}$, assume independence of these, and consequently get a likelihood

$$L(\Theta) = \prod_{ij,kl} \int_0^1 \pi(\theta_{ij,kl}) \theta_{ij,kl}^{A_{ij,kl}} (1 - \theta_{ij,kl})^{U_{ij,kl}} d\theta_{ij,kl} = \prod_{ij,kl} B(A_{ij,kl} + 1, U_{ij,kl} + 1) \quad (4)$$

The MCMC is set up as for the single gene case, except that there now are two clusters, moving about in parameter space independently.

Test for association

To test for a disease association, either for a single gene or a pair of genes, we can use the Bayes factor:

$$B = \frac{L(M_1)}{L(M_0)} \quad (5)$$

where $L(M_1)$ is the likelihood of the alternative model and $L(M_0)$ is the likelihood of the null model of no association between the gene and the disease status. The null model likelihood is given simply by

$$L(M_0) = \int_0^1 \pi(\theta) \theta^A (1 - \theta)^U = B(A + 1, U + 1) \quad (6)$$

where A is the total number of affected individuals and U the total number of unaffected individuals, while the alternative model is obtained by integrating over all parameters:

$$L(M_1) = \int L(\Theta) P(\Theta) d\Theta \quad (7)$$

We estimate $L(M_1)$ using samples of the likelihood from the posterior distribution, obtained during the MCMC run, since

$$\int \frac{1}{L(\Theta)} P(\Theta | \text{Data}) d\Theta = \int \frac{P(\text{Data}, \Theta) / P(\text{Data})}{P(\text{Data}, \Theta) / P(\Theta)} d\Theta \quad (8)$$

$$= \frac{1}{P(\text{Data})} \int P(\Theta) d\Theta \quad (9)$$

$$= \frac{1}{P(\text{Data})} \quad (10)$$

we have

$$P(\text{Data}) = \frac{1}{\int \frac{1}{L(\Theta)} P(\Theta | \text{Data}) d\Theta} \quad (11)$$

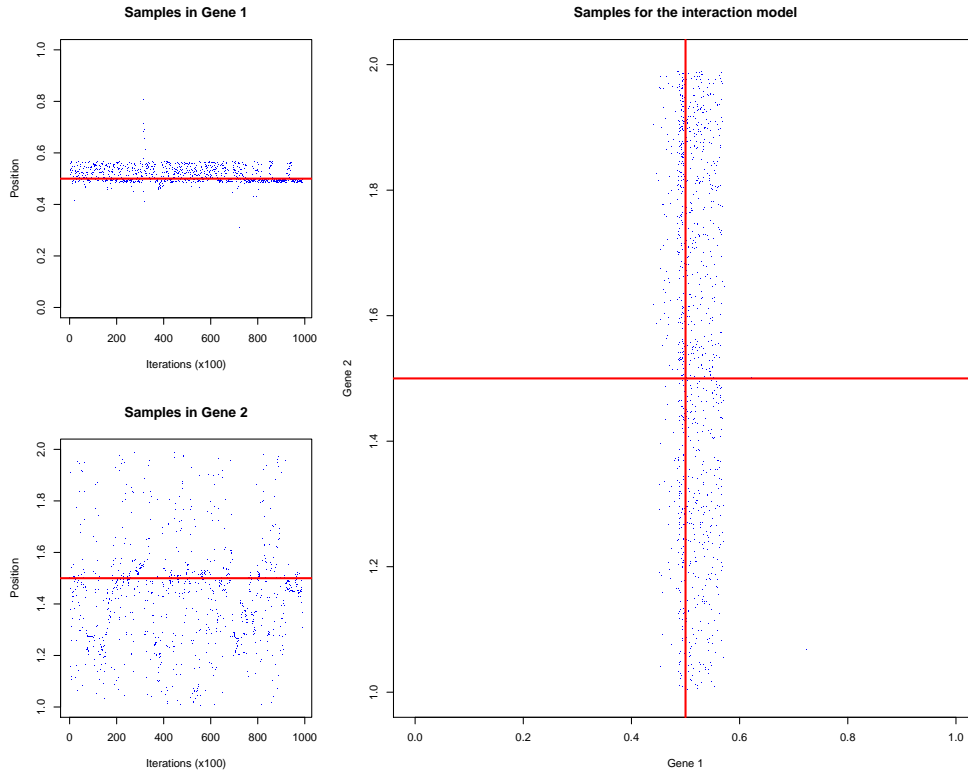


Fig. 1: Example of samples from both marginal and interaction models. The blue dots show samples of the locus (for the marginal model) or pair of loci (for the interaction model). The red lines show the position of the disease markers in the two genes.

where we can estimate the denominator as

$$\int \frac{1}{L(\Theta)} P(\Theta | \text{Data}) d\Theta \approx \frac{1}{N} \sum_{i=1}^N \frac{1}{L(\Theta_i)} \quad (12)$$

where the $\Theta_i \sim P(\Theta | \text{Data})$ are sampled from the posterior during the MCMC run.

Example

Using CoaSim,¹ I simulated two unlinked genes, each with 50 SNP markers and recombination rate $\rho = 200$, and each with a central disease locus. The disease model is dominant model with a risk of 20% for individuals with a mutant at both disease loci and a risk of 5% for the remaining individuals. From this is sampled 250 cases and 250 controls.

Figure 1 shows an example run of such a dataset. On the left is shown samples from a single-gene run for both the two genes, and on the right the sampled pairs of positions from the interaction model. As evident from the graph, there is very little signal in gene 2 in either the marginal model or the interaction

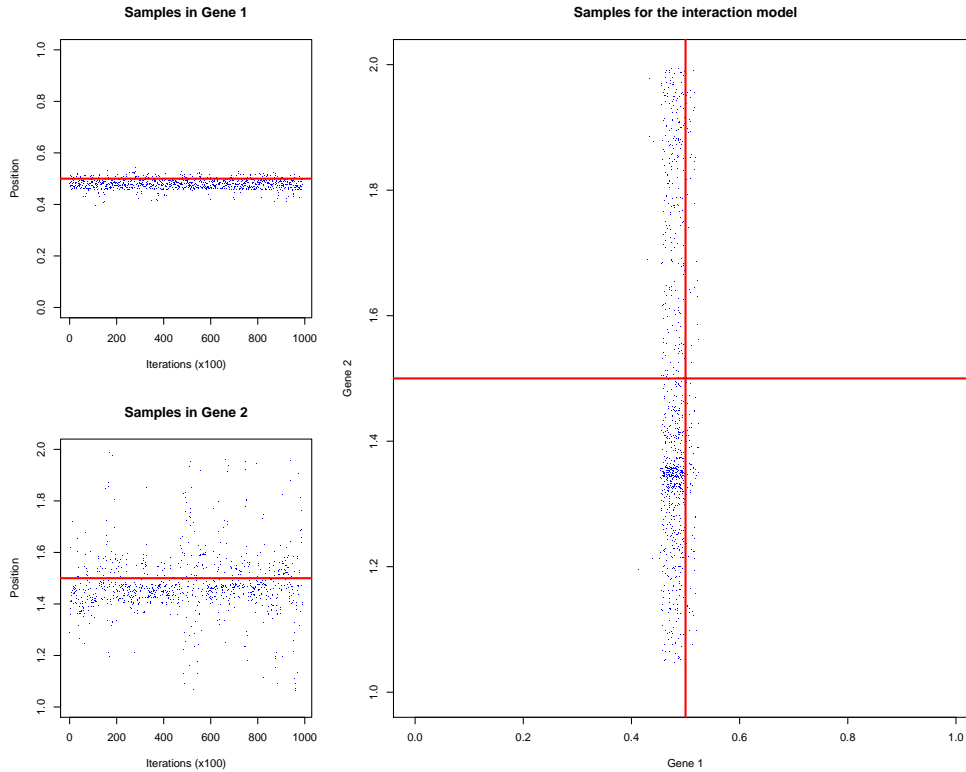


Fig. 2: Another example of samples from both marginal and interaction models.

model. The Bayes factor for association in gene 1 is 1526 (highly significant) but for gene 2 is only 1.22 which is not significant. The Bayes factor for the interaction model over the null model is 100216 (highly significant) and for the interaction model over gene 1 the Bayes factor is 65.7, which is highly significant indicating that the effect seen in the interaction model is not only through gene 1.

Figure 2 shows another example, simulated under the same model. Here both marginal models pick up an interaction (with Bayes factor $3.77 \cdot 10^{19}$ for the first gene and 4.40 for the second (both significant)). The interaction model also detects an interaction, with $1.62 \cdot 10^{20}$ over the null model, and 4.28 and $3.67 \cdot 10^{19}$ over the marginal models for gene 1 and gene 2, respectively.

Table 1 summarises results for five datasets. Of these, the disease association is significant in the interaction model for all but dataset number 4, and in all but number 4 it is also significantly better than either of the marginal models.

Next I simulated datasets where only the marker in gene 1 was associated with the disease (with a dominant model with risk 20% for mutants and 5% for wild-types). The summaries of these experiments are shown in Table 2.

In the next setup, I simulated datasets where disease status was assigned arbitrarily and independent of the genotype in either genes, and thus form the null model of association. Summaries are shown in Table 3. Although the

Data set	B_{10}	B_{20}	B_{I0}	B_{I1}	B_{I2}
1	1 526	1.22	100 216	65.7	82 278
2	$3.77 \cdot 10^{19}$	4.40	$1.62 \cdot 10^{20}$	4.28	$3.67 \cdot 10^{19}$
3	$2.92 \cdot 10^{07}$	0.528	$1.62 \cdot 10^{08}$	5.55	$3.06 \cdot 10^{08}$
4	1.98	0.709	1.14	0.576	1.61
5	23.3	0.323	88.4	3.79	273

Table 1: Bayes factors for models with interacting genes. The columns give the Bayes factor for the marginal model for each gene, the interaction model against the null model, and the interaction model against the two marginal models, respectively.

Data set	B_{10}	B_{20}	B_{I0}	B_{I1}	B_{I2}
1	$2.96 \cdot 10^{14}$	0.518	$2.69 \cdot 10^{14}$	0.91	$5.2 \cdot 10^{14}$
2	$6.60 \cdot 10^{15}$	0.435	$6.93 \cdot 10^{15}$	1.05	$1.59 \cdot 10^{16}$
3	$6.45 \cdot 10^{16}$	0.305	$1.36 \cdot 10^{17}$	2.11	$4.46 \cdot 10^{17}$
4	$3.32 \cdot 10^{11}$	0.389	$5.45 \cdot 10^{10}$	0.164	$1.40 \cdot 10^{11}$
5	$1.18 \cdot 10^{10}$	0.313	$1.26 \cdot 10^{10}$	1.07	$4.02 \cdot 10^{10}$

Table 2: Bayes factors for models with a disease marker in gene 1 only. The columns give the Bayes factor for the marginal model for each gene, the interaction model against the null model, and the interaction model against the two marginal models, respectively.

Data set	B_{10}	B_{20}	B_{I0}	B_{I1}	B_{I2}
1	0.282	0.432	0.136	0.481	0.314
2	0.339	0.445	0.586	1.73	1.32
3	0.452	5.15	5.26	11.6	1.02
4	0.279	0.370	0.121	0.435	0.328
5	0.489	0.301	0.239	0.490	0.796

Table 3: Bayes factors for models where the disease status is assigned independent of genotype. The columns give the Bayes factor for the marginal model for each gene, the interaction model against the null model, and the interaction model against the two marginal models, respectively.

Data set	B_{10}	B_{20}	B_{I0}	B_{I1}	B_{I2}
1	62.5	$9.66 \cdot 10^{13}$	$1.36 \cdot 10^{12}$	$2.17 \cdot 10^{10}$	0.0140
2	47.9	$2.17 \cdot 10^{08}$	$7.07 \cdot 10^{15}$	$1.47 \cdot 10^{14}$	$3.25 \cdot 10^{07}$
3	10 437	1 125	$2.00 \cdot 10^{09}$	191 501	$1.77 \cdot 10^{06}$
4	0.621	$3.21 \cdot 10^{09}$	$1.21 \cdot 10^{09}$	$1.95 \cdot 10^{09}$	37.7
5	7.12	16 382	$1.65 \cdot 10^{06}$	229 999	100

Table 4: Bayes factors for models where the disease status is assigned based on mutations in either gene, but with no interaction between the two genes.

interaction model and gene 2 both have a significant Bayes factor in dataset 3, the remaining, encouragingly, are not significant.

Finally, I simulated datasets where disease status was assigned to individuals if either gene was mutant, but with no interaction between the alleles in the two genes. Results are shown in Table 4. Aside from the first dataset—where the interaction model is not preferred over the second gene—the interaction model is significantly better than either of the genes, indicating that we cannot directly use the three models to distinguish between interaction or an independent effect in both genes.

References

1. T. Mailund, M. Schierup, C. Pedersen, P. Mechlenborg, J. Madsen, and L. Schausser. CoaSim: a flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics*, 6:252, 2005.
2. E. R. B. Waldron, J. C. Whittaker, and D. J. Balding. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30(2):170–179, 2006.