Bruk av genomsekvensdata for å predikere genetiske verdier av husdyr

THEO MEUWISSEN[1]
IHA-UMB, Ås, Norway[1]

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.

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

How to make best use of whole-genome sequence data?

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.

Methods: A computer simulation study

Meuwissen and Goddard (2010) 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).

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.

Extension to practical situations

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):

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

where r is the accuracy of genomic selection; T is the number of training records; L is the genome size; 4NeL 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 Ne doubles, we again need twice as many training animals.

3) If h2 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 Ne=200, L=30 and T=1200 training animals is thus predicted to have about the same accuracy as the aforementioned simulated data set.

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).

Application in breeding programs

General: In contrast to traditional selection, genomic selection does not require that accurate pedigree and phenotypic recording of elite breeding animals. Pedigree recording is not required because genomic selection is based on genotyping. Phenotypic recording of elite animals is not needed because, once the SNP effects have been estimated, accurate breeding values estimates can be obtained by combining the genotypes of the elite animals with the SNP effects, i.e. no phenotypic records are needed. This reduced need of recording of elite

breeding animals makes completely different breeding designs possible together with selection for new traits. In case of an expensive trait, for instance, one can set up an experiment where the trait is recorded and the animals are sequenced, which makes it possible to effects of all SNPs. Next one can select for these SNP effects for several generations, before re-estimation is required. In this way one can select for traits that are not widely recorded, as is required for traditional selection.

Cattle breeding: In cattle breeding, genomic selection is received with great enthusiasm, because it overcomes the problem that the most intensely selected animals, ie. the dairy bulls, do not have performance records. In traditional breeding, this problem is overcome by the progeny testing scheme, where bulls obtain progeny records instead. However, progeny testing of bulls is expensive and takes a lot of time, which slows down the breeding program. GENO will therefore at the end of 2010 introduce genomic selection, as a preselection step to improve the genetics of the young bulls that enter the progeny test. If this is successful, GENO will consider replacing the progeny test based selection by genomic selection.

Pig breeding: In pig breeding, genomic selection could mainly be applied to the selection of test-station boars for maternal traits, which can currently only be measured on female relatives of the boars. In a similar way, the selection for carcass traits will be greatly improved by genomic selection. Traditionally, carcass traits cannot be collected on elite breeding animals since its recording requires slaughtering the animals. Also, in pig breeding, SNP effects could be estimated using crossbred performances data in practical herds. This would make that genomic selection would automatically select for crossbred performance in practice instead of for performance of the purebred pigs at the testing-station.

Salmon breeding: In Salmon breeding, the genetic improvement of disease resistance is very important. Traditionally, sibs of the selection candidates are challenge tested for the disease. The sibs cannot themselves be used for breeding, because of the disease risk. Again this is a situation where the elite breeding animals do not have own performance data, and genomic selection can be used to obtain accurate breeding value estimates for these animals and traits.

References

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
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. (2010) Genetics 185,623-631
Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Genetics, 157:1819-29
Sonesson A.K., Meuwissen T.H.E. (2009). Gen. Sel. Evol., 41:37.