

# Mutation- and Null-clusters in GeneRecon<sup>1</sup>

---

Thomas Mailund

April 25, 2005

We report on initial experiments with using a mutation- and a null-cluster for affected individuals in the GeneRecon LD association mapping tool. The genealogy of affected individuals in the mutation-cluster is explicitly modelled, while affected individuals in the null-cluster is considered to be unrelated and are thus modelled as control cases; an MCMC is used to select which individuals should be included in the mutation-cluster and which should be included in the null-cluster. By only explicitly modelling the genealogy of a subset of the individuals we significantly speed up the running time of GeneRecon, but with little loss of accuracy.

## INTRODUCTION

---

GeneRecon is a tool for analysis of population genetic data from case/control studies. The analysis works on genetic data which is collected from patients with a specific disease and a control group without this disease, and tries to locate genes that are increasing the risk of disease. The analysis, which is based on a Markov-chain Monte-Carlo (MCMC) method, is extremely CPU-demanding, since it requires searching through a very large parameter-space. This often requires several runs of the method for a given set of data, with each run taking days or even weeks.

The algorithm implemented in GeneRecon is based on Morris *et al.*'s shattered coalescent method [3], which explicitly models the genealogy of the set of affected individuals in a candidate locus. In many cases, a large fraction of the affected individuals are not affected due to a mutation in the region being analysed—either they are diseased by other, e.g. environmental, effects, or as a consequence of a mutation in another gene. The shattered coalescent method handles this by considering some parent-child edges in the coalescent tree as if they were removed and the child taken from some background distribution of genotypes. The coalescent tree, however, is still modelled over all affected individuals, resulting in an enormous search space for the algorithm to explore.

The presented work evaluates a method that significantly reduces the running time of the shattered coalescent method by incorporating ideas from Liu *et al.* [1] and Molitor *et al.* [2] by separating the affected individuals into a *mutation cluster*—where affected individuals are assumed to be descendants of

---

<sup>1</sup>The mutation-/null-cluster dichotomy of affected individuals was initially suggested and implemented by Jesper Nymann Madsen in an earlier version of GeneRecon, but were re-implemented for the version used for these experiments. The experiments were conducted on Minimal intrusion Grid (MiG), at the University of Southern Denmark, in cooperation with Jonas Bardino, Henrik H. Karlsen and Brian Vinter.

a common founder—and a *null cluster*—for individuals affected due to other factors. By only explicitly modelling the genealogy of a subset of affected individuals, we reduce the tree space that must be explored by the method and thus potentially greatly reduce the number of iterations needed for the exploration. Also, by sampling, during the run of the MCMC, the distribution of affected individuals among the mutation and null cluster, we hope to be able to infer which individuals carry an increased risk mutation and which are affected solely due to other factors.

## METHODS

---

*The clustering method.* The clustering method extends the shattered coalescent method by splitting the set of affected individuals into a *mutation-cluster*—where the genealogy is explicitly modelled at a locus using a shattered coalescent tree as in Morris *et al.* [3]—and a *null-cluster*—where the affected individuals are treated just as the unaffected individuals in the shattered coalescent method. The MCMC can suggested replacing one individual from the mutation-cluster with an individual from the null-cluster, but moving the null-cluster individual into the coalescent tree at the position of the mutation-cluster individual, and the mutation-cluster individual to the null-cluster.

*The experimental setup.* We have simulated haplotype data sets using the CoaSim simulator <http://www.birc.dk/Software/CoaSim/> under varying recombination rate ( $\rho = 40$  and  $\rho = 400$ , roughly corresponding to 0.1cM and 1cM), with varying marker densities (20 markers on the region, 40 markers on the region, and a twice as wide region with 40 markers, 20 of which are in the middle region—containing the disease marker—and 10 on each side), and varying disease models ( $m/w$  where  $m$  is the fraction of affected individuals among the mutants and  $w$  is the fraction of affected individuals among the wild-types).

For each dataset, four chains of GeneRecon was run for cluster sizes 100 (all affected individuals), 75, and 50. The chains were run on the *Minimal intrusion Grid* (MiG) with the individual jobs managed by a set of custom-built Python and Bash scripts. The scripts were responsible for setting up the runs for the correct data sets, submitting and monitoring the queued and running jobs, and for downloading and collecting the results. Depending on the CPU the chains were run on and the cluster size for the run, each job took from 6 hours to three days to execute. On average, 40–50 CPUs were in use, and the experiments, at the time of writing, have been running for about 4 months.

The final analysis of the results was conducted using a set of R scripts. In calculating the error of disease locus inference, we have used a simple measure of the distance from the point with maximum posterior value to the true disease locus. This does not catch the confidence of the inference, but gives a simple measure of error that lets us compare the accuracy for various cluster sizes.

## RESULTS

For the first part of the experiment we have validated that introducing the mutation- and null-cluster gives a runtime speedup, but does not reduce the methods ability to locate the disease locus significantly. For this study we have, for datasets simulated under various parameters, run several chains of the algorithm both with and without clustering, and then calculated the distance from the true location to the inferred location.

Figure 1 shows the results for  $\rho = 400$ , or  $1\text{cM}$ , for the disease model  $1/0$ —all mutants being affected and no wild-types being affected—for cluster sizes 100, 75, and 50. Little accuracy is lost when going from a mutation-cluster size of 100—corresponding to the original shattered coalescent model—to a cluster size of 75, whereas somewhat more is lost when going to a cluster size of 50. That some accuracy is lost is, perhaps, not surprising considering that with this disease model all affected are mutants and none belong in the null-cluster.

For  $\rho = 40$ , or  $0.1\text{cM}$ , the reduction in accuracy when moving from a mutation-cluster size of 100 to 50, still for the disease model  $1/0$ , is much less, as shown in Figure 2; except for the leftmost column in the figure, there is very little loss in accuracy.

Varying the disease model to  $0.8/0.1$ —that is, mutants are affected with probability 0.8 and wild-types with probability 0.1—we see a small loss in accuracy for some of the settings for  $\rho = 400$ , see Figure 3, but interestingly the loss is

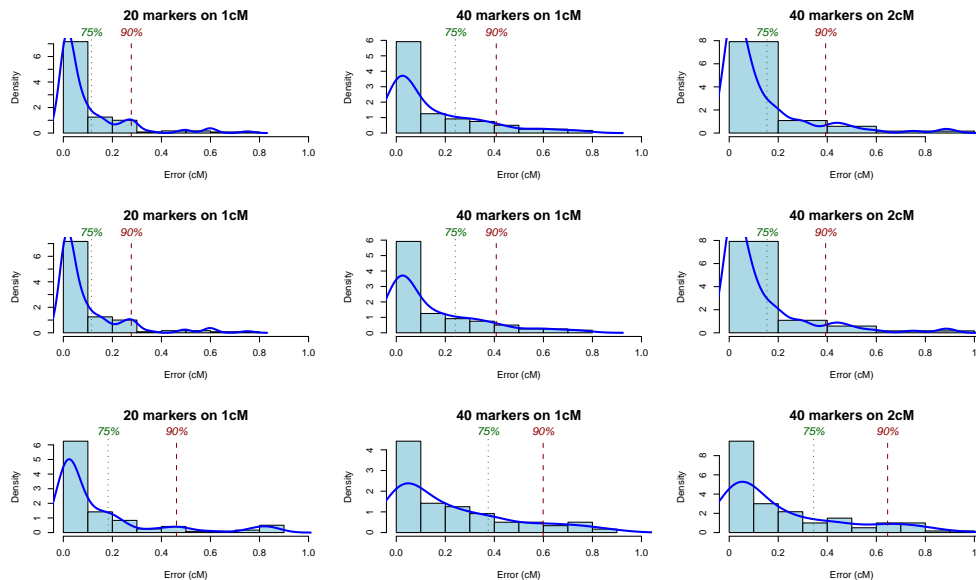


Figure 1: Varying cluster sizes over a  $1\text{cM}$  region. Disease model is  $1/0$ . The plots show the errors (distance from true disease locus to inferred disease locus) for varying mutation cluster sizes (100, 75, 50, from top to bottom, where 100 is the total number of affected) and varying marker densities: 20 markers on  $1\text{cM}$  in the leftmost column, 40 markers in the middle column, and 40 markers spread over  $2\text{cM}$  where the disease locus is in the middle  $1\text{cM}$ .

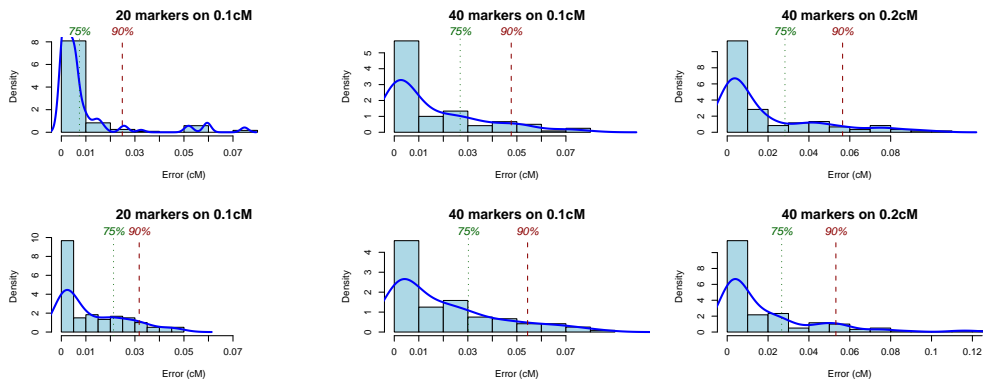


Figure 2: Varying cluster sizes over a  $0.1cM$  region. Disease model is  $1/0$ . The plots show the errors (distance from true disease locus to inferred disease locus) for varying mutation cluster sizes (100, 50, from top to bottom, where 100 is the total number of affected) and varying marker densities: 20 markers on  $0.1cM$  in the leftmost column, 40 markers in the middle column, and 40 markers spread over  $0.2cM$  where the disease locus is in the middle  $0.1cM$ .

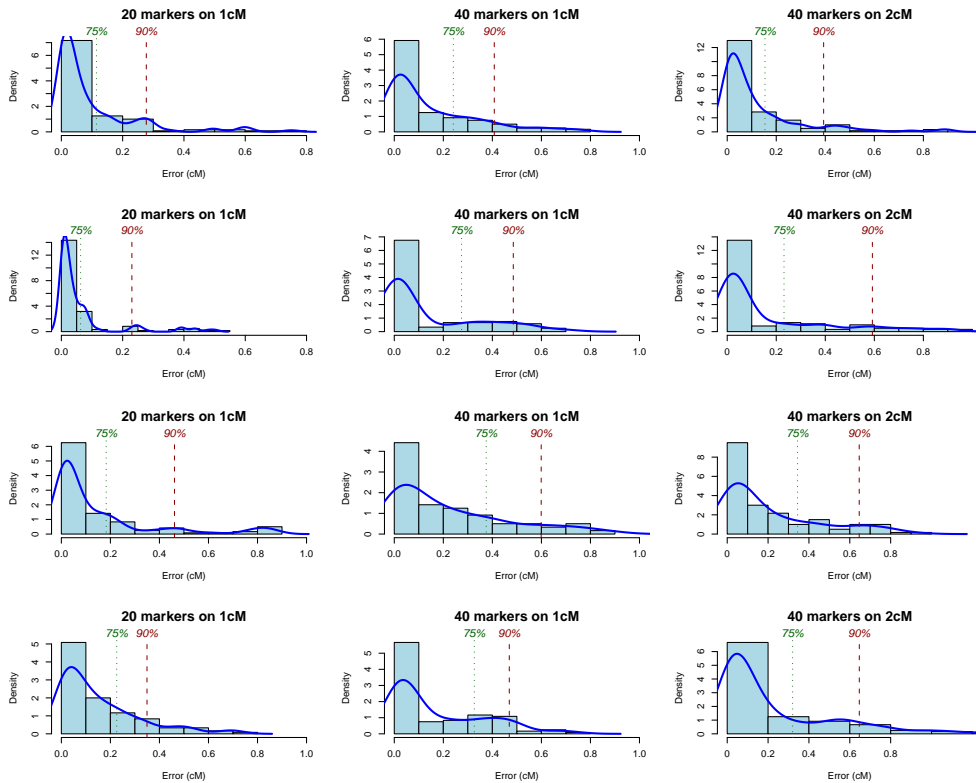


Figure 3: Disease models  $1/0$  and  $0.8/0.2$  over  $1cM$ . The two topmost rows show the errors for a mutation-cluster size of 100, the topmost row for disease model  $1/0$  the second for disease model  $0.8/0.2$ ; the next bottommost rows show the same setting but for a mutation-cluster size of 50.

smaller for the mutation-cluster size 50 than mutation-cluster size 100. Still, cluster size 50 is less accurate than cluster size 100 for both disease models.

---

## CONCLUSIONS

---

Preliminary results show that there is some loss in accuracy when using a smaller mutation-cluster size than the full set of affected individual, but that this loss is not significant for a cluster size of 75 out of 100 for  $\rho = 400$  or 50 out of 100 for  $\rho = 40$ . This despite that the running time is reduced from several days to one day or a quarter of a day—the exact measurement of the running time has not been done yet since the heterogeneous nature of MiG complicates this.

The disease model 0.8/0.2 is currently being run for  $\rho = 40$ , and following this we will run with a disease model that makes half the affected individuals mutations,  $P(\text{mutant} | \text{affected}) = 0.5$ , which is achieved for any  $P(\text{affected} | \text{mutant}) = 4 \cdot P(\text{affected} | \text{wild-type})$  for the simulations where  $P(\text{mutant}) = 0.2$  and  $P(\text{wildtype}) = 0.8$ . In this setting we will also sample the mutation-cluster and evaluate if the mutants are over-represented in the cluster.

---

## REFERENCES

---

1. J.S. Liu, C. Sabatti, J. Teng, B.J.B. Keats, and N. Risch, *Bayesian analysis of haplotypes for linkage disequilibrium mapping*, *Genome Research* **11** (2001), 1716–1724.
2. J. Molitor, P. Marjoram, and D. Thomas, *Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques*, *Am. J. Hum. Genet.* **73** (2003), 1368–1384.
3. A.P. Morris, J.C. Whittaker, and D.J. Balding, *Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies*, *Am. J. Hum. Genet.* **70** (2002), 686–707.