Use of whole genome sequence data for QTL mapping and genomic selection

Theo Meuwissen[*]

Introduction

Current generation genome sequencing technology is about two-orders of magnitude faster and more cost effective than the technologies used for the sequencing of the human genome. It is more correct to call this resequencing technology, since it is used on species with known reference sequence. Future technologies are predicted to reduce cost by another 100 fold so that sequencing an entire human genome for ~$1,000 is considered achievable in the future (Mardis, 2008). Hence, in the near future we can expect to have whole genome sequence data available on substantial numbers of animals.

Whole genome sequence data differs fundamentally from current dense SNP-chip data in that (i), in contrast to traditional marker data, it is expected to contain all causal polymorphisms; (ii) the density of the polymorphisms is one to two orders higher; and (iii) some of the polymorphisms will be indels and other copy number variants (CNVs). In the case of high density SNP data, the SNPs are used to trace the causal polymorphism, whilst in the case of sequence data the neutral SNPs are merely a nuisance and hindrance for the detection of the causal mutation. Hence, prediction of genetic value using genomic selection reduces to a model selection problem, where the relatively few causal polymorphisms need to be selected amongst millions of others. But the detection of causal mutations is also the aim of QTL mapping, so QTL mapping and genomic selection have very similar goals, and the techniques for achieving these goals are also similar. Although, that QTL mapping remains more focused on positioning individual genes with large effects, and genomic selection remains more focused on addressing all genetic differences.

The aim of this paper is to explore how effective whole genome sequence data is in the mapping of QTL and in genomic selection, and how to make best use of this new type of data.

QTL mapping

Current state of the art. Current QTL mapping efforts follow very much a two stage approach, where step (i) maps the QTL to a region, which is ~1 cM when dense SNP-chip genotyping is used; and in step (ii) the most likely position candidate genes are pinpointed, i.e. genes which are within the QTL regions and which (Ron and Weller, 2007): (a) have a known physiological role within the phenotype studied; (b) affects the trait in knock-outs,

[*] Norwegian University of Life Sciences, Box5003, 1432 Ås, Norway

mutations or transgenics in other species; (c) the gene differentially expressed in organs related to the trait; and (d) the gene is differentially expressed during developmental stages related to the phenotype. Next these positional candidates are sequenced together with their promoter regions. The latter is because often the promoter region contains the causative mutation instead of the coding regions of the gene. The availability of a reference sequence very much helps to pinpoint the genes that are within the QTL region, and our ever improving knowledge about gene functions helps to identify the positional candidates. The sequencing results in a number of newly identified polymorphisms and genotyping the mapping population for these is expected to identify the most likely causative mutation.

There are a number of problems with this approach: Firstly, QTL mapping methods are quite poor in providing confidence regions for the QTL, in particular when a genome wide association study (GWAS) is used (Meuwissen and Goddard, 2007). Although bootstrapping can be used to estimate confidence regions (Visscher et al., 1996) this requires the mapping of the QTL in many bootstrap samples which is often computationally not possible. Secondly, many candidate genes fulfill some of the criteria of Ron and Weller (2007), whilst even known causative mutations often do not fulfill all of these criteria. Thirdly, the putative causative mutations need to be validated using genetic and functional studies. The proof of that a polymorphism is functional relies on multiple pieces of evidence, each in itself insufficient, but put together consistently point to the causative mutation (Mackay, 2001). Important genetic pieces of evidence are: (a) similar effects are found in other populations; (b) when fitted in a model no other polymorphism in LD with the mutation is still significant; (c) it accounts for the full effect of the QTL; (d) the CST test (see below).

The Causative SNP Test (CST) tests whether a putative SNP is causative: CST tests whether the SNP shows significantly more LD with the QTL than any other SNP (Uleberg and Meuwissen, submitted Gen. Sel. Evol.). In the CST, the estimates of the SNP effects are allocated to each SNP in the region in turn (e.g. 1000), and after adding noise, 1000 replicated, simulated data sets are obtained. Next these 1000 data sets are analysed in the same way as the real data, whilst leaving the known causative SNP out of the analysis. This provides a null-hypothesis distribution for the test-statistic, i.e. $\text{LogLik}(\text{SNP}_{(1)})$ minus $\text{LogLik}(\text{SNP}_{(2)})$, when the causative SNP was not included, where $\text{LogLik}(\text{SNP}_{(i)})$ is the log-likelihood of the SNP with the i-th highest likelihood. Next the real data test-statistic can be compared to these null-hypothesis test-statistics to find the P-value of the test that the original SNP is causative. The idea is that a causative SNP is not only Identical-by-Descent (IBD) with the QTL but also IBS (Identical-by-State), and thus gives a significantly higher test-statistic than (the best) linked SNP ($\text{LogLik}(\text{SNP}_{(2)})$. The power of CST was only 28%, but in case it gives a significant result it yields strong evidence for the causality of the SNP.

Another test whether SNPs are causative is the concordance test. In livestock QTL mapping designs either fullsib or halfsib family structures are used, which provides a quite accurate estimate of whether the parent is heterozygous or not. And, in case of a heterozygous parent, the phase between the the Q and q allele and any SNP allele. The concordance test, tests whether the SNP and the QTL allele have the correct phase in each of the heterozygous parents. Equations have been derived to calculate the probability of concordance under the null-hypothesis of independence between the putative SNP and the QTL, which provides P-

values for the test. However, the assumption of independence is obviously violated, otherwise the SNPs would not have given a significant QTL mapping signal. When mapping a coat color locus in dogs, Karlsson et al. (2007) could reduce the number of candidate causative mutations from 124 to 46, i.e. 37% past the test despite the null-hypothesis being true for all but one of the mutations, whilst the P-value was <1%. Although many genes passed the test, it resulted in a significant reduction of the list of candidate mutations.

State of the art in monogenic trait mapping. Monogenic traits are traits that are determined by single gene, and are not affected by the environment. Often they are disease traits, where the disease is either present or absent. Mapping of monogenic traits has benefited greatly from current dense SNP chip and sequencing technology (Karlsson et al., 2007; Charlier et al. 2008). The big advantage of monogenic trait mapping is that the (homozygote) carriers of the mutant allele are unequivocally identified. The small effective population size in livestock, makes that haplotype blocks are big and that highly significant SNP effects can even be found, when using relatively sparse SNP-chips. In the next step the confidence region is either sequenced entirely or candidate genes within the confidence region are sequenced together with their promoter regions. The unequivocal detection of the carriers leads to the requirement that a potentially causative mutation has to go always with the cases and the opposite allele always with the controls. This requirement often leads to few or just one identified potentially causative SNP. If the trait also segregates in other breeds, the potential SNPs can be tested in these breeds in order to further reduce the shortlist of SNPs. Finally, functional information and/or functional studies can be used to select or confirm the causative mutation.

Lessons from mono-genic trait mapping. The biggest difference between monogenic and QTL mapping is that in QTL mapping the phenotype is only an indicator of the QTL genotype, instead of identifying it. This makes it much more difficult to compare genomes with the mutant allele to those with the wild-type allele. But not impossible: as mentioned above, in livestock QTL mapping designs it is quite well possible to identify parents which are heterozygous for the QTL, especially when family sizes are large. Although, there are complications, such as which of the two haplotypes of one parent agree to those of the other, and the probability of heterozygosity is <100%, it is possible to compare haplotypes across familes (Andersson and Georges, 2004). Perhaps the biggest problem is that ~10 or more large heterozygous families are needed, which implies more than $10/(2q(1-q))$ large families in total, where q is the QTL allele frequency. This huge QTL mapping design compares to just ~5 cases and ~5 controls in a monogenic trait mapping experiment!

Another important lesson is that candidate causative SNPs need to be confirmed in other populations. For complex traits, the allelic effect of the SNP may be different in another population, due to interactions with background genes and different allele frequencies, but the direction of the effect is most likely be the same. Validation of candidate SNPs in other populations seems a very useful tool to identify causative mutations.

How whole-genome sequence data will revolutionize QTL mapping. The use of whole genome sequence data in QTL mapping makes that the aforementioned QTL mapping steps (i) and (ii) can be combined into one step. The most significant polymorphism is

immediately expected to be the causative mutation, because all causative mutations are in the data. Confidence regions will more often than not be so small that only one gene is implicated, so we do not need to choose between several positional candidates. The proof that the mutation is causal remains however as difficult a task as before. Although, in some fortunate cases the most significant mutation might be much more likely than the second most significant polymorphism, in which case the CST test will show a significant result.

As always, one of the SNPs will be most significant, but it will be important to define a set of SNPs which could also be causative. One way to do this is to define a 95% confidence SNP set, which is not a closed confidence interval, but a set of SNPs that together have a probability of 95% of containing the causative SNP. Let LR(i) be the likelihood-ratio of SNP_i where the SNPs are ordered from the most likely to the least likely, the q most likely SNPs are entering the 95% confidence SNP set, where q is the smallest number such that: $\frac{\sum_{j=1}^{q} LR(j)}{\sum_{i} LR(i)} \geq 0.95$, where the summation in the denominator is over all SNPs in the region, that is considered to contain the causative SNP. It should be attempted to test the 95% confidence SNP set in other breeds in order to confirm their association with the trait, which will most likely further reduce the number of potentially causative SNPs, hopefully to a single one.

The confirmation in other breeds is important, because SNPs that show an effect due to LD instead of being causative, may have a different LD coefficient in other breeds and thus will not be confirmed. Also, the QTL may not be segregating in another breed, but in this case the causative SNP should also not be segregating. In human genetics many of the detected SNPs have been confirmed in independent samples, but here sample sizes often exceed 10,000 genotyped and phenotyped individuals. Due to the genotyping efforts involved in genomic selection, such large data sets will also come available in livestock.

GWAS versus LDLA. In animal breeding, combined Linkage Disequilibrium and Linkage Analysis (LDLA) QTL mapping (e.g. Meuwissen et al., 2002) is popular for four reasons: (i) strong family relationships make that the use of linkage analysis information is very important; and (ii) Linkage Disequilibrium (LD) information is crucial because of few genotyped generations; (iii) family structures are complex and LDLA can deal with this complexity; and (iv), in the past, marker maps were often relatively sparse so that a haplotype based method is needed to pick up the LD. But LDLA mapping is computationally complex and demanding, and probably too demanding to apply to genome sequence data.

In human genetics, when faced with millions of SNPs, the simplest and fastest analysis possible is Genome Wide Association Study (GWAS), i.e. the phenotype is regressed upon the marker genotype for every SNP in turn, and the most significant SNPs are taken as an indication for a QTL. Since family structures are less prominent in human genetics, they can more easily be ignored, although this lead to problems with spurious associations (Seldin et al., 2004). In livestock, family structures are too important to be ignored, and a GWAS model should fit a polygenic effect next to the SNP genotype (Goddard and Hayes, 2009). The latter creates two problems: (i) the polygenic effect is fitted by a variance component, which implies one REML (Residual Maximum Likelihood) variance component analysis per SNP, which requires massive parallel computing. (ii) When the SNP is fitted as a fixed

effect, the degrees of freedom due to error is poorly defined and thus the test-statistic is not really known. The latter can be resolved by fitting the SNP genotypes as a random effect. The resulting loglikelihood-ratio test with half a degree of freedom (Stram and Lee, 1994) provides a relatively well described null-hypothesis distribution.

Both GWAS and LDLA suffer from a massive multiple testing problem: one test per SNP genotype is performed. The permutation test (Churchill and Doerge 1994), where data are randomly permuted in order to simulate data under the null-hypothesis, accounts for the multiple testing, but this is computationally not possible and the randomization of the data would also disrupt the complex family structure, i.e. it is only possible in the absence of family structure or in situations with a simple family structure. Fortunately, the deterministic approach of Piepho (2001) can be used to correct for the multiple testing. This approach also shows that LDLA is expected to suffer less from the multiple testing since the variability of the test-statistic across the genome is less than that of an association study (Meuwissen and Goddard, 2007), i.e. LDLA mapping is expected to have more power to detect the QTL. The high variability of the GWAS test-statistic, which is due to the high variability of LD (Hill and Weir, 2004), makes that the true QTL position is less likely to be enclosed in a 2-loglik-drop off confidence interval than when a similar confidence interval is constructed using LDLA. However, in case of whole genome sequence data, confidence regions are less important because the two-step procedure for mapping QTL is abandoned.

In the case of whole genome sequence data, GWAS is clearly the method of choice, since it considers the correct model, i.e. the model where the causative polymorphism is fitted, whereas LDLA uses markers only to predict IBD (Identity By Descent) at a position, i.e. it does not consider the mutations as causal. At lower densities, GWAS may still be the method of choice because (i) there is still a possibility that some SNPs are causal (e.g. when using >500k SNP-chip), and (ii) some SNPs may show a very high LD with the causative mutation. At a SNP density of $1*N_e$ SNPs per Morgan, where N_e is the effective population size, the mapping precision of GWAS and LDLA was similar, but LDLA showed a higher log-likelihood ratio test statistic, and thus had more power to detect the QTL (Meuwissen and Goddard, 2007; Uleberg and Meuwissen, J. Anim. Breed. Genet. 2010, in press). Marker density is expressed here in proportion to N_e and distance, since the LD, which is used to map the QTL, is a function of N_e times the distance between the markers (d), i.e. $E(R^2) = \frac{1}{1+4N_e d}$ (Sved 1971). Thus, if N_e in livestock is ~200 and using a 60 kSNP-chip, mapping precision will be similar for GWAS and LDLA, but LDLA may detect some more QTL. However, when using genome sequence data, GWAS will have both the highest precision and power, since it tests not only for IBD at the putative position but also for AIS (alike-in-state), i.e. the causative SNP must also be AIS with the causative mutation (Uleberg and Meuwissen, J. Anim. Breed. Genet. 2010, in press).

Population stratification. Population stratification is an important problem in both GWAS and to a lesser extend in LDLA. To a lesser extend because LDLA is careful in requiring a linkage and a LD signal at the putative QTL position instead of solely relying on LD. Population stratification is often remedied in livestock mapping by fitting a polygenic effect next to the QTL effect in the model. The covariance structure of the QTL effect is based on the pedigree of the mapping population. In case of a mixture of breeds, a genetic group effect

can be added (Westell et al., 1988). However this only accounts for the known structure in the population. In case of dense SNP data, the relationships between the animals can be based on the SNP genotypes (VanRaden, 2008). In this situation, the putative SNP has to prove itself against all other SNPs fitted in the model, which are fitted as a polygenic effect. In case of stratification, some SNPs can pick up the stratification effect, and this effect will thus not be allocated to the putative SNP, i.e. the false positive rate will be as expected.

Single vs. multiple QTL mapping approaches. Multipoint QTL mapping is expected to reflect the underlying genetic model better than single QTL mapping. Particularly, the MCMC multiQTL mapping approaches make the QTL peaks more sharp. This is because they test whether there is a QTL at the putative position conditional on all other QTLs in the model, whereas single QTL mapping tests whether there is a QTL at the putative position or no QTL at all at this chromosome. The former seems the more appropriate test. The latter results in a larger confidence interval, and a larger confidence interval is often more realistic in QTL mapping. In a comparison, Uleberg and Meuwissen (2007) found that single QTL mapping is somewhat more precise in mapping a single QTL, but multiple QTL mapping could disentangle two closely linked QTL, which were mapped by single QTL mapping as one ghost QTL in between the two real QTL. Hence, it seems that both single and multiple QTL mapping have a role: single QTL mapping provides the most accurate position and multiple QTL mapping may disentangle closely linked QTL. An additional advantage of single QTL mapping is the greater simplicity of the model.

Genomic selection using whole genome sequence data

How to make best use of whole-genome sequence data in animal breeding? Whole-genome sequence data provides information on the genes, and thus it's use will be to predict genetic value at a young age of the animal. Genomic selection models generally fit SNP effects as if they were the genes. Thus presenting these models with sequence data makes their assumption that the SNPs have effects more realistic: at least some of the SNPs are causative and thus have a direct effect. The problem with sequence data is however that the number of SNPs is increased one or two orders of magnitude, which makes it harder to distinguish the causative SNPs from all the others, i.e. the n<<k problem is greatly enhanced.

A simulation study. Meuwissen and Goddard (Genetics, 2010, Epub. ahead of print) studied the accuracy of genomic selection when using whole genome sequence data. The simulated population had evolved at Ne=1000 for 10,000 generations, and had a 1 Morgan genome. Mutation rate was 10^{-8} per base-pair per meiosis, which resulted in ~33,000 SNPs, of which 30 were chosen to be causative. Next 20 more generations G0 – G20 were created, where effective size was increased to 10,000 in order to avoid close family relationships. In generation G10, 200 training animals were sequenced and phenotyped to estimate the SNP effects. Also 500 evaluation animals were sampled and genotyped from generation G10, and similarly 10 generations later (generation G20). Heritability was 0.5. BayesB or BLUP (Meuwissen et al., 2001) was used for prediction of genome-wide EBV (GWEBV).

Interpretation of simulation results. The above and other simulation studies usually simulate a very particular situation, which may be different from practice. Fortunately,

accuracies of simulation can easily be extended to different circumstances due to the following equation (Daetwyler et al., 2008; Goddard 2009):

$$r^2 = \frac{Th^2}{Th^2+4N_e L\nu} \quad [1]$$

where r is the accuracy of genomic selection; T is the number of training records; L is the genome size; $4N_eL$ is the expected number of segments in the genome; and ν translates the actual number of segments to the effective number (some segments are very small, which reduces their importance, ie. ν<1). Equation [1] results in a number of predictions:

1) If the genome size doubles (triples), we need twice (three times) as many training animals to maintain the accuracy (assuming constant marker and QTL density).
2) If historical N_e doubles, we again need twice as many training animals.
3) If h^2 halves, we need twice as many training animals

These predictions were tested by Meuwissen (2009) and found approximately correct. As an example, a cattle population with N_e=200, L=30 and T=1200 training animals is thus predicted to have about the same accuracy as the aforementioned simulated data set.

Simulation results. The accuracy of the GWEBV of valuation animals from generation G10 was 0.826, and this reduced to 0.806 when the causative SNPs were omitted from the data. Hence, the inclusion of causative SNPs did improve the accuracy even at a density of 33,000 SNPs per chromosome, but not all that much. It also shows that, even in sequence data, not all SNPs are tagged by another SNP, ie. are in perfect LD with at least one other SNP. If the evaluation animals were from generation G20, i.e. there were 10 generations between the training and evaluation animals, accuracy reduced marginally to 0.824. This is in sharp contrast to the findings in simulation studies with less dense marker maps, where the accuracy decreased markedly with the number of generations between training and evaluation animals (e.g. Habier et al., 2007; Sonesson and Meuwissen, 2009). In case of sequence data, either the causative SNPs are found or the LD between the SNPs with estimated effects and those with true effects is very high, such that the LD does not decay over time. Using sequence data may be important in breeding schemes where we need to predict GWEBV of remotely related animals, since accuracy hardly decreases with distance.

When BLUP instead of BayesB was used to predict the GWEBV, accuracy decreased from 0.826 to 0.493. This shows that, if the GWEBV estimation method does not give extra weight to the most important SNPs, the effect of those SNPs is diluted by all the ~33,000 other SNPs, and the advantage of using sequence data is lost. The same reasoning may hold when moving from a 50k to a 600k SNP chip in cattle: this will only increase accuracy if extra weight is given to the most important SNPs, as BayesB does. The accuracy of using whole genome sequence versus that of using a SNP chip with 1,000 SNPs per Morgan was 0.826 versus 0.443, showing that there is a very significant gain in accuracy when moving to sequence data.

Cost of sequence data. Even at a costs of 1,000$ per sequenced genome, it is still very costly to genotype 10,000's of training and evaluation animals. These costs may however be alleviated by sequencing a relatively small number of founder animals in the pedigree, i.e.

animals that have together contributed the majority of the genes that are present in current animals. Current animals may then be genotyped by a relatively sparse SNP chip, and the missing genotypes may be imputed (Hayes and Goddard, 2009).

The Future.

Genomic selection and QTL mapping require the same kind of data, such that the huge genotyping efforts that are and will be spent in genomic selection will also greatly improve our abilities to identify causative mutations. This will result in a great increase in the understanding of the biology of traits that are important for livestock production. However, it will hardly increase genetic improvement: Odegaard et al. (2009) found that knowing the causative mutation of a major gene hardly improved total genetic gain in a genomic selection breeding program, although it did change the direction of the selection, which was more directed towards increasing the frequency of the major gene. Perhaps more important for genetic gain is the increased knowledge about distributions of additive, dominant, epistatic and pleiotropic effects of QTL and biological pathways, which may lead to further improvements of the models used for genomic selection.

References

Andersson, L., Georges, M. (2004). *Nature Reviews, Genetics,* 5:202-212.
Charlier, C. et al. (2008). *Nature Genetics,* 40: 449-454.
Churchill, G., and Doerge, R. (1994). *Genetics,* 138: 963–971
Daetwyler H.D., Villanueva B., Woolliams J.A. (2008). PLoS One, 3:e3395
Goddard, M.E. (2009). *Genetica,* 136:245-57
Goddard, M.E., Hayes, B.J. (2009). *Nature Reviews, Genetics,* 10: 381-391.
Habier D., Fernando R.L., Dekkers J.C. (2007). *Genetics,* 77:2389-97
Hill W.G., Weir B.S. (1994). *Am J Hum Genet.,* 54:705-14.
Karlsson, E.K., et al. (2007). *Nature Genetics,* 39: 1321-1328.
Mackay, T.F. (2001). *Annual Review of Genetics,* 35:303-339.
Mardis, E. R. (2008). *Annu. Rev. Genomics Hum. Genet.,* 9: 387-402.
Meuwissen, T.H.E. (2009). *Gen. Sel. Evol.,* 41:35
Meuwissen, T.H.E., Goddard M.E. (2007). *Genetics,* 176:2551-60.
Meuwissen T., Karlsen A., Lien S., Olsaker I., Goddard M.E. (2002) *Genetics,* 161:373-9.
Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Genetics, 157:1819-29
Odegård J., Sonesson A.K., Yazdi M.H., Meuwissen T.H.E. (2009). *Gen.Sel.Evol.,* 41:38
Piepho, H-P. (2001). *Genetics,* 138: 963–971
Ron, M., Weller, J.I. (2007). *Animal Genetics,* 38: 429-439.
Seldin, M.F. et al. (2004) *Genome Research,* 14: 1076-1084.
Sonesson A.K., Meuwissen T.H.E. (2009). *Gen. Sel. Evol.,* 41:37.
Stram, D. O., and Lee, J. W. (1994). *Biometrics,* 50: 1171–1177.
Sved, S.A. (1971). *Theor. Pop. Biol*., 2: 125-141.
Uleberg E., Meuwissen T.H.E. (2007). *Genet. Sel. Evol.,* 39:285-99
VanRaden P.M. (2008). *J. Dairy Sci.,* 91: 4414-4423
Visscher, P.M., Thompson, R., Haley C.S. (1996) *Genetics,* 143:1013-20.
Westell, R.A., Quaas R. L., and Van Vleck L. D. (1988) *J. Dairy Sci.,* 71: 1338-1345