VARIABILITY IN THE DISTRIBUTIONS OF SINGLE NUCLEOTIDE POLYMORPHISM EFFECTS IN LIVESTOCK POPULATIONS

E.J. Smith and J.M. Henshall

F.D. McMaster Laboratory Chiswick, CSIRO Livestock Industries, Armidale NSW 2350

SUMMARY

Variability in the distributions of single nucleotide polymorphisms (SNP) effects were investigated by simulation. Realised distributions after thousands of generations of selection were found to be sensitive to parameters relating to the genome size, SNP density, population size and the distribution of sampled SNP effects. In particular, the distributions were not generally exponential, and in some cases SNPs of smaller effect were less likely to segregate than SNPs of larger effect.

INTRODUCTION

In association studies with dense SNP and phenotype data, Bayesian methods require assumptions regarding the distribution of SNP effects. However, data available to gain an understanding of the true but unknown distribution are limited. Firstly, the minimum sized quantitative trait locus (QTL) that can be detected will vary with the experimental design. Also, the properties of the trait and closely linked QTL can result in misleading estimates of size and effect. A meta-analysis of QTL effect distributions on pig and dairy data indicated effects were skewed with a few QTL of large effect (Hayes and Goddard 2001), and a review of QTL analyses in *Drosophila* concluded that the distribution of homozygous QTL is exponential with the larger effects contributing to most of the variation between parental lines (Mackay 2004); however, there has been no study of any statistical precision on the distribution of QTL effects (Roff 2007).

In this paper we present results of a simulation study that explored the sensitivity of the distribution of SNP effects to factors such as population size, genome size, SNP density and sampling distributions. Although the simulated population structures are modelled on sheep or cattle populations, we make no claims that the realised distributions are applicable to real livestock populations. Rather, the purpose of the study is to identify factors that might influence the distribution of SNP effects.

MATERIALS AND METHODS

Five population structures loosely based on a ruminant animal with annual joining were simulated for 105000 years (after equilibration) and data collected at years 1, 50000 and 100000 for a 5000 year period. All base animals were assigned as homozygous with allele "0" and equilibration was deemed to be the first instance when the frequency, q, of SNP's with allele "0" was equal to the frequency, p, of SNP's with allele "1", where p and q sum to one. Each year, 100 or 50 males were joined to 500 or 250 females, respectively, with animals eligible to enter the breeding herd between the ages of 2 and 6. Females had only one progeny per year. Selection was on a trait with a heritability of 25%. Animals were ranked according to their distance from a selection target.

Genomes with varying number of SNPs distributed over chromosomes of differing lengths (see Table 1 for details) were simulated with a SNP mutation rate of 3.1×10^{-4} per gamete. SNP effects were drawn from a uniform distribution, U, over (-5,5) or (-10,10); however, new effects were only sampled if the SNP was fixed (i.e., p = 0.0 or p = 1.0) in the current simulated population, reducing the realised mutation rate. Where the SNP was fixed with p = 1.0 the simulated population mean was adjusted to account for the change in the SNP effect. No polygenic variance

was simulated; all of the genetic variance was due to the SNPs. The desired heritability was achieved by holding the environmental variance constant at 60.0 and tuning the simulation as it ran to achieve the appropriate genetic variance of 20.0. To create genetic variance the selection target was moved a small amount each year, with the step-size dynamically tuned according to the size of the current realised genetic variance in relation to the desired genetic variance of 20.0. A second trait was also simulated, and after animals had been selected for breeding each year, mate allocations were assortative based on the second trait. Such an assortative mating system increases genetic diversity, and can be justified as accounting for a spatial component in wild or domestic populations. SNP transmission between animals as well as mutation and recombination events were recorded and written to a file after the simulation. The differences in parameters for the five simulated populations are outlined in Table1.

Table 1 – Population parameters used in each simulation (cM = centimorgans)

Simulation	# SNPs	SNP / cM	Dams	Sires	U
1	1600	16	500	100	(-5,5)
2	800	16	500	100	(-5,5)
3	1600	32	500	100	(-5,5)
4	1600	16	250	50	(-5,5)
5	1600	16	500	100	(-10,10)

RESULTS AND DISCUSSIONS

The distributions for the contribution of each individual SNP to the additive variance, $2pq\alpha^2$, and the absolute value of SNP effect sizes, α, are shown in Figure 1. The data is collated for $0.1 \leq p \leq 0.9$. Results for the three 5000 year periods of data collection (first two not shown) were similar, indicating the simulations had stabilized.

SNP effect distributions used as priors in Bayesian approaches are often assumed to be a decaying exponential function, which is a function whose second derivative is always greater than zero. Whilst the SNP effect distributions obtained for simulations 1, 3, 4 and 5 generally indicate that there are a greater number of SNPs with smaller effects than SNPs with large effect, the distributions contain inflexion points (second derivative is zero) and sections where the second derivative is less then zero (concave "down" shape) and therefore are functionally different from a decaying exponential distribution. In contrast, simulation 2 displays a uniform distribution. These results indicate that the distribution of α may differ depending on the simulation parameters and may not necessarily be exponential. In simulation 5 the distribution of SNP effects also shows that mutations on SNPs with effects sampled with a value of less than -6 or greater than 7 never survive beyond a frequency of 0.1. Further analysis also indicated that in simulation 5, mutations on SNP's with effects greater than the absolute value of 6 never became fixed in the population, indicating an upper limit to the effect size for mutations that can survive in a population.

The distribution of $2pq\alpha^2$ values were exponential in behaviour and do not vary significantly between the different populations. The possible exception is simulation 2 where the distribution appears more normal. This is due to the behaviour of the SNP effect distribution as opposed to the frequencies, as these are almost identical for simulations 1 and 2 (not shown). It should be noted that the frequency distributions were not identical for all 5 simulations. Increasing the SNP density increased the frequency of mutations segregating in the population, and as discussed below, increasing the range of the sampling distribution of SNP effects decreased the frequency of mutations segregating.

Figure 2 shows the distribution of $2pq\alpha^2$ as a function of p for simulations 1 and 5. Qualitative inspection indicates that mutations move through the population with varying "velocities" and that

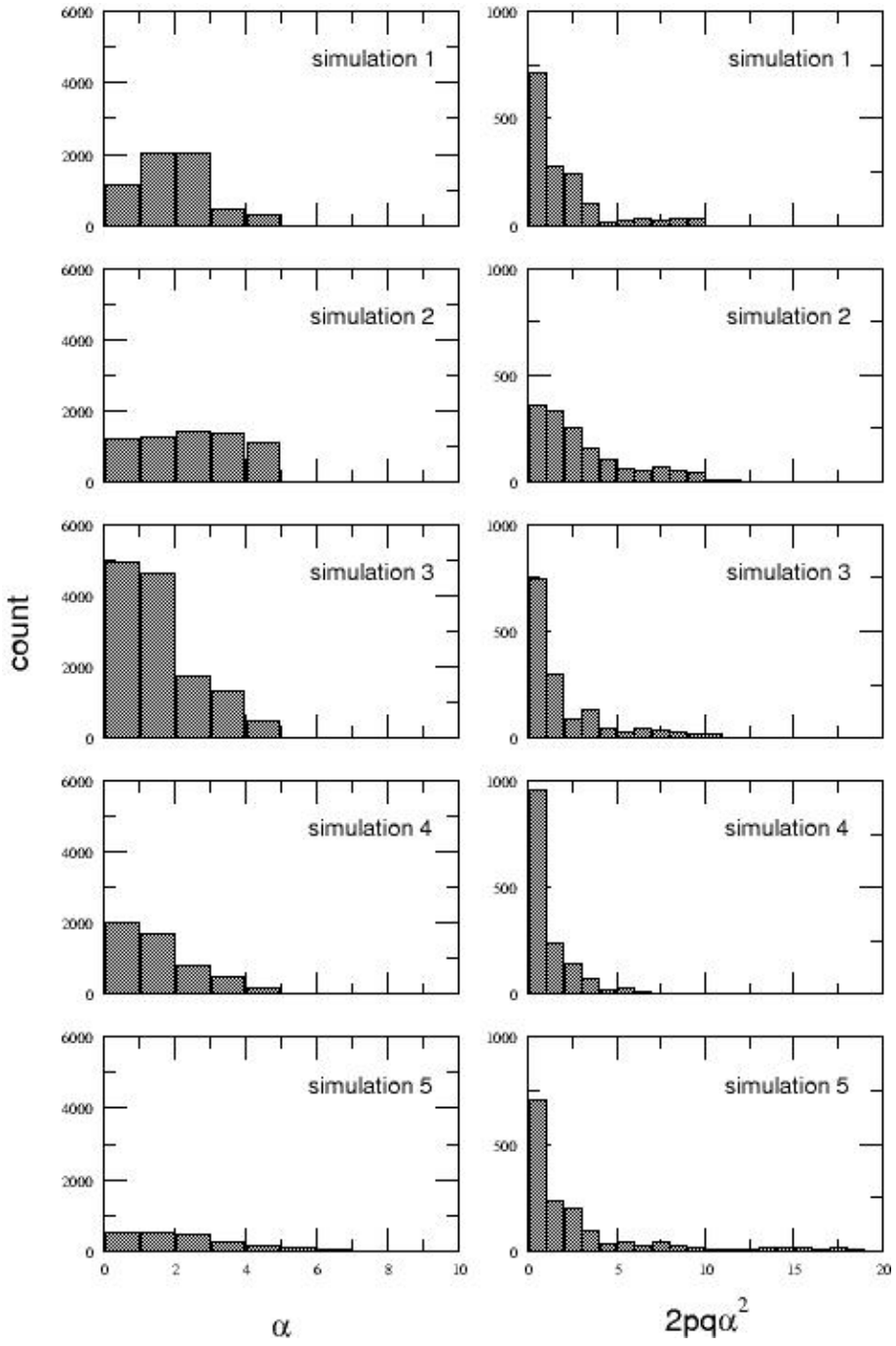

Figure 1. Distribution of the absolute value of the effect size, α, for $0.1 \leq p \leq 0.9$, and the contributions to the additive variance from individual SNPs, $2pq\alpha^2$, collected during the last 5000 years of the simulated populations.

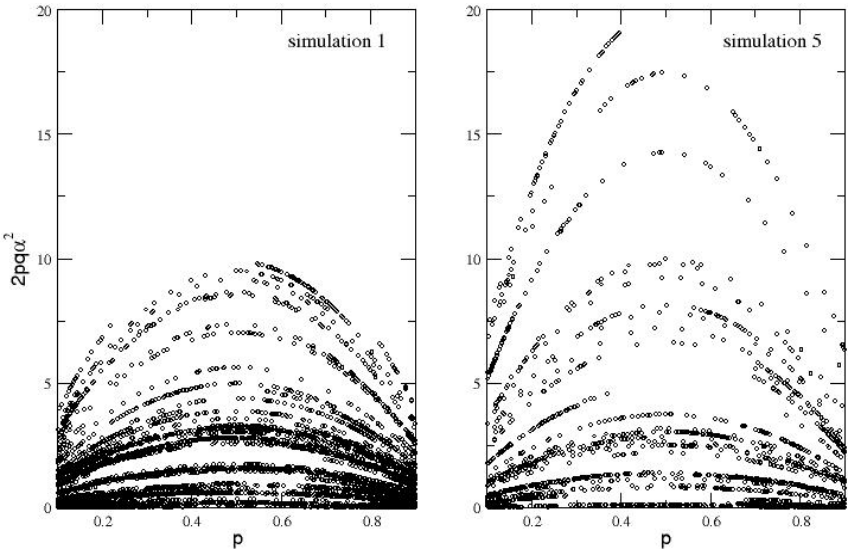

Figure 2. Data from simulations 1 and 5 showing the contribution of each SNP, $2pq\alpha^2$, to the additive variance as a function of the frequency, p, of SNP's with allele "1", collected during the last 5000 years of the simulated populations.

this observation is similar for both simulations. Quantitative analysis of the data showed that the average age to fixation and deletion was approximately the same for all simulations, and was also reasonably independent of the size of the SNP effects. However, Figure 2 also indicates that there are fewer mutations moving through the population for simulation 5 than simulation 1, and this was confirmed by extracting a histogram of the frequencies. The explanation may lie with the sampling distribution. As SNP effects are sampled from a uniform distribution, they are all initially equally as likely. Effects in simulation 5 are sampled from a distribution twice as broad as simulation 1, but the mutations with effects greater than those from the sampling distribution used for simulation 1 mostly do not segregate beyond p = 0.1. In other words, while the same number of mutations are occurring in each simulation, less of those in simulation 5 actually exist in the population for any significant period of time.

In this study we considered only one approach to producing a population with a SNP based genetic variance consistent with heritabilities observed in livestock populations. We recognise that this may also have a large influence on the distribution of SNP effects.

CONCLUSIONS

The distributions of SNP effects and their contributions to additive variances in livestock populations were investigated for populations that were simulated with different parameters. Results indicate not only a variability in the SNP effect distributions, but that the distributions do not consistently follow an exponential decay as effect size increases. This study supports that there is little to justify using particular distributions as priors for Bayesian analyses of SNP effects.

REFERENCES

Hayes, B. and Goddard, M.E., (2001) *Genet. Sel. Evol.* **33**:209
Mackay, T.F.C. (2004) *Cur. Opin. Genet. Dev.* **35**:253
Roff, D.A. (2007) *Evolution* **61**:1017