Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 122-125

MARKER-BASED FAMILY ASSIGNMENT WITHOUT PARENT GENOTYPES

K.G. Dodds[1], J.E. Symonds[2], S.J. Rowe[1] and J.C. McEwan[1]

[1] AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand
[2] Cawthron Institute, Nelson, New Zealand

SUMMARY

We investigate how to assign individuals to families (and therefore, implicitly, to parents) when parent genotypes are not available. In the first case, a set of progeny were assigned to putative full-sib families using an algorithm based on clustering and expected relatedness values. A plausible set of full-sib families was found, but the method is subject to some arbitrary thresholds. In the second case, individuals were putatively assigned to their (paternal) grandparents based on relatedness or the frequency of opposing homozygotes in the pair and these were compared to the true pedigree. The assignments were good when they were restricted to the set of paternal grandparent mate pairs and/or when dam genotypes were available to infer sire alleles, which could then be matched to the paternal grandparents using methods designed for low-depth sequencing-based genotypes.

INTRODUCTION

Sometimes breeders or researchers want to infer family relationships using genetic marker information, but do not have genotypes or DNA samples from the putative parents. Two such use cases are to 1) assign a set of progeny to full- or half-sib families, or 2) assign individuals to their grandparents. In the absence of parents, exclusion methods, such as relying on opposing homozygotes, are not available to unequivocally exclude some potential relationships. Modern technologies like SNP chips or sequencing-based assays allow inference based on relatedness estimates (Moore *et al*. 2019) and/or on the distribution of genotype combinations in pairs of individuals (VanRaden *et al*. 2013). These approaches are applied to two real data sets to assign individuals to families using genomic data.

MATERIALS AND METHODS

Example 1: full-sib families. For this research, a set of 5,216 Chinook salmon were obtained from a breeding company for undertaking a research project (Symonds *et al*. 2025). The company advised that this population comprised 133 full-sib families, however, the family membership was unknown, and the parental DNA was not available. The progeny were genotyped using genotyping-by-sequencing (GBS) and QC filters applied as described in Scholtens *et al*. (2023). A genomic relationship matrix (GRM), **G**, was calculated following the method of Dodds *et al*. (2015) and then rescaled by dividing each row and each column by the square root of the diagonal element. This was to allow focus on the relatedness derived from parents alone and not additional common ancestry (creating inbreeding).

Fish were assigned to full-sib families as follows:

1) k-means clustering was applied using a distance matrix calculated as the Euclidean distance between pairs of rows in **G**, with 120 clusters and using the *kmeans* function in base R software with 50 random starting sets (*nstart*=50 in the *kmeans* function) but default settings otherwise.

2) split families further cutting hierarchical cluster dendrogram (on sqrt(**G**)-min(**G**)) of each family at 1.6

3) split families with mean relatedness < 0.2 into singleton families

4) join families related at 0.4

5) split families with evidence of two groups, based on the top join of a hierarchical cluster and where the resulting two groups have mean relatedness < 0.35, or a mean squares ratio (original group

size -2) times squared difference between groups divided by the sum of squared differences within groups > 4.

Example 2: finding grandparents. A set of 263 sheep (the 'progeny') from an experimental flock (Rowe *et al*. 2019), born in 2023, along with their parents and grandparents were genotyped using various SNP chips containing at least 16,000 SNPs. The 12,682 autosomal SNPs that were common across all these SNP chips, had less than 30% missingness and were not monomorphic were used to construct a GRM using the imputation-free method described by Dodds *et al*. (2015). The 2023-born cohort and their parents previously had their parents assigned based on their SNP chip genotypes. Only those progeny whose parents were not also in the set of grandparents were included for this study. These progeny had 10 sires with unique parentage and 145 dams, which had 24 sires and 118 dams. In total, there were 31 grandsires and 126 granddams.

Paternal grandparents (PGPs) were assigned using different methods and levels of associated information ('scenarios', Table 1). In all cases all true PGPs were present, but scenarios without an 'a' in the label (Table 1) also included the maternal grandparents. Grandparents were putatively assigned to the progeny either as the highest related ('relatedness method') or the progeny-grandparent pairings with the lowest opposing homozygote rate ('EMM method'). EMM refers to the excess mismatch rate compared to the expected rate for the given level of sequencing depth for methods such as GBS; the expected mismatch rate with actual genotypes (e.g. chip-based) is zero. For the '1' scenarios (Table 1), only the grandparents were available. For scenarios labelled as '2' where the dams were present, the additional criteria of grandparent-dam relatedness or grandparent-dam-progeny mismatch rates were examined to help determine the PGPs. If the best match failed to meet the criteria, the 2nd best match was then considered.

Inferred sire alleles were used in the '3' scenarios (Table 1). These alleles are inferred from progeny and dam genotypes. For example, if the progeny is AB and the dam is AA the sire must have a B allele. The inferred alleles were then treated in the same way as a GBS genotype with depth of 1 or 0 corresponding to when an allele could be inferred or not, respectively, allowing a search for the sire's parents. Assignments were made using the parentage functionality within the KGD software (Dodds *et al*. 2019).

RESULTS AND DISCUSSION

Example 1: full-sib families. There were 148 families found (compared to 133 recorded families), ranging in size from 1 (11 families) to 51. However, 14 of the assigned families had five or fewer members while the remaining 134 families had at least 17 members. It is possible that a few fish from outside these families may have escaped into this group and that two of the assigned families are actually a single family.

The large family sizes in this study have likely assisted the clustering algorithm to define these groups effectively. The likely presence of relatedness between families (such as different families being half-sibs) has made the process of assigning families more difficult as the closest related half-sibs may have similar relatedness to the least related full-sibs, especially if there is some relatedness between the non-common parent for the half-sibs.

The method is not guaranteed to produce the same groupings if rerun (as the k-means method uses randomised starting sets). However, in this case, a repeat run of the method did produce the same results. Several arbitrary criteria have been used in this process; the initial number of families to use depends on the expected number of families, but it is not clear if the other thresholds will need to be adjusted for other situations. For the current example it appeared more effective to initially assign fewer than the expected number of families and then subdivide them if needed.

The proposed method could also be adapted (using different thresholds) for finding half-sib families. This is relevant in livestock populations when only the progeny generation is genotyped, and dams have only one or a few progeny each. Assigning with a target relatedness of 0.25 (for non-

inbred half-sibs) may present more challenges than the full-sib case, but it should still be possible for reasonably sized families.

Example 2: finding grandparents. A summary of the PGP assignment is shown in Table 1. In this study the best matches were always 'assigned', which is sensible if the true relatives are present in the data (as known in this case). In practice, assignments depend on the measures passing a certain threshold (e.g. for relatedness and/or excess mismatch rate). The KGD software reports only the top two matches; however, it might be possible to find suitable matches (Scenarios 2, 2a) in lower ranked matches when additional criteria (to exclude the maternal relatives) are used.

Table 1. Paternal grandparent assignment rates

Scenario*	Match method	Additional criteria†	PGS % correct	PGD % correct	PGS & PGD % correct	% unassigned
1	relatedness		80	44	30	0
	EMM		80	44	30	0
1m	relatedness		94	94	93	0
	EMM		94	93	93	0
1am	relatedness		100	100	100	0
	EMM		100	100	100	0
2	relatedness		41	44	11	0
		DamRel	83	85	59	26
		TrioMM	87	88	69	32
	EMM		41	54	17	0
		DamRel	73	90	58	22
		TrioMM	88	93	76	30
2a	relatedness		76	65	49	0
		DamRel	97	83	79	11
		TrioMM	100	89	89	14
	EMM		78	75	57	0
		DamRel	94	90	82	8
		TrioMM	98	95	92	11
3	relatedness		98	97	95	0
	EMM		97	98	95	0
3a	relatedness		100	99	99	0
	EMM		99	99	98	0
3m	relatedness		100	100	100	0
	EMM		100	100	100	0
3am	relatedness		100	100	100	0
	EMM		100	100	100	0

* Scenario labels with 'a' include only PGP in the test set, otherwise all GPs; scenario labels with 'm' use known grandparent mating pair information. In scenarios 1, 1m, assignments were considered correct if the PGP was among the best two matches (allows the maternal grandparent to be the best match). In scenarios 2(a), the dam genotypes were available, so the 2nd best female match is assigned as PGD (best match is the dam). In scenarios 3(a/m) the matching is to the inferred sire alleles.

† DamRel criterion is that an assigned paternal grandparent must have a relatedness less than 0.4 with the dam. TrioMM criterion is that the progeny, dam and proposed grandparent should have a parent-progeny mismatch rate less than 0.1.

PGS is paternal grandsire and PGD is paternal grandam.

Although haplotypes were inferred for the sires, no SNP positional information was subsequently used, i.e., the fact that the sire alleles were together in the same gamete was disregarded.

Additionally, there was no attempt to collate sire haplotypes from different progeny, as at that stage it was not known which progeny have the same sire. It might be possible to find paternal half-sibships through clustering methods and then collate the sire haplotypes. The corresponding 'read depths' would be the number of offspring with inferred sire alleles for that SNP. The use of haplotypes (VanRaden *et al*. 2013) and/or the number of shared identity-by-descent segments (Jewett *et al*. 2021) could provide a more powerful assignment tool, as might be required in more extensive searches.

The approaches to making assignments (relatedness or EMM) performed similarly with neither being uniformly better. In practice combining these approaches is likely lead to better results. Moreover, it is also possible that half-sib clustering could aid the assignments, however this approach was not investigated.

CONCLUSIONS

Using genetic data to assign family relationships is simpler for parent-progeny than for sibships or more distant relationships. The assignment process can be helped with the use of ancillary information such as known mating pairs or expected number of family groups. The proposed method for assigning full-sib families placed most of the fish into almost the expected number of groups, but required a set of arbitrary thresholds. Finding paternal grandparents is difficult when dams are also related as in the sheep dataset. If the dams are genotyped (and assigned), a useful approach is to infer the sire haplotype and then apply a parentage assignment to that haplotype, using methods designed for low-depth sequence-based genotyping. The proposed methods are a useful addition to the tools available for parentage testing.

ACKNOWLEDGEMENTS

We thank the GenomNZ (AgResearch) laboratory for genotyping the animals in both examples. Sheep were funded by the New Zealand Agricultural Greenhouse Gas Centre (NZAGRC) and the Pastoral Greenhouse gas research consortium (PGgRc). The salmon research was supported by the NZ Government Ministry for Business Innovation and Employment (contract no. CAWX2304).

REFERENCES

Dodds K.G., McEwan J.C., Brauning R., Anderson R.A., Van Stijn T.C., Kristjánsson T. and Clarke S.M. (2015) *BMC Genomics* **16**: 1047.

Dodds K.G., McEwan J.C., Brauning R., Van Stijn T.C., Rowe S.J., McEwan K.M. and Clarke S.M. (2019) *G3: Genes Genom. Genet.* **9**: 3239.

Jewett E.M., McManus K.F., Freyman W.A. and Auton A. (2021) *Am. J. Hum. Genet.* **108**:.2052.

Moore K.L., Vilela C., Kaseja K., Mrode R. and Coffey M. (2019) *J. Anim Sci.* **97**: 35.

Rowe S.J., Hickey S.M., Jonker A., Hess M.K., Janssen P.H., Johnson T., Bryson B., Knowler K., Pinares-Patiño C., Bain W., Elmes S., Young E., Wing J., Waller E., Pickering N. and McEwan J.C. (2019) *Proc. Assoc. Advmt Anim. Breed.Genet.***23**: 306.

Scholtens M., Dodds K., Walker S., Clarke S., Tate M., Slattery T., Preece M., Arratia L. and Symonds J. (2023) *Aquaculture* **563**: 738936.

Symonds J.E., Scholtens M.R., Dodds K.G., Costilla R., Clarke S.M., Arinez J., Kenny N., Walker S.P. and Waddington Z. (2025) *Proc*. *Assoc. Advmt Anim. Breed. Genet.* **26**: *These proceedings*.

VanRaden P.M., Cooper T.A., Wiggans G.R., O'Connell J.R. and Bacheller L.R. (2013) *J. Dairy Sci.* **96**:1874.