

LIMITS TO GENOTYPIC PROBABILITIES FOR SINGLE LOCI

Bruce Tier* and John M. Henshall⁺

*Animal Genetics and Breeding Unit¹, University of New England, Armidale NSW 2351

⁺CSIRO Livestock Industries, Locked Bag 1, Armidale NSW 2350 Australia

SUMMARY

Methods for determining genotype probabilities were originally developed for use with single bi-allelic loci in small populations where one allele is at low frequency. Now they are being applied to larger populations with multi-allelic loci. Using livestock populations, this paper examines the proportions of individuals whose genotype cannot be inferred with certainty when the genotypes of its parents, progeny and mates are known. It is concluded that when allele frequencies are moderate the genotypes of most individuals cannot be inferred with certainty.

INTRODUCTION

Genotypic information is now becoming more available for some loci. Nevertheless, it is still expensive to genotype individuals and, while it remains so, many animals will not be genotyped for all loci. When the pedigree of a population is recorded accurately it can be used to extend genotypic information to other, un-genotyped individuals. For example the genotypes of progeny of two homozygous parents can be inferred with certainty. Similarly progeny exhibiting two different alleles can be used to infer the genotype of a common parent. This can result from two different homozygotes among the progeny (e.g. progeny with AA and BB genotypes), or because the parental allele can be ascertained as the one not coming from the mate (e.g. two progeny have the same genotype AB, but one mate has the genotype AA and the other BB). However, it is impossible to infer the genotypes of other individuals, such as the progeny of heterozygous individuals, with certainty. For these, and other un-genotyped individuals there are methods for assigning probabilities for all possible genotypes. These methods include peeling (Elston and Stewart, 1971) and sampling descent graphs (Henshall et al. 2001). Certainty can extend a number of generations if the progeny of the genotyped individuals are themselves homozygotes.

Practitioners may want to know the genotypes of all, or only a subset of the population – such as all progeny born this year or all potential parents. Because genotypes of some individuals can be inferred it may be unnecessary to genotype all the individuals in the subset. Kinghorn (1999) and Macrossan (2004) have both examined the question of how to maximise the amount of information about the genotypes of the desired subset for a given amount of genotyping.

Elston and Stewart's (1971) method for computing genotype probabilities was developed for use in human populations where family sizes and populations considered are small in number, and the locus of interest has a deleterious allele at low frequency. The purpose of such analyses is to assign probabilities to carriers. Van Arendonk et al. (1989) extended this idea to use in large livestock populations. Currently genotypes can take many forms. Microsatellite markers are chosen because

¹ AGBU is a joint venture of the NSW Department of Primary Industries and the University of New England.

they have large numbers of alleles, in moderate frequencies, and can be used to determine parentage and/or to trace inheritance of chromosomes from one generation to the next. Single nucleotide polymorphisms (SNPs) take only two forms, however multiple linked SNPs can also be used to emulate microsatellites. SNPs exist in coding and non-coding regions. Direct markers are generally a short sequence of nucleotides used to identify alternative alleles.

Using major livestock species this paper examines the proportion of population that cannot be inferred from the genotypes of animals adjacent to it in the pedigree (ie. parents, progeny and mates).

MATERIALS AND METHODS

Real pedigrees for cattle, sheep and pig populations were obtained from BREEDPLAN, LAMBPLAN and a PIGBLUP user respectively. Base alleles were assigned to founders according to the gene frequency in the base population (see Table 1). Alleles were dropped through the pedigree according to Mendel’s first law. If possible, the genotype of each individual was determined from the genotypes of the adjacent set of animals – parents, progeny and mates.

Table 1: Number of alleles and their relative proportions in the founder populations.

Case Label	Number of Alleles	Relative Proportions
1	2	1:1
2	2	1:2
4	2	1:4
8	2	1:8
16	2	1:16
32	2	1:32
64	2	1:64
128	2	1:128
41	4	1:1:1:1
42	4	1:2:3:4
43	4	1:2:4:8
81	8	1:1:1:1:1:1:1:1
82	8	1:2:3:4:5:6:7:8
83	8	1:2:4:8:16:32:64:128

Table 2: Structure of sample populations.

Population	Beef cattle	Meat sheep	Pig
Number (percentage) of			
Individuals	9,648	5,979	20,408
Full-sib families	6,865	3,651	2,492
Founders	1,449(15.0)	746(12.5)	174(0.85)
Sires	1,081(11.2)	484(8.1)	168(0.82)
Dams	3,608(37.4)	1,997(33.4)	752(3.7)
Parents in last year	433(4.5)	467(7.8)	198(0.97)
Progeny in last year	394(4.1)	696(11.6)	2,545(12.5)
Family sizes:			
Full-sib	1.2	1.4	8.1
Paternal half-sib	7.6	10.8	120.4
Maternal half-sib	2.3	2.6	26.9
Mean generations of pedigree in last year’s drop	4.4	6.7	5.0

The effect of different numbers of alleles with different frequencies in the base population was examined and the list is shown in Table 1. The genotypes for each case and species were replicated 100 times. The mean proportions of the genotypes that could not be inferred, given each individual's adjacent set of genotypes was calculated for four subsets of the population. The subsets were: the whole population, the founders, the parents of progeny born in the most recent year, and progeny born in the most recent year.

Table 3: Percentages of subsets of three livestock populations (BC: Beef Cattle, MS: Meat Sheep, P: Pigs) whose genotypes could not be inferred with certainty if their parents, progeny and mates genotypes were known. Mean of 100 replicates. Shading indicates proportion of individuals that cannot be inferred (no shading 0-10%, light grey 10-25%, grey 25-50%, dark grey 50-100%). Superscripts show the range of standard errors for each estimate (unmarked: 0-0.1, a: 0.1-1%, b: 1-2.5%, c: 2.5<5%, d: 5-10%, e: 10-15% and f:15-20%).

Case	Total Population			Founders			Most recent parents			Most recent progeny		
	BC	MS	P	BC	MS	P	BC	MS	P	BC	MS	P
1	71 ^b	71 ^c	72 ^c	97 ^a	99 ^a	58 ^b	51 ^c	51 ^d	5.7 ^b	75 ^d	75 ^d	75 ^d
2	66 ^c	66 ^d	66 ^d	97 ^a	99 ^a	58 ^b	47 ^c	47 ^d	4.7 ^b	68 ^d	70 ^e	69 ^e
4	55 ^c	53 ^d	51 ^e	97 ^a	99 ^a	60 ^b	36 ^c	34 ^d	3.9 ^b	54 ^d	51 ^f	53 ^f
8	42 ^c	39 ^d	35 ^e	98 ^a	99 ^a	61 ^b	26 ^c	22 ^d	2.5 ^b	36 ^d	32 ^f	35 ^f
16	30 ^c	27 ^d	20 ^e	98 ^a	99.5	62 ^a	16 ^c	12 ^d	1.4 ^b	19 ^d	18 ^e	21 ^f
32	24 ^c	21 ^d	8.7 ^d	99 ^a	99.5	63 ^a	11 ^c	6.8 ^d	0.6 ^b	12 ^d	11 ^e	8.3 ^e
64	20 ^b	16 ^c	6.3 ^d	99 ^a	99.6	63 ^a	7.3 ^c	2.8 ^c	0.4 ^a	5.7 ^d	4.5 ^d	7.0 ^e
128	17 ^b	15 ^c	3.3 ^c	99 ^a	99.6	64 ^a	5.9 ^b	1.6 ^c	0.2 ^a	3.1 ^c	3.1 ^d	2.9 ^d
41	82 ^a	82 ^b	90 ^b	94 ^a	99 ^a	53 ^b	49 ^b	52 ^b	5.3 ^b	94 ^c	93 ^c	93 ^d
42	80 ^b	81 ^b	87 ^c	94 ^a	99 ^a	54 ^b	50 ^c	52 ^c	5.3 ^b	91 ^c	90 ^d	90 ^d
43	77 ^b	77 ^c	82 ^c	95 ^a	99 ^a	55 ^b	50 ^c	50 ^c	5.3 ^b	86 ^c	85 ^d	85 ^d
81	83 ^a	85 ^a	94 ^a	93 ^a	98 ^a	51 ^b	44 ^b	48 ^b	4.6 ^b	98 ^c	98 ^b	98 ^b
82	83 ^a	84 ^a	94 ^a	93 ^a	98 ^a	51 ^b	46 ^b	49 ^b	4.8 ^a	97 ^b	97 ^c	98 ^b
83	78 ^b	79 ^c	85 ^c	95 ^a	99 ^a	55 ^b	49 ^c	51 ^c	5.2 ^b	88 ^c	87 ^d	87 ^d

RESULTS

The population structure for each species is summarised in Table 2. Full-sib and half-sib families were larger and there were correspondingly fewer founders and parents in the pig population than in the sheep and cattle populations. The mean number of generations in the pedigrees for the set of most recent progeny ranged between 4.4 (cattle) and 6.7 (sheep).

The proportions of individuals whose genotypes could not be inferred are shown in Table 3. There is an inverse relationship between the proportion of the individuals that can be genotyped and the numbers of alleles at the locus. Increasing differences among the founder allele frequencies increases the proportion of the population whose genotypes can be determined with certainty. The genotypes of

many founders cannot be determined. Similarly the genotypes of many of the most recently born individuals and their parents cannot be determined unless there are few alleles at unequal frequency.

DISCUSSION

In most cases it will be necessary to genotype individuals if their genotype is required, as only when there are few alleles in unequal proportions is it possible to determine genotypes with certainty. The results for two alleles with extremely different frequencies illustrate the case for which methods for determining genotype probabilities were developed. In these cases the genotypes of the most recent individuals and their parents could be determined with certainty. Few individuals were carriers. When there were larger numbers of alleles and they were at similar frequency in the base population the genotypes of only a very small proportion could be inferred.

The genotypes of most of the founders of the sheep and cattle populations could not be determined from their offspring and mates. This reflects the few progeny from the founder females. Although the genotypes of many more founders can be determined in the pig population, the majority of them cannot be. In the most recent generation the proportions that cannot be determined (U) is proportional to the proportion of matings between homozygous parents and is approximated with the formula $U = \left(\sum_i p_i^2 \right)^2$, where p_i is the frequency of the i th allele.

If there are errors in the pedigree then the situation will be worse than depicted in Table 3. However, the use of DNA testing to verify the pedigree can increase confidence when inferring genotypes. Multiple linked loci present an entirely different type of problem. Depending upon the proximity of the loci to each other, genotypes can be inferred for many un-genotyped loci given partial information on other loci on all adjacent individuals.

REFERENCES

- Elston R.C., Stewart J. (1971) *Hum. Hered.* **21**:523.
Henshall J.M., Tier B., Kerr R.J. (2001) *Genet. Res.* **78**:281.
Kinghorn B.P. (1999) *J. Anim. Breed. Genet.* **116**:175.
Macrossan P. (2004) Ph.D. Thesis, University of New England.
van Arendonk J.A.M., Smith C., Kennedy B.W. (1989) *Genet.* **78**:735.