Haplotype Frequency Estimation From Pooled DNA For Between Family Selection In Aquaculture

J.M. Henshall[*]

Introduction

The potential to significantly reduce the cost of genotyping has driven research into estimating SNP and haplotype frequencies from samples of pooled DNA (Barratt et al. 2002; Craig et al. 2009; Homer et al. 2008; Kirkpatrick et al. 2007; Pirinen et al. 2008; Sham et al. 2002; Zhang et al. 2008). Barratt at al. (2002) and Sham et al. (2002) suggested a method to estimate haplotype allele frequencies from individual SNP allele frequency estimates from pooled DNA when there is prior knowledge of the set of segregating haplotype alleles. The method was developed further by Gasbarra et al. (2009). Gasbarra et al. (2009) restricted their discussion to the situation where there are more haplotype alleles than there are SNP in the haplotype. For human genetics this appears to be reasonable as the number of haplotype alleles for a DNA interval is proportional to 4 times the effective population size (Ne) (Sved 1971), and Ne is estimated to be 3,000 to 7,500 for human populations (Tenesa et al. 2007). However, in livestock populations, within breed Ne may be considerably less than this: estimates for cattle are in the vicinity of 100 (Gibbs et al.), and in an aquaculture trial or commercial cohort all animals may be progeny of a small number of mating pairs. Under these circumstances there may be more SNP in a haplotype than haplotype alleles segregating. In this paper we explore the implications of having fewer haplotype alleles than the number of SNP in the haplotype, with a particular focus on an example from aquaculture.

Material and methods

Suitable populations. Although it is unlikely that all haplotype alleles segregating in an ungenotyped population can be known with certainty, if all individuals are the progeny of known parents and the phased genotype of all of the parents is known, then, with the exception of recombinant haplotypes, the haplotype alleles existent in the progeny are known. Examples of when this might be the case are individuals from a cage of salmon or a pond of shrimp, comprising only 3 or 4 fullsib families. Where individuals are from a small number of families the number of haplotype alleles is not related to Ne but to the number of parents, and the SNP density does not need to be particularly high for there to be more SNP in the haplotype than there are haplotype alleles in the population. The density of existing SNP for Atlantic salmon (Salmo salar) or Pacific White Shrimp (Litopenaes vannamei) may be sufficient (Ciobanu et al. 2009; Kent et al. 2009).

[*] CSIRO Livestock Industries and Food Futures Flagship, FD McMaster Laboratory Chiswick, Armidale, 2350, NSW, Australia

Pooled DNA. It is assumed that quantitative SNP allele frequency estimates have been obtained for diallelic SNP in the haplotype, from a pooled sample of DNA for a population of interest. It is assumed that there may be error associated with these estimates.

Least squares estimation. As in Gasbarra et. al. (2009), we assume that each locus is diallelic, code the alleles 0 and 1, and form an n x p matrix where n is the number of SNP in the haplotype and p is the number of haplotype alleles. The dependent variable is the n x 1 vector of quantitative SNP allele frequency estimates. Like Gasbarra et. al. (2009) we treat these as continuous variables rather than as ratios of integers. The parameters of interest are in the p x 1 vector of haplotype frequencies. Unlike Gasbarra et. al. (2009), and as suggested by Barratt et al. (2002) and Sham et al. (2002) we require that $n \geq p$, and a constrained least squares approach is suitable. The model is illustrated for an example in Table 1, with $n = 11$ and $p = 6$. To simulate quantitative SNP allele frequency estimates we added a random error term (standard deviation 0.01) to the frequency expected given the true haplotype frequencies. Haplotype allele frequencies were then estimated from the perturbed SNP allele frequencies using the "Solver" tool in Microsoft Excel to minimize the squared error subject to the constraints that estimated frequencies were ≥ 0.0 and sum to 1.0.

Results and discussion

For 10 random replicates of the example in Table 1, each with different haplotypes and frequencies, the mean of standard deviation of the difference between the true and estimated haplotype frequencies was 0.006, with a maximum of 0.010. The estimated haplotype allele frequencies can then be used to provide more accurate estimates of the frequencies of the SNP alleles that comprise the haplotype if these are of interest.

Table 1: Least squares model for an analysis with 6 haplotype alleles (A to F), each of 11 SNP. The columns of haplotypes form the design matrix, the dependent vector y contains the vector of frequency estimates for the SNP allele coded "1".

	Haplotypes						
	A	B	C	D	E	F	**y**
	0	0	1	1	1	1	0.542
	0	0	1	1	1	0	0.483
	1	0	0	0	0	0	0.040
	0	1	1	0	1	1	0.799
	1	0	1	0	1	1	0.415
	0	1	1	1	0	0	0.573
	1	1	1	1	1	0	0.969
	0	1	0	1	1	1	0.966
	0	0	1	1	0	0	0.166
	1	0	1	1	0	1	0.257
	1	0	1	0	0	0	0.050
True frequency	0.047	0.422	0.002	0.155	0.323	0.052	
Estimated frequency	0.043	0.419	0.000	0.162	0.330	0.046	

The main reason for pooling DNA samples is to reduce the cost of genotyping, and in an aquaculture setting the savings are potentially significant. For example, in a typical Atlantic Salmon enterprise broodstock spend their lives in a freshwater hatchery, and are not challenged in the marine growout environment of their progeny. DNA parentage of marine cohorts can be used to estimate breeding values on parents, allowing between family selection. If DNA from phenotypic extremes of marine cohorts is pooled and haplotype frequencies estimated as described above, then by assaying only a small number of pooled samples, thousands or even tens of thousands of fish can still contribute to the estimation of between family breeding values. The assay of hundreds or thousands of SNP for the pooled sample may be more expensive than the assay of a few dozen microsatellites as occurs today, but many fewer assays are required. In the example in Table 2, the cost of the SNP assay is 5 times the cost of the microsatellite assay, and pooling of blood as in Craig et al (2009) is assumed, and the total cost of the program is less when SNP are used. However, less information can be extracted from pooled samples than from individual samples, and whether the two approaches produce estimates of family effects of equal accuracy will depend on the genetic parameters of the trait of interest and on the number and population structure of the progeny tested. A requirement is that the haplotypes in the parents are phased, which is a reasonable assumption for an ongoing program, as the grandparents will have been genotyped as part of the selection program a generation before.

Table 2: Example of the costs of a SNP based program and a microsatellite based program.

	Individual microsatellites		Pooled SNP	
	Number	Cost (at $20 per assay)	Number	Cost (at $100 per assay)
Parents Tested	144	$2,880	144	$14,400
Progeny Tested	1,000	$20,000	1,000+	
Pools Tested	NA		50	$5,000
Total		$22,880		$19,400

For the example in Table 2 the SNP are not used for genomic selection (Meuwissen et al. 2001), but essentially to estimate parent breeding values from phenotyped progeny, thus allowing between family selection. However, the SNP haplotypes do contain additional information that could be used to make within family selection decisions, by selecting broodstock that carry haplotypes associated with desirable trait values recorded in their marine tested fullsibs. This would add to the cost of the program as potential parents have to be assayed as well as fish that are actually selected and used as parents.

The example above is applicable in aquaculture, where pools of siblings are common. However the method of estimating haplotype allele frequencies is equally applicable to livestock, where phenotypes and DNA pooling would be for relatively unrelated individuals. A density of at least (4Ne + 1/c) SNP per Morgan would be required, as this is when, for an interval with recombination frequency c, the number of SNP in the interval is approximately equal to the expected number of haplotype alleles in the population. A variation on genomic selection (Meuwissen et al. 2001) would be required, in which an estimated breeding value would be based on an animal's relationship to phenotyped and genotyped pools of animals, with the relationships estimated using haplotypes.

Conclusion

SNP densities are already sufficient to make haplotype allele frequency estimation from pooled DNA feasible for some aquaculture species, and ongoing SNP discovery suggests that the potential for the method will grow. If used for pedigree reconstruction the SNP based method is potentially no more expensive than existing microsatellite based methods, but with SNP haplotypes there is also scope to implement more sophisticated selection programs. The method will also be feasible in some livestock breeds once the next generation of dense SNP panels becomes available.

References

Barratt, B. J., Payne, F., Rance, H. E. *et al.* (2002). *Ann. Hum. Genet.*, 66:393-405.

Ciobanu, D. C., Bastiaansen, J. W. M., Magrin, J. *et al*. (2009). *Animal Genetics*, 9999.

Craig, J. E., Hewitt, A. W., McMellon, A. E. *et al.* 2009. *Genome Res*., 19:2075-2080.

Gasbarra, D., Kulathinal, S., Pirinen, M. *et al.* (2009). *IEEE Transactions on Computational Biology and Bioinformatics,* In Press.

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P. *et al.* (2009). *Science*, 324:528-532.

Homer, N., Tembe, W. D., Szelinger, S. *et al.* (2008). *Bioinformatics*, 24:1896-1902.

Kent, M. P., Hayes, B., Xiang, Q. *et al*. (2009). *Plant and Animal Genomes XVII.*

Kirkpatrick, B., Armendariz, C. S., Karp, R. M. *et al*. (2007). *Bioinformatics*, 23:3048-3055.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). *Genetics*, 157:1819-1829.

Pirinen, M., Kulathinal, S., Gasbarra, D. et al. (2008). *Genetics Research*, 90:509-524.

Sham, P., Bader, J. S., Craig, I. *et al*. (2002). *Nature Reviews Genetics*, 3:862-871.

Sved, J. A. (1971). *Theoretical Population Biology*, 2:125-141.

Tenesa, A., Navarro, P., Hayes, B. J. *et al*. (2007). *Genome Research*, 17:520-526.

Zhang, H., Yang, H. C. & Yang, Y. N. (2008). *Bioinformatics*, 24:1942-1948.