Proc. Assoc. Advmt. Anim. Breed. Genet. 20:122-125

GENETIC ARCHITECTURE AND EVOLUTION OF QUANTITATIVE TRAITS

M.E. Goddard[1,2]

[1]Department of Environment & Primary Industries, AgriBio, Bundoora, VIC 3083, Australia
[2]Melbourne School of Land & Environment, University of Melbourne, Parkville, VIC 3010, Australia

SUMMARY

Genome wide association studies in livestock and in humans typically find many SNPs of small effect and intermediate allele frequency associated with each quantitative trait. This is in contrast to theories that predict most variance to be due to rare alleles of large effect. This paper reports a computer simulation of the evolution of a quantitative trait under the effects of mutation, selection and genetic drift. The simulation can approximate the experimental findings but only by assuming that there are >1,000,000 sites at which mutation can affect a typical trait, mutation at these sites is much more likely to cause an allele of small effect than of large effect and selection against the mutant allele increases more than linearly with the size of the mutation's effect on the quantitative trait. Thus the experimental results are consistent with the theory for the control of genetic variation in quantitative traits at least under these assumptions.

INTRODUCTION

Quantitative or complex traits are important in agriculture, medicine and evolution but we have no good understanding of the forces that control genetic variation in these traits. Most theories explaining quantitative genetic variation assume that it is controlled by a balance between mutation, which creates new variants, and selection which eliminates these mutant alleles. Consequently, most versions of this theory predict that the genetic variance will be mainly due to rare mutations of large effect (eg Eyre-Walker 2010).

Until recently we had little knowledge of the genes that cause variation in quantitative traits and so this prediction was difficult to test. However, in the last 6 years assays for thousands of genetic markers or single nucleotide polymorphisms (SNPs) have become available for livestock and humans and this has allowed a new type of experiment known as a genome wide association study (GWAS). In a GWAS, individuals are measured for a trait and genotyped for thousands of SNPs. Then the SNPs are searched for those that are significantly associated with the trait. A SNP that is significantly associated with the trait is assumed to 'tag' a nearby mutation that causes variation in the trait because it is in linkage disequilibrium (LD) with this causal mutation (or quantitative trait locus, QTL). GWASs typically find many SNPs with small effects and intermediate allele frequencies but very few with large effects. This seems to contradict the theory that predicts that most of the variance will be due to rare mutations of large effect. In this paper I consider how the theory and the experimental results can be reconciled by using a computer model which simulates the evolution of a quantitative trait.

The simulation requires inputs concerning the number of mutations that can affect a typical quantitative trait, the size of their effects and the selection to which they are subject. Prior to the era of GWASs, some relevant information about the genetic architecture of quantitative traits was available. The variance added each generation by mutation is in the range 0.001 to 0.01 times the environmental variance (V_e) for most traits studied. This variance could be due to many mutations of small effect or few mutations of large effect or a mixture of both. In mice, many experiments have been reported in which a gene is 'knocked out' or replaced by an inactive form. As many as 1 in 3 of these knock outs affect body size (Reed *et al.* 2008). If this applies to most quantitative traits, it implies that over 5000 genes can affect each trait. Some mutations do have a large effect on a quantitative trait. For instance, mutations in the gene FBN1 in humans can cause Marfan's

syndrome which includes a large increase in height. Over 500 different mutations in FBN1 cause Marfan's syndrome (Kemper *et al.* 2012). FBN1 may be unusual, but if even 200 mutations can cause a large change in height, it is likely that even more mutations can cause a small effect on height. Therefore, it seems reasonable to expect that at least 5000 x 200 = 1,000,000 different mutations might affect a trait such as height.

It is usually assumed that natural selection favours an intermediate value for many quantitative traits. That is, individuals with extreme phenotypes are less fit than individuals with intermediate phenotypes. It is also likely that some mutations, which affect a quantitative trait, are detrimental regardless of their effect on phenotype for a particular quantitative trait. Zhang and Hill (2002) call this model of selection a joint effect model because it combines stabilising selection directly on the trait with selection directly against a deleterious mutation, and it is a joint effect model of selection that I have used in the simulation.

In this paper I compare GWAS results for weight in cattle with a simulation of the evolution of a quantitative trait under the influence of mutation, selection and genetic drift. The aim is to find a theory for the control of genetic variation in quantitative traits that is consistent with experimental GWAS results.

MATERIAL AND METHODS

GWAS. Bolormaa *et al.* (2013) analysed data from the Beef CRC on 6000 animals that were measured for post weaning weight and had genotypes for 700,000 SNPs. The method of analysis was called "BayesR' by Erbe et al (2012) and it fits all the SNPs simultaneously. The effects of the SNPs are assumed to be random variables drawn from a mixture of 4 normal distributions. The 4 distributions are such that SNPs in each of the distributions explain on average either zero, 0.0001, 0.001 or 0.01 of the genetic variance of the trait. The analysis used a Gibbs sampling chain and in each cycle the number of SNPs in each of the 4 distributions was counted. In this way the distribution of SNP effects and the variance they explain was estimated. The distribution of variances explained was compared to the simulation results.

Simulation. The computer simulation assumed a constant population size of 10,000. Each generation gametes are formed by recombination between the paternal and maternal gametes of the parent. Mutations occur in these gametes at a rate of 10^{-8} per site and there are 10^6 sites in the genome where mutation affects the trait. The effect of each mutation is drawn from a gamma distribution with a shape parameter of 0.1 and a scale parameter such that the variance added by mutation each generation is 0.001 times the environmental variance. The effect of the mutation is negative in a random 50% of cases. The parents mate at random and the offspring are subject to selection. The fitness of each offspring is obtained by multiplying together a fitness due to stabilising selection and a fitness which is constant for the mutation. The fitness from stabilising selection is $\exp(-0.5\ y^2/V_s)$ where y is the phenotype in environmental standard deviations and V_s is 200Ve. The constant fitness component of a mutation is 1-s where $s = 0.8\ x^2$ where x is the effect of the mutation on the trait in units of environmental standard deviations. After the simulation reaches an equilibrium state it is run for 1000 generations and the number of mutations segregating, their effects sizes and allele frequencies recorded each generation.

RESULTS

In the simulation, the effect of a mutation is drawn from a gamma distribution with shape parameter of 0.1. This is a distribution with many very small effects but a long tail of larger effects (Fig 1). Although mutations of large effect (eg. > 1 standard deviation) are rare they explain most of the mutation variance (Fig 1). However, mutations of large effect are strongly selected against and so they are kept rare and eventually eliminated by natural selection. Eventually when an equilibrium is reached, as much variance is lost each generation by genetic drift and selection as is

added by mutation. At this point the heritability of the trait was 0.33. Table 1 gives the distribution of mutation effects when an equilibrium state has been reached. Table 1 shows that mutations explaining less than 0.0001 of the total variance are common and mutations explaining more than 0.1 of the variance are rare.

The simulation includes the mutations that cause variation in the trait and these have been counted in Table 1 regardless of their allele frequency. However, a GWAS is based on SNPs that are not the causal variants but are hopefully in LD with them. Most of the SNPs on commercial SNP 'chips' such as used in our cattle GWAS, have a minor allele frequency (MAF) in the range 0.1 to 0.5. Therefore a causal mutation with MAF <0.1 cannot be in complete LD with a SNP with MAF > 0.1 and so the SNPs will underestimate the true effect of the causal mutation. The most optimistic assumption would be for causal mutations with MAF > 0.1 to be in complete LD with one of the SNPs and for causal mutations with MAF = q <0.1 to be in LD with a SNP that explains a fraction q/0.1 of the variance explained by the causal mutation. When this assumption is used to calculate the number of SNPs in each variance class (Table 1) the number of SNPs explaining >0.01 of the variance is much less than the number of causal mutations because many of these mutations are rare and hence incompletely detected by the SNPs. Consequently, the number of SNPs explaining <0.0001 of the variance is more than the number of causal mutations because it includes some causal mutations that explain a greater variance but are incompletely 'tagged' by the SNPs.

The BayesR analysis of weight in cattle provides a distribution of the effects of SNPs on weight. The distribution has been summarised (Table 1) by calculating the number of SNPs that fall into each proportion of variance class. The results are broadly similar to those predicted by the simulation model but the real data has even more SNPs explaining <0.0001 of the variance than predicted by the simulation.

Table 1. Number of segregating sites in the computer simulation and number of SNPs in the Bayes R analysis of cattle weight classified by the proportion of genetic variance that they explain.

Proportion of variance explained	Number of causal sites in simulation	Number of simulated SNPs	Number of SNPs from BayesR
$< 10^{-4}$	1466	1562	3166
10^{-4} to 10^{-3}	190	161	1492
10^{-3} to 10^{-2}	145	91	52
10^{-2} to 10^{-1}	21	8	5
$> 10^{-1}$	0.3	0	0.05

DISCUSSION

The simulation parameters might be regarded as extreme in certain respects. For instance, I assumed 1,000,000 sites in the genome affect a typical trait when mutated, that the distribution of their effects is very leptokurtotic (ie has a long tail) and that selection against mutations rises with the effect of the mutation squared. These assumptions all act to increase the importance of SNPs with small effects and decrease the number of SNPs with large effects when an equilibrium is reached. Despite this, the simulated data has fewer SNPs of small effect than the real data on cattle weight. To mimic the real GWAS results more closely the simulation would need to assume that >1,000,000 sites in the genome affect a typical trait when mutated. Thus the true parameters may be even more extreme than those assumed in the simulation.

If the theory of Eyre-Walker (2010) is applied to the parameters assumed in the simulation it also predicts that most of the variance is due to mutations of small effect. However, most authors have ignored this conclusion perhaps because they regarded the input parameters to be too extreme to be realistic. The conclusion of this paper is that they are not extreme enough.

Qualitatively the simulation matches an important feature of real data on quantitative traits. Mutations of large effect (> 2 standard deviations) occur for many traits. For instance, mutations causing dwarfism are known in many species but they are usually kept rare by natural selection so that they explain little of the total genetic variance.

The simulation results make a prediction with important practical consequences. The simulation predicts that there are a number of QTL segregating that explain > 1% of the variance but which go undetected by GWAS because their MAF is too low. This could explain the 'missing heritability' discussed in human genetics (Yang *et al.* 2010) and in cattle (Haile-Mariam *et al.* 2013). If this is indeed the case, the use of genome sequence data instead of SNP genotypes or haplotypes of SNPs should lead to the discovery of more QTL of medium size effects and their exploitation in genomic selection.

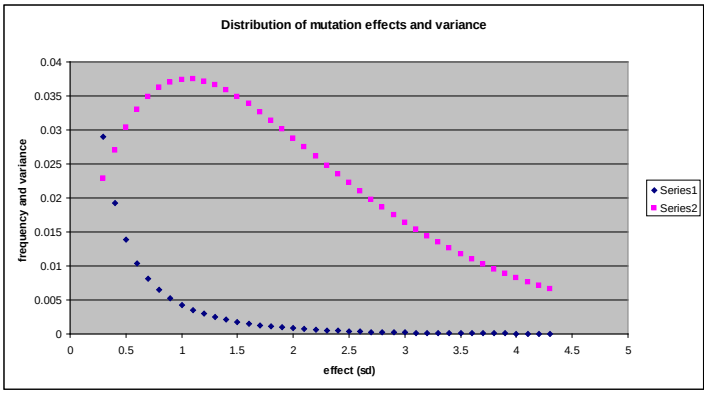

Figure 1. Distribution of the effects of mutations (series 1) and the mutation variance explained (series 2) by mutations of different size measured in units of environmental standard deviations.

REFERENCES

Bolormaa S., Pryce J.E., Kemper K., Savin K., Hayes B.J., Barendse W., Zhang Y., Reich C.M., Mason B.A., Bunch R.J., Harrison B.E., Reverter A., Herd R.M., Tier B., Graser H-U and Goddard M.E. (in press) *J. Anim. Sci.*

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich, C.M., Mason B.A., and Goddard M.E. (2012) *J.Dairy Sci.* **95**: 4114.

Eyre-Walker A (2010) *Proc, Nat. Acad. Sci*. **107**: 1752.

Haile-Mariam M., Nieuwhof G.J., Beard K.T., Konstatinov K.V. and Hayes B.J. (2013) *J. Anim Breed Genet.* doi:10.1111/j.1439-0388.2012.01001.x

Kemper K.E., Visscher P.M. and Goddard M.E. (2012). *Genome Biology*. **13**: 244.

Reed D, Lawler M, Tordoff M (2008) *BMC Genetics* **9**: 4.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A. Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, *et al* (2010) *Nature Genet.* **42**: 565.

Zhang X-S. and Hill W.G. (2002). *Genetics* **167**: 459.