Proc. Assoc. Advmt. Anim. Breed. Genet. 19:1-6

GENETIC ARCHITECTURE OF COMPLEX TRAITS

M.E. Goddard[1,2]

[1] Department of Agriculture and Food Systems, University of Melbourne, VIC 3010
[2] Department of Primary Industries, 1 Park Drive, Bundoora, VIC 3083

SUMMARY

The long term, almost linear response to selection observed in experiments and commercial livestock suggest that many genes control variation in most complex or quantitative traits and modern experiments with genetic markers support this conclusion and suggest that hundreds or even thousands of polymorphisms affect a typical trait. In fact, the variance explained by most quantitative trait loci is so small that single nucleotide polymorphisms (SNPs) near them do not reach significance even in large genome wide association studies. Even when the variance explained by all SNPs together is estimated in human genome wide association studies they only explain about half the genetic variance. This could be explained if quantitative trait loci tend to have a lower minor allele frequency than the SNPs used in experiments. This picture of the genetic architecture of complex traits is consistent with our knowledge of mutations that cause genetic variation. There are many sites in the genome at which mutations affect a complex trait and their effects vary from small to large. The mutations with a large effect tend to be eliminated by selection so that large numbers of small mutations explain the standing genetic variance but this selection causes the spectrum of allele frequencies to be biased towards low frequencies.

INTRODUCTION

Complex or quantitative traits are those controlled by many genes and by environmental factors. They are of great importance in agriculture (eg milk yield), evolution (eg clutch size of birds) and medicine (eg suffering from diabetes or not). Although, by definition, complex traits are not controlled by a single gene, we do not know which genes control most complex traits or even how many there are. (The polymorphisms affecting a trait may be in the DNA between genes but I will refer to genes for simplicity). The genetic architecture of these traits refers to the number of loci affecting them, the sizes of their effects and their allele frequencies, and the occurrence of non-additive interactions within and between loci. The genetic architecture, in turn, depends on the characteristics of mutations that affect a complex trait and the selection pressure and genetic drift that control their subsequent evolution. The genetic architecture is important in livestock improvement. For instance, if only a few genes influence a trait we might concentrate all our selection on those few genes, whereas if there are many genes for each trait, we need a strategy, such as genomic selection (Meuwissen *et al.* 2001), that uses markers covering the whole genome without necessarily identifying the genes involved. The efficiency of genomic selection is influenced by other features of the genetic architecture. For instance, if genes affecting quantitative traits (QTL) typically have one very rare allele and one common allele, it will be difficult to estimate breeding value for these traits because the single nucleotide polymorphisms (SNPs) on commercial assays do not have a very rare allele and so cannot be in high linkage disequilibrium with such a QTL.

Information about the genetic architecture of quantitative or complex traits comes from traditional experiments prior to the availability of genetic markers (classical information) and, more recently, from experiments using genetic markers. In this paper I will review the information from both sources on the genetic architecture of complex traits. Other reviews of this topic can be found in Hill (2009) and Mackay (2009).

INFORMATION FROM CLASSICAL EXPERIMENTS

If a trait shows a normal distribution and a high heritability it is usually safe to conclude that more than one gene is causing variation. However, it is almost impossible, from such simple information, to tell how many genes are affecting the trait. Other experiments such as F2 crosses between high and low lines provide some information. They suggest that many genes (>50) influence most traits (Mackay and Lyman 2005), but have little power to distinguish 50 from 1000.

Potentially selection experiments contain information about genetic architecture. If selection response continues for many generations it suggests that many genes are involved. If there are only a few genes causing variation, selection would soon fix the favourable allele at each gene, there would be no remaining genetic variation and selection response would stop. This is not the observed outcome in experiments with effective population size >50. In most experiments, that are carried on long enough, selection response lasts >100 generations (Hill and Bunger 2004) and the genetic variance is never extinguished, partly because new mutations continually add to it. Similarly, the response to selection continues unabated in livestock and poultry (Havenstein *et al.* 2003). In fact, Zhang and Hill (2005) found that an infinitesimal model fitted the data on selection response in Drosphila as well as models with a finite number of QTL. Selection does drive favourable alleles towards fixation which reduces genetic variance, but for a time at least, this is compensated by increasing the frequency of initially rare, favourable alleles, which increases the genetic variance until a frequency of 0.5 is reached (Goddard 2009).

The variance added by mutation each generation (V_m) is approximation $1/1000^{th}$ of the environmental variance (V_E) for most traits studied (Keightley and Halligan 2009). Considering that mutation is rare (about 10^{-8} per nucleotide per generation), V_m is surprisingly high, implying that mutations at many sites can affect a given trait and/or the effects of those mutations are large. For instance, if mutations at 25,000 sites in the genome affect milk yield and if these mutations changed milk yield by 1000L (over 1 standard deviation) each, then $V_m = 0.001\ V_E$. Of course, not all mutations will have the same effect, but if some mutations have only a small effect, the number of sites affecting milk yield (or any other complex trait) must be even greater than 25,000.

These surprising estimates of the number of sites affecting a complex trait and their large effects, are supported by experimental evidence. In Drosophila, mutations can be induced by random insertion of P elements. About 20% of insertions affect abdominal or sternal bristle number (Mackay and Lyman 2005). This could be interpreted to mean that >2000 genes affect a relatively simple trait such as bristle number. Even then, there would need to be many sites in these genes that can affect the trait if there are 25,000 sites altogether.

Experiments on the size of effect of mutations appear contradictory (Keightley and Halligan 2009). Experiments that detect mutations by their effect on the trait (mutation accumulation experiments) find large effects. However, experiments that examine the effect of a DNA polymorphisms segregating in the population find small effects. For instance, the effects of mutations that change an amino acid in a protein (non-synonymous coding mutations) have a leptokurtotic distribution of effects with many small effects and rare large effects. It may be that the mutations of large effect are eliminated by selection and so are not present in population samples (Keightley and Halligan 2009).

If these mutations affecting a complex trait did not affect fitness (i.e. if they were neutral) they would accumulate until the number being added each generation balanced the number lost through genetic drift. If the effective population size is N_e , this balance occurs when genetic variance $V_g = 2N_e\ V_m$. Consequently if N_e =10,000 as in humans, V_g would reach an equilibrium at $2*10000*0.001\ V_E = 20\ V_E$ and h^2 would be 0.95. This is not what we observe so selection must be acting to eliminate many of the mutations that occur and which affect a complex trait. This selection is expected to cause allele frequencies at QTL to be close to zero or 1.0, that is, the minor

allele frequency (MAF) will be very low.

Inbreeding depression and heterosis are most easily explained by (partial) dominance of alleles that increase fitness at many loci. Unfortunately, we have little knowledge as to the number of genes contributing to inbreeding depression or heterosis. Selection is expected to oppose any increase in frequency in a deleterious allele caused by genetic drift and this should reduce inbreeding depression if inbreeding is slow (Ehiobu *et al.* 1989). Consequently, heterosis in crosses between breeds, which have been inbred only slowly, should be small. The fact that heterosis is not small suggests that it is due mainly to genes of small effect that are subject to very weak natural selection (Goddard and Ahmed 1982).

Accurate estimates of dominance and epistatic variance are very hard to make, but those available suggest these variances are less than the additive variance for most traits (Hill *et al.* 2008).

INFORMATION FROM EXPERIMENTS WITH GENETIC MARKERS

Genome wide association studies (GWAS) use thousands of single nucleotide polymorphisms (SNPs) that cover the whole genome to map QTL, based on the assumption that all QTL will be in linkage disequilibrium with nearby SNPs and so create an association between the SNPs and the trait. GWAS find SNPs associated with almost every complex trait scattered over the whole genome implying that there are many QTL affecting each trait. If there are many QTL for each trait, the variance explained by each QTL must be small, but just how small the effects of most QTL are has come as a surprise. Few QTL explain more than 1% of the genetic variance in each trait. To increase the power of experiments to detect genes of small effect, large sample sizes have been used in meta-analyses that combine many independent GWAS. In humans, records on 183,727 people were used in one meta-analysis of height (Lango Allen *et al.* 2010). Even when the most stringent significance tests are applied ($p<5x10^{-8}$), 180 QTL for human height were detected and confirmed in additional populations. The low power of even this huge study implied that there were about another 600 QTL similar to those already detected. The largest QTL explained 0.4% of the genetic variance.

The 180 significant and replicated SNPs together explain only 12% of the genetic variance of human height (Lango Allen *et al.* 2010). This has been called the missing heritability paradox. Using a different approach, we estimated that all 300,000 SNP together explain half the genetic variance in human height (Yang *et al.* 2010). The difference between 12% and 50% is due to SNPs with such a small effect that they were not significant. The remaining 50% of the genetic variance is missing because the QTL are not in complete LD with the SNPs on the commercial chip. Ten percent out of the missing 50% is due to the finite number of SNPs used (300,000 is not enough) and the other 40% is because the QTL have different properties to the SNPs. For instance, if the QTL had minor allele frequencies <0.1 this could explain the lack of complete LD with the SNPs (Yang *et al.* 2010).

The genetic variance explained by most SNPs is small. The most that any SNP explains for human height is 0.004 of the genetic variance. However, there are exceptions to this generalisation. The polymorphism in DGAT explains 40-50% of genetic variance in fat% in milk of Holsteins (Hayes *et al.* 2009) and, when double muscling mutants are segregating, they explain a lot of the variation in the proportion of muscle in the carcase. Although mutations of large effect are usually selected against and so remain rare, occasionally one of these mutations is favoured by natural or artificial selection. When this happens, the mutant allele increases in frequency and so, for a time, explains a large amount of variance in the trait. However, if the selection remains constant the mutant allele will eventually be fixed and therefore no longer contribute to the variance. As well as the QTL that explain a small but significant part of the genetic variance (eg 0.004) there are likely to be other QTL with even smaller effects. Therefore the distribution of

effects appears to be J-shaped with many QTL of very small effect and a small number with large effects.

The variance explained by a QTL depends on its effect on the trait and the allele frequency. A small variance explained could be due to a large effect but a very small minor allele frequency (MAF). This does occur as shown by major mutations that are typically deleterious and kept rare by natural selection. However, the SNPs associated with a trait are usually not rare. It is possible that the QTL that the SNP is tracking has a much lower MAF than the SNP. Nevertheless, it seems likely that many QTL of small effect are not especially rare. For instance, QTL segregating in multiple breeds are unlikely to be rare or the rare allele would have been lost in most breeds.

Dominance and epistatic effects are common among genes with large effects (Carlborg and Hayley 2004). However, non-additive variances are usually not large. This could occur despite the existence of non-additive gene effects because non-additive variances are only large if all the alleles are at intermediate frequencies (Hill *et al.* 2008). It is also possible that QTL of small effect tend to act additively. In most GWAS no evidence of non-additive effects has been found (Lango Allen *et al.* 2010).

Imprinting and even more bizarre patterns of inheritance, as shown by the calligyge gene, are known to occur but there is no convincing evidence that they explain a large part of the genetic variance. Even if non-additive variance was important, this would not be an explanation for the "missing heritability" because narrow sense heritability does not include non-additive variance. Epimutations (heritable changes in DNA methylation or chromatin acetylation) that last only a few generations would cause "missing heritability" in GWAS but they can not be a major source of heritable variation or else selection responses would not accumulate over many generations as they do.

GENOMIC SELECTION

Genomic selection is the use of a panel of genome wide, dense SNPs to predict the breeding value of an individual (Meuwissen *et al.* 2001). For many traits, a method of genomic selection in cattle that assumes an infinitesimal model performs as well as other methods, implying that the number of QTL must be large (Hayes *et al.* 2009).When methods of prediction analogous to genomic selection are used in humans, the accuracy of prediction improves as more SNPs are added and doesn't reach a plateau until >1000 are used (Lango Allen *et al.* 2010).

Within a breed, such as Holstein, <50,000 SNPs are needed to provide an accurate prediction of breeding value (Goddard *et al.* 2010). This is because the recent effective population size of Holsteins is small (~100) and so LD extends for a long distance and SNPs can be used to track QTL some distance away on the chromosome. Also multiple SNPs could combine to track a QTL even if its MAF were lower than any of the SNPs. However, the accuracy of the prediction equation may decline rapidly over generations, because favourable SNP alleles are pushed to fixation despite the QTL continuing to segregate and because rare favourable QTL alleles are not being selected at all (Muir 2007; Goddard 2009). This problem of declining accuracy can be partially overcome by continually re-estimating the prediction equation using data on recent animals.

If multiple breeds or breeds with higher N_e are considered, then much larger numbers of SNPs are needed. In this case, genomic selection is probably more accurate if only some of the polymorphisms are included in the prediction equation as is the case with the method called 'Bayes B' (Meuwissen *et al.* 2001). There is also an additional increase in accuracy if the actual QTL or causal polymorphism is included in the data. This can be achieved by using full genome sequence data instead of SNP genotypes (Meuwissen and Goddard 2010).

Genomic selection is a 'black box' method in that it a statistical approach with no attempt to find the genes and mutations actually causing variation. This has been an advantage of genomic

selection to date. However, with large number of polymorphisms available today and more in the future, a major part of the statistical analysis becomes deciding which polymorphisms have some association with the trait and which have none. We will be helped in this endeavour by biological knowledge of the genes and sites within genes where mutations affect our trait of interest.

Lango Allen *et al.* (2010) found that SNPs associated with human height were often near genes known to affect skeletal growth and they are often in LD with DNA polymorphisms causing an amino acid change in a protein or changing the expression of the nearby gene. Speliotes *et al.* (2010) found many SNPs near genes expressed in the hypothalamus and possibly controlling food intake, were associated with body mass index (a measure of overall fatness) in humans. However, Heid *et al.* (2010) found that SNPs associated with waist to hip ratio (a measure of fat distribution) were in different genes to those associated with body mass index and more likely to be in genes expressed in fat tissue.

Therefore I expect that more biological knowledge will be used in genomic selection in the future. This will also result in statistical models from genomic selection being used to find causal mutations rather than the 'one SNP at a time' models that have been used for GWAS to date. Consequently, genomic selection will be a major source of new biological discoveries that may have uses quite different to their use in genetic improvement such as new pharmaceuticals.

CONCLUSIONS

New mutations occur every generation and mutations at many sites in the genome can affect a typical complex trait. Many mutations have a very small effect on a given complex trait but there is a spectrum of effect sizes with some mutations having a large effect. Most of these mutations decrease fitness and are kept rare and eventually eliminated by natural selection. The mutations with large effects are especially likely to be unfavourable and face negative selection pressure that prevents them from contributing greatly to genetic variance. Mutations of small effect may be subject to such small selection pressure that their frequency drifts randomly due to finite population size. Some drift to intermediate frequency and collectively explain a fraction of the genetic variance. Occasionally a mutation of large effect is favoured by artificial or natural selection and increases in frequency until it causes a large part of the genetic variance for one or more traits. In total, QTL display a range of MAF from mutations that are very rare to those that are common in multiple breeds. However, even if QTL were neutral most of them would have low MAF and, since they are not all neutral, the majority must have low MAF.

IMPLICATIONS

Most genetic variance is due to QTL which, individually, explain a very small proportion of the genetic variance. Consequently, very powerful experiments are needed to find these QTL and this requires large numbers of animals with phenotypes and genotypes. In human genetics, this is being achieved by meta-analyses that combine several independent experiments to maximise sample size. This collaborative approach would also be very beneficial in livestock.

QTL will typically have lower MAF than the SNPs on commercial SNP chips. Consequently, GWAS will underestimate the size of QTL effects and fail to detect some. This problem can be overcome by using genome sequence data instead of SNP genotypes. Using genome sequence data, which contains the causal mutation or QTL, we will be able to detect QTL that have no SNP in high LD with them. However, each QTL will still only explain a small amount of the variance and so we will still need powerful experiments with large numbers of animals.

Genomic selection will use full genome sequence data to predict the breeding value of individuals. The prediction equation will select, from millions of polymorphisms, those that affect a particular trait and this selection will utilise biological knowledge about the genes and sites in the genome affecting a trait and contribute greatly to increasing our knowledge of these genes and

sites. Selection candidates will be genotyped with an inexpensive SNP chip but will have full genome sequence imputed by using a reference population that have been sequenced.

REFERENCES

Carlborg O. and Haley C.S. (2004) *Nat. Rev. Genet*. **5**: 618.
Ehiobu N.G., Goddard M.E. and Taylor J.F. (1989) *Theoretical and Applied Genetics* **77**: 123.
Goddard M.E. (2009) *Genetica* **136**: 245.
Goddard M.E. and Ahmed A.M. (1982 *2nd World Cong. Genet. Appl. Livestock Production* **8**: 377.
Goddard M.E., Hayes B.J. and Meuwissen T.H.E. (2010) *9th World Congress on Genetics Applied to Livestock Production (Leipzig)* Paper 0701
Havenstein G.B., Ferket P.R. and Qureshi M.A. (2003) *Poultry Sci.* **82**: 1500.
Hayes B.J., Pryce J., Chamberlain A.J. and Goddard M.E. (2010) *PLoSGenetics* **6**: e1001139.
Heid I.M. *et al.* (2010) *Nat. Genet.* **42**: 949.
Hill W.G. (2009) *Phil. Trans. R. Soc. B* **365**: 73.
Hill W.G. and Bunger L. (2004) *Plant Breed. Rev*. **24**:169.
Hill W.G., Goddard M.E. and Visscher P.M. (2008) *Plos Genetics* **4** (2) e1000008.
Keightley P.D. and Halligan D.L. (2009). *Genetics* **136**: 359.
Lango Allen H. *et al.* (2010) *Nature* **467**: 832-838.
Mackay T.F.C. (2000) *Genetica* **136**: 295.
Mackay T.F.C. and Lyman R.F. (2005) *Phil. Trans. R. Soc. B* **360**: 1513.
Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001). *Genetics* **157**: 1819.
Meuwissen T.H.E. and Goddard M.E. (2010). *Genetics* **185**:623.
Muir, W. M., (2007). *J. Anim. Breed. Genet*. **124**: 342.
Pryce J.E., Bolormaa S., Chamberlain A.J., Bowman P.J., Savin K., Goddard M.E. and Hayes B.J. (2010) *Journal of Dairy Science* **93**: 3331.
Speliotes E.K., *et al.* (2010). *Nat. Genet*. **42**: 937.
Yang J., Beben B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.F., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E. and Visscher P.M. (2010) *Nat. Genet.* **42:** 565.
Zhang X.-S. and Hill W.G. (2005) *Genetics* **169**: 411.