Haplotype Tests for Diagnosis of QTL and Genes

***J. M. Henshall*[1], E.K. Piper[2] and B. Tier[3]**
[1]Food Futures Flagship, CSIRO Animal, Food and Health Sciences, Armidale, New South Wales, Australia
[2]TheUniversity of Queensland, School of Veterinary Science, Gatton, Queensland, Australia
[3]Animal Genetics and Breeding Unit, University of New England, Armidale, Australia
(Animal Genetics and Breeding Unit is a joint venture with NSW Department of Primary Industries)

ABSTRACT: Tests based on haplotypes are usually considered to be temporary measures until causal mutations are found. An alternative view is that for livestock improvement they may be sufficient. Haplotype tests can be delivered quickly and cheaply, requiring no knowledge of the causal mutation. Appropriately applied, they are robust to uncertainty in the number of alleles at the mutation. With a haplotype based test for polled in beef cattle we found that most animals carried common haplotypes, and for these there was very high concordance in their association with the underlying genotypes. Phenotyped animals tested commercially can contribute to the accuracy of haplotype effect estimates, especially for rare or previously unseen haplotypes. For livestock, haplotype based tests may be competitive with causal mutation tests in delivering genetic improvement.
Keywords: haplotype; QTL; beef cattle; polled

Introduction

Tests based on the known causal mutation are the goal of most research efforts into QTL and gene diagnostic tests. Advantages of such tests include the requirement for only a single assay, high sensitivity and specificity across populations, greater likelihood of securing IP protection and subsequent return on investment through licensing, and greater potential to target journals of higher impact for publication of results. Haplotype based tests may lack all of these advantages: assay of multiple markers is required, results cannot be generalized to populations unrelated to the discovery and validation populations, sensitivity and specificity depend on the distribution of alleles, IP protection is more difficult and publication options more restricted. Despite these real limitations, haplotype based tests have some potential advantages: lower density assays are all that are required for the discovery population, fewer assumptions regarding the number of alleles at the mutation are required, uncertainty can be explicitly acknowledged in the test results, and tests can be delivered to market in a short timeframe. Importantly, the time to market is not necessarily dependent on the complexity of the causal mutation, which is unknown at the commencement of the research.

In this paper, as a case study we use the development and commercialisation of a haplotype based test for polled in Australian beef cattle (Henshall, Piper and Tier, (2014); Piper, Tier and Henshall, (2014)). For both the case study and more general situations we discuss the investment required to develop the test, the underlying assumptions required by the test, the statistical methods used in the test, and the ongoing investment required to maintain the test.

Materials and Methods

By 2005 the polled locus had been mapped to a 1Mb interval on Bovine chromosome 1 (Georges, Drinkwater, King et al. 1993; Brenneman, Davis, Sanders et al. 1996; Drogemuller, Wohlke, Momke et al. 2005). Recently a commercial test was released based on a 202 bp insertion-deletion event (Medugorac, Seichter, Graf et al. (2012)). The development of the haplotype test for polled in Australian beef cattle pre-dated this test and is described in Henshall, Piper and Tier (2014) and Piper, Tier and Henshall (2014). The test is based on 10 microsatellite markers on BTA1, located between 1,495,504 bp and 2,119,315 bp (Bta4.0). These markers were all discovered by the end of 2008 (Mariasegaram, Harrison, Bolton et. al. (2012)). As at late 2013, for the least polymorphic marker 7 alleles had been observed in Australian populations, and for the most polymorphic marker 36 alleles. These are from an animal resource of 1,759 cattle from 16 beef breeds (including British, European and Zebu derived breeds) and Holstein Friesian. The animals came from two sources; for around half, DNA had already been extracted for an earlier use, and the other half were being tested commercially for the CSAFG29 single marker test (Mariasegaram, Harrison, Bolton et. al. (2012)). Most, but not all animals had phenotypes (polled, scurred or horned), and these were of variable reliability. The samples submitted for commercial testing with CSAFG29 were almost all from polled animals, while the other samples were chosen to span the full range of phenotypes. Where available, samples from sires that had progeny-tested poll genotypes were included. Samples of unknown phenotype were only included where required to help resolve haplotypes. No animals were phenotyped specifically as part of the trial; labor costs were incurred only in genotyping, DNA extraction for around half of the samples, and in analysis.

Haplotypes were estimated using the haplo.em function from the haplo.stats (Sinnwell and Schaid (2013)) package in R (R Core Team (2013)). The haplotype test requires assumptions about the expression of phenotype based on genotype. We assume alleles at the polled locus with two distinct effects described by the upper case words POLLED and HORNED. Diploid genotypes are described by pairs of upper case letters (PP, PH and HH). POLLED and HORNED were linked to phenotypes through a penetrance function (Table 1). Each row sums to one, and contains the probability of diploid polled genotype (in columns) given observed phenotype (in rows). Penetrance

values for progeny tested animals are given more weight than for animals with only a phenotype, but even then we do not make any probability equal to zero, as we acknowledge the possibilities of phenotyping errors, of genotyping errors, and of sample mislabeling errors. Strictly speaking, a penetrance function is the proportion of individuals in each phenotype class given the genotype, but as we have progeny tested individuals and known genotypes we express it here as the proportion of animals in each genotype class given the phenotype.

Table 1. Penetrance function relating diploid polled genotype to phenotype.

Genotype Phenotype	PP	PH	HH
Horned	0.05	0.10	0.85
Polled	0.49	0.49	0.02
Scurred	0.20	0.79	0.01
Progeny test PP (and Angus)	0.97	0.02	0.01
Progeny test PH	0.03	0.94	0.03
Progeny test HH	0.01	0.02	0.97

It is necessary to estimate the linkage with POLLED or HORNED for hundreds of unique haplotypes, and to do this we used an MCMC sampler applying the Metropolis-Hastings algorithm (Hastings 1970)). The sampler is described in detail in Henshall, Piper and Tier (2014). Probabilities of POLLED were estimated for each haplotype, and these estimates were used when the test was launched commercially. Users of the test are encouraged to submit a phenotype along with the tissue sample, so commercial samples contribute to refining the test as more data accumulate. Ideally, the sampler would be run every time a new batch of samples with associated phenotypes is received.

Results

In the 1,759 animal sample set, 448 distinct haplotypes were observed. The distribution of haplotype frequencies is displayed in Figure 1. In panel A it can be seen that while the most frequent haplotype was seen over 300 times, over 200 haplotypes were seen only once. In panel B it can be seen that the first 100 haplotypes account for around 80% of observations. The 200 haplotypes that were seen only once account for less than 10% of observations.

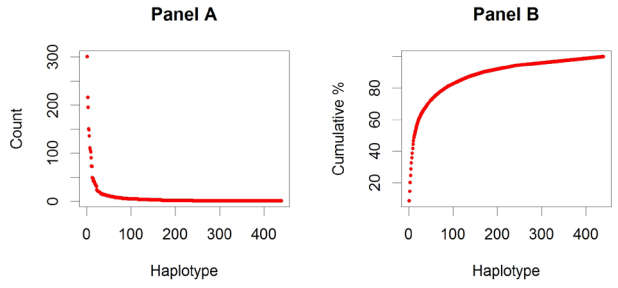

Figure 1: Frequencies of observed haplotypes in the 1759 animal calibration population.

As illustrated in Figure 2 (panel A), for haplotypes that were seen often in phenotyped animals there is little ambiguity in the genotype at the polled locus. Around 250 alleles have very low probabilities of being POLLED, around 65 alleles have very high probabilities of being POLLED, and around 150 alleles have an intermediate probability, but most of these were seen fewer than 3 times.

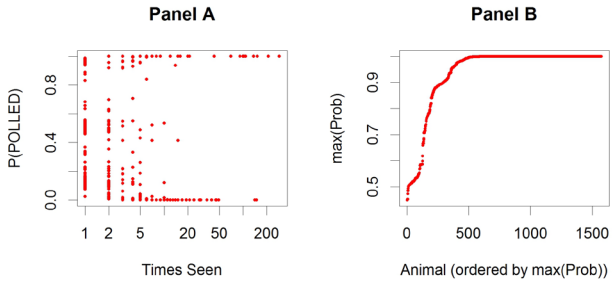

Figure 2: Estimated probability that haplotype alleles are POLLED, in panel A as a function of the number of times the haplotype was seen, and in panel B the probability of the most likely diploid genotype for each animal (max(Prob)).

The impact of the distribution of estimated haplotype effects on diploid genotype probabilities is shown in Figure 2 panel B, where the probability of the most likely diploid genotype is plotted for all animals with a phenotype. Out of 1,573 animals with phenotypes, for 295 animals the estimate has a probability of less than 90%, for 358 animals the estimate has a probability of less than 95% and for 469 animals the estimate has a probability of less than 99%.

Two months after the analyses reported above, 90 genotype samples were submitted for a Sanga derived breed not in the original sample set. There were 44 new haplotypes, with only 7 animals (6 polled, one horned) carrying no new haplotypes. Only 3 (out of 41) of the polled animals carried two new haplotypes, while 32 (out of 49) horned animals carried two new haplotypes. The samples were from a single breeder, with the genotyping of samples from the horned animals subsidized.

Discussion

Investment required to develop the test. In developing the test it became clear that it really didn't matter which markers were chosen from the region around the polled locus, any polymorphic markers would do. That means that had we had the statistical method and software we could have released this test much earlier than we did. For QTL discovered using high density SNP assays it might be that data for a haplotype based test are already available from the discovery population, allowing validation to take place using a smaller general purpose or custom assay. Discovery and validation population sizes are larger than required for causal mutation tests, but phenotypes need not be known with as high precision and assays during development are potentially much cheaper per animal.

Underlying assumptions required by the test. We assumed alleles having two distinct effects at the polled locus but we did not assume only two alleles. If there are more than two alleles then each will have been allocated to be either POLLED or HORNED, and having multiple alleles grouped together is not a problem provided that their effects are similar. Observations of some haplotypes behaving in an inconsistent manner could be evidence of either incomplete linkage, or an additional allele at polled with a different effect. If an additional allele was the cause then the penetrance function could be easily extended to accommodate it.

Statistical methods used in the test. Given the difficulty of estimating individual effects for multiple haplotypes, especially when effects are conditional on the other allele carried, sampling linkage to an unknown locus with a small number of effects provides a tractable solution. This is the case not just for discrete effects such as polled, but for loci affecting quantitative traits as well. Estimates are less precise for rare haplotypes, but this is entirely appropriate. We assume that users will accept a proportion of uninformative tests and a degree of uncertainty in others. This is less of an issue for breeders or buyers who are using or buying many tested animals, provided that the estimated genotype probabilities turn out to be unbiased in the long run. The region spanned by our test was determined by marker availability, and appears to be adequate at this stage. A wider haplotype is likely to have more alleles with less data per allele, making estimation more difficult. With a narrower haplotype fewer alleles means easier estimation, but more risk that alleles that are IBS are not IBD, and more risk that all important genomic features at the causal mutation are not spanned.

Ongoing investment required to maintain the test. Ongoing refinement of the test is an integral part of the design of the software and marketing. A central database is required, but this is no different to the requirements of other genetic improvement systems. The system also depends on the continual goodwill of users of the test. This may even be an advantage of the approach; engaging breeders in something of immediate importance to them may lead to increased awareness and adoption of other centralized genetic improvement services. Sometimes it may be necessary to subsidize the assay of samples from horned animals as these will not usually be submitted for commercial testing. In the case of the addition of the Sanga derived breed it would be fairly safe to assume that any new haplotype was most likely HORNED, given the breed history and the observation that almost all polled animals carried a haplotype we had seen before. This information though is not used by the sampler as it currently operates, and a much better approach is to source samples from horned animals from the new population, which the breeder was happy to provide.

Conclusion

Unlike in human health, where the individual is of primary importance, in livestock breeding the intent is to shift the population mean. As such a less accurate test brought to market earlier or more widely applied, may have greater impact than a more accurate test that takes longer to get to market or that is more costly and hence less widely applied. The mechanism by which the causal mutation affects the phenotype is interesting but not essential for genetic improvement.

Acknowledgments

The development of the test for polled was funded by Meat & Livestock Australia, and has benefited from data from numerous breeders and breed societies.

Literature Cited

Brenneman, R. A., Davis, S. K., Sanders, J. O. et al. (1996). *J. Hered*. 87:156–161.

Drogemuller, C., Wohlke, A., Momke, S., et al. (2005). *Mamm. Genome* 16: 613–620.

Georges, M., Drinkwater, R., King, T. et al. (1993). *Nat. Genet.*, 4:206–210.

Hastings, W. K. (1970). *Biometrika,* 57(1):97–109.

Henshall, J. M., Piper, E., and Tier, B. (2014). (in prep.).

Mariasegaram, M., Harrison, B. E., Bolton, J. A. et al. (2012). *Animal Genetics* 43:683-688.

Medugorac, I., Seichter, D., Graf, A., et al. (2012) *PLoS ONE* 7(6):e39477.

Piper, E. K., Tier, B. and Henshall, J. M. (2014). These proceedings.

R Core Team. (2013). R: A Language and Environment for Statistical Computing.

Sinnwell, J. P. and Schaid, D. J. (2013) haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. R package version 1.6.3. edition; 2013.