# A COMPARISON OF EFFICIENT GENOTYPE SAMPLERS FOR COMPLEX PEDIGREES AND MULTIPLE LINKED LOCI

**C. Stricker[1,2], M. Schelling[2], F. Du [3] I. Hoeschele[3], S. A. Fernández[4] and R. L. Fernando[4]**

[1] Applied Genetics Network, 8852 Altendorf, Switzerland
[2] Institute of Animal Sciences, ETH Zurich, 8092 Zurich, Switzerland
[3] Departments of Dairy Science and Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[4] Department of Animal Science, Iowa State University, Ames, IA, USA

## INTRODUCTION

QTL mapping usually involves estimating the location of as well as the effect and frequency of alleles, or the variance contribution, for the QTL. Specific experimental designs are established to increase the statistical power, to simplify the analysis and to control the associated costs. However, such designs are not always available in livestock and humans. Furthermore, data on many economically important traits are routinely available in livestock. Therefore, appropriate statistical methods to map QTL under various genetic models in general pedigrees are needed.

The major issues in QTL mapping are (i) the complexity of the pedigree structure, (ii) the number of loci that are included in the model, and (iii) the amount of missing marker data. There are two major approaches in model-based QTL mapping: Deterministic maximum likelihood or Markov Chain Monte Carlo (MCMC) simulation used for maximum likelihood (MCEM, see e.g. Guo and Thompson, 1994) or Bayesian inference (e.g. Thomas and Cortessis, 1992, Heath, 1997). These two approaches will be discussed below in the context of issues (i) to (iii).

With respect to issue (i), Elston and Stewart (1971) proposed an efficient algorithm to calculate deterministically the likelihood for large pedigrees without loops and only a few segregating loci. Their algorithm is referred to as 'peeling'. Although for small pedigrees, extensions to account for loops have been proposed (Cannings et al., 1978; Lange and Elston, 1975; Lange and Boehnke, 1983), these approaches are not suitable for the size and complexity of general livestock pedigrees. Thus, Wang et al. (1996) proposed an approximation to the likelihood for oligogenic models in large and complex pedigrees with loops based on the original work of Elston and Stewart (1971) and iterative peeling (Janss et al., 1992). As an alternative to these deterministic approaches, Markov Chain Monte Carlo (MCMC) methods have been proposed to accommodate complex pedigree structures. An early example of the MCMC approach is Thompson (1991).

Issue (ii), i.e. the number of genetic loci in a model, has two aspects: the number of marker loci and the number of QTL. Since in deterministic maximum likelihood, summation is over unobserved phase known genotypes, models accounting for more than a few QTL soon become computationally unfeasible. Alternatively, when the residual QTL are modeled as a normally distributed polygenic component, the likelihood could be computed only for small pedigrees

(Fernando *et al*., 1994). An approximation for such mixed inheritance has been proposed by Hasstedt (1982). For a model with a few major QTL and several residual QTL segregating, Fernando et al. (1994) proposed an approximation to calculate the likelihood. MCMC on the other hand is completely general with respect to the genetic model assumed. Realizations based on any number of QTL and any distribution assumed can be simulated by an appropriate sampler.

Order refers to the maternal or paternal origin of an allele, whereas phase refers to the placement of alleles across several loci onto haplotypes without specifying the maternal or paternal origin of the haplotype. Thus for multiple loci, order includes the information about phase. The state space for all possible ordered or phase known genotypes to sum over in deterministic likelihood calculation or to sample from by MCMC methods becomes large even when only a few incompletely observed multiallelic marker loci are considered. Thompson (1994) and Sobel and Lange (1996) argue that linkage information is solely in the transmission pattern of alleles from parents to offspring and not in the specific alleles that were transmitted. In some instances, this may lead to a substantial reduction in the size of the state space. Their idea will be further discussed in the theory section in this paper.

Issue (iii) is connected to issue (ii) in the sense that when only a small portion of genotypes at the marker loci are observed, the state space of all possible multilocus transmission patterns or genotypic configurations increases. Consequently, unobserved genotypes at several marker loci make deterministic likelihood calculation and MCMC simulations increasingly inefficient. However, computational efficiency differs considerably between the various MCMC approaches, partly depending on the unobservable genetic variable that is sampled. Therefore, in addition to the approximation necessary to account for complex pedigrees and mixed inheritance, the number of incompletely observed marker loci is a major problem for deterministic maximum likelihood in multipoint linkage studies. MCMC approaches bear the potential to become the state of the art for such multipoint analyses. Below, we will discuss a few properties of MCMC approaches to sample genetic information from different state spaces and present different MCMC approaches for inference about genotypes in pedigrees.

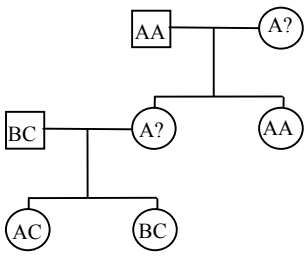


**Figure 1. Example for a reducible Markov chain generated by sampling genotypes by a Gibbs sampler when only a single parent is missing. Alleles A, B and C segregating. Two non-communicating states B or C for allele ? possible (Sheehan & Cannings, 2002)**

## SAMPLING GENOTYPES

Initially, sampling genotypes by the Gibbs Sampler was thought to be a major breakthrough in QTL mapping. However, it was soon realized that for loci with more than two alleles, sampling single genotypes by a Gibbs sampler does not guarantee an irreducible Markov chain. The example of a nuclear family with unknown parents and two offspring with genotype AB and CC, respectively has been frequently used to illustrate this. Thus, several authors claim that the Markov chain is irreducible, if at least one parent has observed marker genotypes, a situation frequently encountered in livestock pedigrees where DNA on male parents is usually available. The counter example in Figure 1

taken from Sheehan and Cannings (2002) demonstrates that this statement is not true in general. Updating blocks of variables (genotypes) jointly may provide an irreducible Markov chain. However, there is no general algorithm available to construct blocks of genotypes such that irreducibility is guaranteed in general pedigrees, except when the block consists of the entire pedigree. The latter is the basis of the approach taken by Fernández et al. (2002a, 2002b). For simple pedigrees, the entire pedigree is 'peeled' exactly using the Elston-Stewart algorithm, then ordered genotypes are sampled by reverse peeling (Heath, 1998). When a pedigree is complex and cannot be 'peeled' exactly, they use a modified pedigree to draw joint genotypic realizations from. Genotypes are peeled exactly until loops make peeling too inefficient. Then, the remaining loops are cut and the pedigree is extended at the cuts. Finally, samples are generated from the modified pedigree, and these are accepted with the probability given by the Metropolis-Hastings ratio. Since the proposal distribution (i.e. the modified pedigree) is very close to the true pedigree, an independence sampler is used. Fernández et al. (2002a, 2002b) call this the ESIP-sampler, combining the **E**lston-**S**tewart algorithm and **i**terative **p**eeling. In Fernández et al. (2002b) they show that the Markov chain generated by the ESIP-sampler is irreducible and aperiodic. Extension of the ESIP-sampler to multiple loci is straightforward, implying that peeling needs to be performed over several loci jointly. However, as the sampler relies on peeling, efficiency decreases exponentially with increasing number of loci considered in the model. For multipoint linkage studies using flanking markers, the ESIP-sampler does provide a fast mixing, efficient sampler as illustrated below.

## SAMPLING ALLELIC ORIGIN

Thompson (1994) and Sobel and Lange (1996) distinguish two types of genetic variables in a pedigree: 1) the grand maternal or grand paternal origin of each non-founder allele, which have been called segregation indicators, and 2) the ordered genotype of each individual. Individuals with observed genotypes put restrictions on the size of the state spaces for these variables. The state space for the latter, which is the set of all ordered genotypes compatible with the observed data, is usually substantially larger. The information that is useful for making inferences on the linkage between loci is contained entirely in the transmission indicators.

In the algorithms proposed by Thompson (1994) and by Sobel and Lange (1996), segregation indicators for non-founders were sampled without sampling the genotypes of the founders. To do so, they first computed the joint probability of the founder genotypes and the non-founder segregation indicators; then, the marginal probability for the segregation indicators was obtained by deterministically taking the sum of these probabilities over all the possible founder genotypes that are compatible with the observed data. Segregation indicators were sampled from this marginal distribution.

In recent applications to livestock and plant pedigrees, in addition to segregation indicators, the genotypes of founders were also sampled (Jansen et al., 1998; Bink and Van Arendonk, 1999, Yi and Xu, 2001). The size of the state space for segregation indicators and for ordered genotypes depend on the amount of unobserved marker genotypes and the size of the pedigree. There is no simple rule to identify which state space is smaller. In the recent applications listed above performing simulations on the state space of segregation indicators, the Gibbs sampler

was used. As illustrated in Figure 2, even when all genotypes are observed, the Gibbs sampler does not guarantee an irreducible Markov chain.

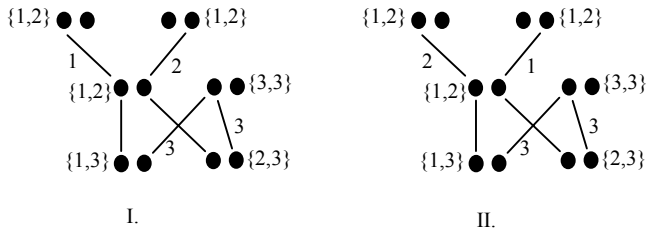


**Figure 2. Simple pedigree demonstrating vertical dependence. Each individual's single locus is depicted with 2 alleles as black dots. Unordered observed genotypes are listed in curly brackets. Assuming the left allele at each locus to be of maternal origin, the lines indicate segregation indicators. The numbers at the lines denote the alleles travelling on the paths. Applying a Gibbs sampler results in a reducible Markov chain with non-communicating sets I. and II. I. corresponds to the ordered genotype of [1,2] for the middle offspring, II. to [2,1]. Both are equally likely given the data**

Note that for the pedigree in Figure 2 the genotypes of all individuals are known. The Gibbs sampler fails even for sampling the order of known genotypes, when segregation indicators are sampled. It is due to strong vertical dependence, i.e. dependence between individuals' loci. Therefore, the state space of segregation indicators does not ensure irreducibility in general when the Gibbs sampler is used. Sobel and Lange (1996) do not simulate from the full conditional distribution of a segregation indicator given all remaining segregation indicators and the data like in the Gibbs sampler. They use a Metropolis sampler to simulate from the conditional distribution of segregation indicators only. Their sampler will mix poorly due to only small changes being compatible with the data. To improve mixing, they update segregation indicators for certain blocks of segregation indicators jointly with the blocks chosen based on genetic relationship between individuals (full- or halfsibs). Although this resolves the problem inherent in Figure 2, it will not in general guarantee an irreducible Markov chain (c.f. Figure 7 in Sobel and Lange, 1996). To achieve this, they use a random number of updating steps to propose a new realization of the Markov chain. The random number is taken from a geometric distribution with mean 2. Such multiple updating steps allow to step through illegal transmission patterns, revealing a theoretically irreducible Markov chain. Du and Hoeschele (2002) have modified the sampler of Sobel and Lange to improve its mixing properties further. They introduced the concept of pedigree splitting and grouping. Some segregation indicators are independent of others and can thus be grouped and updated independently. They extended the idea of grouping meioses; segregation indicators that can only take on a single value compatible with the data are grouped. For all remaining segregation indicators, they implemented 4 different grouping strategies. The first two strategies group according to genetic relationship the segregation indicators of full-sibs or offspring from the same parent as in Sobel and Lange. Grouping methods 3 and 4 are based on observed marker data. Some segregation indicators restrict the grandparental origin of other segregation indicators, while for some other segregation indicators, changing their state forces a change in the number of

permissible states for other segregation indicators. These segregation indicators are grouped (grouping strategy 3 and 4, respectively) and updated jointly. The main purpose of meiosis grouping is to obtain groups with a small number of transmission patterns compatible with the data. Jointly updating these grouped segregation indicators is performed by sampling from all legal transmission patterns in a group of segregation indicators, conditional on the observed data for that group. Compared to the original approach of Sobel and Lange where only small changes in the transmission pattern will be accepted with high probability, i.e. mixing is poor, updating a random number k of groups of segregation indicators jointly conditional on data is expected to improve mixing. Acceptance rate for the sampler is further improved, when updating all groups is done sequentially and after each group the resulting incomplete transmission pattern is evaluated conditional on all data. This strategy will generate independent proposals. To generate proposals conditional on the current transmission pattern, Du and Hoeschele consider a third strategy, which represents a combination of the previous two samplers. All three sampling strategies of Du and Hoeschele reveal an irreducible Markov chain, since in two of their strategies, the number k of groups to update may be equal to the total number of groups in a pedigree in some cycles. Their strategy to update all groups of segregation indicators sequentially generates an independent legal proposal and thus an irreducible Markov chain. Until here, we only considered a single locus. Considering multiple loci independently will generate a poorly mixing sampler for tightly linked loci. The grouping strategies of Du and Hoeschele must thus be extended to include multiple loci. Research on this strategy is currently in progress and results will be presented at the WCGALP.

Schelling et al. (2000) and Schelling (2002) implemented the original sampler of Sobel and Lange into the software package MATVEC (Wang et al., 2002) treating different loci independently in multipoint analysis. As expected, the sampler revealed poor mixing, i.e. practical reducibility in some situations for tightly linked loci. Examples are presented below. Based on the state space of segregation indicators, Schelling (2002) developed a haplotype sampler considering the segregation indicators at linked loci. Assume a single segregation indicator is chosen and updated according to Sobel and Lange (1996), this segregation indicator is called the SL-indicator. Then all segregation indicators adjacent to the SL-indicator are generated based on an appropriate mapping function conditional on the SL-indicator. The SL-indicator is chosen at random and updated according to Sobel and Lange, thus retaining the property of theoretical irreducibility. Results from the haplotype sampler are presented below. Thompson and Heath (1997) introduced a haplotype updating scheme based on the Gibbs sampler. As was demonstrated in Figure 2 even for a single locus with complete genotype information, the Gibbs sampler does not guarantee an irreducible Markov chain. Schelling (2002) realized that a haplotype sampler may show poor mixing for tightly linked loci when recombination between such loci are present in the data. The reason is that generating segregation indicators based on a mapping function for tightly linked loci will generate non-recombinant haplotypes with high probability. Such haplotypes are not compatible with the data, if recombinations are present in the data. Thus, these realizations get rejected by the Metropolis sampler. To account for this problem, Schelling (2002) developed an alternative approach to sample segregation indicators of haplotypes jointly. As in his haplotype sampler, an SL-indicator is updated according to the principles described in Sobel and Lange (1996). According to a mapping function, recombination probability $\theta$ is calculated between the updated locus i-1 and the

adjacent locus i. With probability 1-θ the same "update" as on segregation indicator i-1 is performed on segregation indicator i. With probability θ the alternative action as on segregation indicator i-1 is performed on segregation indicator i. This process is used also used to sample segregation indicators at locus i-2, and the process is continued to the left and right until all loci are sampled. Unlike the haplotype sampler of Schelling (2002), this approach will not tend to loose recombination events between tightly linked loci in its proposals. We will refer to this sampler as **c**ascading **o**rigin sampler (**CO**-sampler) in the results presented below.

## DEMONSTRATING THE SAMPLERS

The two examples presented will represent simple pedigrees with all individuals genotyped at two very tightly linked loci 0.01 cM apart. For further simplification, only the maternal meioses of the terminal offspring are informative. In all examples, there are 4 ordered possible genotypes for mother 10, but only the orders 1-1/2-2 and 2-2/1-1 (1-1 indicates a haplotype) are likely to be sampled, due to the loci being tightly linked and 10 terminal offspring receiving a haplotype 1-1 from mother 10. We expect these two ordered genotypes to be sampled with equal probability of virtually 0.5. The terminal offspring 100-110 are expected to get their 1-1 haplotype from the maternal grandmother or maternal grandfather with probability 0.5. The data for the first example is in Table 1. Note that the structure of the example in Figure 2 is maintained but extended by one tightly linked locus and 9 additional terminal offspring. The results in Table 2 show that the original Sobel and Lange (SL) sampler is locked in the starting configuration. This is due to strong dependence between loci that are tightly linked (horizontal dependence): for the sampler to move to the other phase for the genotype of mother 10, it needs to move through a state with 11 recombinant offspring 100-111 (reasoning not shown here, see T1-rule in Sobel and Lange). Such a proposal has a probability of being accepted which is virtually zero. All other samplers perform equally well in this example of strong horizontal dependence, although with quite different acceptance rates. Note that the acceptance rate for the ESIP sampler is 1.0, since the pedigree is simple and can be peeled exactly. Now let us consider the second example: Assume the data in Table 1 except that terminal offspring 111 has now the unordered genotypes {1,3} at the first and {2,3} at the second locus. Following the same reasoning as above, offspring 111 is now likely a recombinant. The results in Table 3 indicate that for the

**Table 1. Pedigree & genotypes of example 1**

| Ind. | Mother | Father | Genotype at ... | |
|---|---|---|---|---|
| | | | Marker 1 | Marker 2 |
| 1 | 0 | 0 | {1,2} | {1,2} |
| 2 | 0 | 0 | {1,2} | {1,2} |
| 10 | 1 | 2 | {1,2} | {1,2} |
| 20 | 0 | 0 | {3,3} | {3,3} |
| 100 | 10 | 20 | {1,3} | {1,3} |
| 101 | 10 | 20 | {1,3} | {1,3} |
| 102 | 10 | 20 | {1,3} | {1,3} |
| 103 | 10 | 20 | {1,3} | {1,3} |
| 104 | 10 | 20 | {1,3} | {1,3} |
| 105 | 10 | 20 | {1,3} | {1,3} |
| 106 | 10 | 20 | {1,3} | {1,3} |
| 107 | 10 | 20 | {1,3} | {1,3} |
| 108 | 10 | 20 | {1,3} | {1,3} |
| 109 | 10 | 20 | {1,3} | {1,3} |
| 110 | 10 | 20 | {1,3} | {1,3} |
| 111 | 10 | 20 | {2,3} | {2,3} |

**Table 2. Proportion of ordered genotypes for mother 10 in example of Table 1. 100'000 rounds for each sampler**

| | Ordered Genotypes | | Acceptance |
|---|---|---|---|
| | 1-1/2-2 | 2-2/1-1 | rate |
| ESIP | 0.4977 | 0.5023 | 1.0 |
| Original Sobel & Lange | 1 | 0 | 0.078 |
| Haplotype sampler (HT) | 0.5005 | 0.4995 | 0.438 |
| Cascading Origin (CO) | 0.4973 | 0.5027 | 0.274 |

**Table 3. Proportion of ordered genotypes for mother 10 in example of Table 1, but offspring 111 having genotype {1,3} and {2,3}. 100 000 rounds for each sampler**

| | Ordered Genotypes | | Acceptance |
|---|---|---|---|
| | 1-1/2-2 | 2-2/1-1 | rate |
| ESIP | 0.4977 | 0.5023 | 1.0 |
| Original Sobel & Lange | 1 | 0 | 0.084 |
| Haplotype sampler (HT) | 0.4025 | 0.5975 | 0.369 |
| Cascading Origin (CO) | 0.5061 | 0.4939 | 0.274 |

SL-sampler this has no effect, i.e. the chain is locked in the starting configuration due to the same reasons as above. Offspring 111 being a likely recombinant, the HT-sampler does not reach the true posterior distribution, since it tends to eradicate all recombinants between tightly linked loci. Acceptance rates stay as in the previous example, except that the recombinant offspring 111 lowers it for the HT-sampler. Note that all samplers demonstrated here are theoretically irreducible. Thus it is practical irreducibility encountered in these examples caused by strong horizontal dependence. More complex pedigree structures and genetic models, when a QTL is sampled also, will likely cause the SL- and the HT-sampler to fail in practice. In these examples, the CO-sampler performed well, although considerably less efficient that the ESIP-sampler. Therefore, when only a limited number of markers is considered (e.g. interval mapping), then the ESIP-sampler performs best since it samples the whole pedigree jointly conditional on the data. If more loci need to be accounted for, jointly sampling all genotypes for the whole pedigree becomes inefficient. Combining the SL-, HT- and the CO-sampler is then expected to be more efficient. Such a combined sampling strategy is implemented in preliminary version of MATVEC with the number of SL-, HT- and CO rounds to be determined by the user. In an upcoming contribution, we will compare all the samplers presented here, including the grouping strategies of Du and Hoeschele (2002) on more complex pedigree structures. This will then also access the properties of the ESIP-sampler when proposals are drawn from a modified pedigree and those of the Du and

Hoeschele grouping strategies updating groups of segregation indicators conditional on observations.

**ACKNOWLEDGEMENTS**

This research was supported in part by NSF (DBI-9723022), USDA National Research Initiative Competitive Grants Program (96-35205-3662) and the Swiss National Science Foundation.

**REFERENCES**

Bink, M.C.A.M. and Van Arendonk, J.A.M. (1999) *Genetics* **151** : 409 – 420.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1978) *Adv. Appl. Prob.* **10** : 26-61.

Du, F.X. and Hoeschele I. (2002) *Technical Report, Department of Dairy Science, Virginia Tech., U.S.A.*

Elston, R.C and Stewart, J. (1971) *Hum. Hered.* **21** : 523-542.

Fernando, R.L., Stricker, C. and Elston, R.C. (1994) *Theor. Appl. Genet.* **88** : 573-580.

Fernández, S.A., Fernando, R.L., Gulbrandtsen, B., Totir, L.R. and Carriquiry, A.L. (2002a) Genet. Sel. Evol. 33 : 337-367.

Fernández, S.A., Fernando, R.L., Gulbrandtsen, B., Stricker, C. Schelling, M. and Carriquiry, A.L. (2002b) *Genet. Sel. Evol*. (submitted)

Guo, S.W. and Thompson, E.A. (1994) *Biometrics* **50** : 417-432.

Hasstedt, S. J (1982) *Comput. Biomed. Res.* **15** : 295-307.

Heath, S.C. (1997) *Am. J. Hum. Genet*. **61** : 748-760.

Heath, S.C. (1998) *Hum. Hered.* **48** : 1-11.

Jansen, R.C. Johnson, D.L. and Van Arendonk, J.A.M. (1998) *Genetics* **148** : 391 – 399.

Janss, L.L.G., Van der Werf, J.H.J. and Van Arendonk, J.A.M. (1992) *Proc. 43rd EAAP Annual Meeting, Madrid, Spain* **1** : 144

Lange, K. and Elston, R.C. (1975) *Hum. Hered.* **25** : 95-105.

Lange, K. and Boehnke, M. (1983) *Hum. Hered.* **33** : 291-301.

Sobel, E. and Lange, K. (1996) *Am. J. Hum. Genet.* 58 : 1323 – 1337.

Sheehan, N.A. and Cannings, C. (2002) pers. comm. and *Genetics (in press)*

Schelling, M., Fernando, R.L. Kuenzi, N, and Stricker, C. (2000) *Proc. 51th EAAP Annual Meeting, The Hague, The Netherlands* **6** : 20.

Schelling, M. (2002) Ph.D. thesis, ETH Zurich, available in PDF-format from http://www.genetics-network.ch/downloads/

Thomas, D.C. and Cortessis, V (1992) *Hum. Hered.* **42** : 63-76.

Thompson, E.A. (1991) *Proc. 23rd symposium on the interface* 321-328.

Thompson, E.A. (1994) *Proc. of the 1994 Interface Conference, SAS, Cary, North Carolina, USA.*

Thompson, E.A. and Heath, S.C. (1997) *Stat. Mol. Biol. Genet.* **33** : 95- 113.

Wang, T., Fernando, R.L., Stricker, C. and Elston, R.C. (1996) *Theor. Appl. Genet.* **93** : 299-1309.

Wang, T., Southey, B.R., Katchman, S., Schelling, M. and Fernando, R.L. (2002*) MATVEC, Experimental version. Dpt. Anim. Sci., Iowa State University, Ames, U.S.A.*

Yi, N. and Xu, S. (2001) *Genetics* **157** : 1759 – 1771.