SNP Based Parentage Verification via Constraint Non-Linear Optimisation

Vinzent Boerner and Robert Banks
Animal Genetics and Breeding Unit (AGBU), Armidale, Australia
corresponding author: vboerner@une.edu.au

Abstract

Since the introduction of parentage verification by molecular markers this technique is based mainly on short tandem repeat markers (STR). With the advent of single nucleotide polymorphism (SNP), advances in genotyping technologies and decreasing costs, SNPs have become the marker of choice for genotyping projects. This is because the genotypes have a wide range of applications and imputation technologies provide well a developed compatibility layer between different types of SNP genotypes. Thus, the subsequent step is to use SNP genotypes for parentage verification as well. However, algorithms for parentage verification mostly date back to the STR era, and recent developments of SNP based algorithms such as evaluating opposing homozygosity have drawbacks, for example the inability of rejecting all animals of a sample of potential parents. This paper describes an algorithm for parentage verification via non-linear optimisation which overcomes the latter limitations and proofs to be very fast and highly accurate even with number of SNPs as low as 100. The algorithm was tested on a sample of 90 animals with 100, 500 and 40k SNP genotypes. These animals were evaluated against a pool of 12 putative parents containing random animals only, random animals and the true dam, and random animals, the true dam and the true sire. Assignment quality of the algorithm was evaluated by the power of assignment (Pa , probability of picking the true parent when true parent was among the putative parents) and the power of exclusion (Pe, probability of rejecting all parents if the true parent was not among the putative parents). When used with 40k genotypes, the algorithm assigned parentage correctly for all 90 test animals. That is, if one or both parents were among the putative parents they were correctly identified. If both were absent parentage was ruled out for the whole set of putative parents. A similar result was achieved when shrinking the genotypes to 500 randomly selected SNP, with Pe = 0.99 and Pa = 1. When only 100 SNP, randomly selected but the sample space narrowed by the minor allele frequency >0.3, were used, Pe and Pa were still 0.99 and 0.96, respectively. The described method is an easy to implement, fast and accurate algorithm to assign parentage using genomic marker data of size as low as 100 SNP. It overcomes limitation of methods such as evaluation of opposing homozygosity by not relying on the presence of a true parent in the pool of putative parents.

Introduction

The advent of DNA markers has facilitated verification of nominated parents enabling more accurate pedigrees for genetic evaluation, conflict resolution in breeding animal trading and basic parent identification in extensive production systems. For the last two decades this verification was based on short tandem repeat markers (STR), commonly called micro-satellites, which are highly polymorphic allowing to discriminate between individuals even if the total number of markers used is low. Due to their highly polymorphic character, parentage assignment on the basis of STRs can be done by simple exclusion or by categorical allocation. For an exhaustive review of parentage assignment algorithms see Jones *et al*. (2010).

However, in the last decade single nucleotide polymorphism (SNP) have been established as a new bi-allelic marker class. SNPs became quickly the markers of choice for genotyping projects because their sheer abundance made them much more suitable for genome wide association studies. In addition, imputation techniques provided a compatibility layer between different types of SNP genotypes relieving researchers from the necessity of re-genotyping ancient animals if SNP panels change. Eventually, SNP marker became the backbone of genomic selection which is now to replace pedigree selection as the dominant form of animal breeding (Mäntysaari *et al.*, 2014).

Animal breeding happens within an economic environment, and it was only

consequential and a matter of time that questions arose about the necessity of genotyping animals twice, STRs for parentage verification and SNPs for genomic selection. Since many ancient animals have STR genotypes only, a first step to merge both approaches was to impute STRs from SNP marker genotypes (McClure *et al.,* 2012). While this provides a necessary compatibility layer between STR parentage verification and SNP marker genotypes during a transition period, parentage verification should omit imputation and rely on SNPs only as soon as SNP genotypes for both, parents and offspring, are available. However, the bi-allelic nature of SNPs requires much more markers for successful parent identification. Although initial simulations found 100 SNPs to be sufficient (Baruch & Weller, 2008), more recent experience with real data suggest at least 500 SNPs for successful parentage assignment (McClure *et al.,* 2015). In addition, new algorithms had to be developed which could exploit information in SNP genotypes for that purpose. One method is the evaluation of the number of opposing homozygous marker loci as a possible measure of parentage (Wiggans *et al.,* 2009; Hayes, 2011). That is, parents are identified by having the least number of loci with a homozygosity status opposite to that of the offspring because opposing homozygosity between parents and offspring is theoretically impossible, but introduced by genotyping errors. Although necessary speed ups of the initial slow implementation of that method were developed (Ferdosi & Boerner, 2014), and it has been used in some studies already (Heaton *et al.,* 2014; Strucken et al., 2014), its main short coming remains: the true parents must be among the pool of suggested parents (Boichard *et al.,* 2014). Boichard et al. (2014) pointed out that likelihood based methods (Kalinowski *et al.,* 2007; Marshall *et al.,* 1998), originally developed to deal with STRs, can allow for the absence of the true parents but suffer speed limitation, and made necessary adjustments to that technique to make it suitable for SNP genotypes. However, they found that method to have difficulties finding the correct parent if the number of SNP markers approached 100 (Boichard *et al.,* 2014). This article describes a non-linear optimisation approach for parentage assignment, in the remainder called "constraint genomic regression" (CGR), which overcomes the limitations of opposing homozygosity evaluation. The algorithm is easy to implement, very fast, scales to any size of marker genotypes and provides both, a high power of exclusion (rejecting wrong parents) if true parents are not among the putative parents, and a high power of assignment (picking the true parent) if at least one parent is among the putative parents, even for SNP genotypes with as little as 100 SNP. The algorithm was tested on a data set of 4612 Australian Angus beef cattle SNP genotypes using 90 as animals with uncertain parentage.

Methods

Model

The problem to solve can be written as:

$$\arg\min_{b} f(b) = y'y - 2y'Xb + b'X'Xb$$

$$s.t.$$

$$b_i \geq 0 \ \{i = 1, ., N\}$$

$$\sum_{i}^{N} b_i = 1$$

where y is the marker genotype of the animal with uncertain parentage (explained animal) and X is a matrix of marker genotypes of possible parents (explanatory animals). Columns in X can be genotypes of single animals (e.g. sire, dam), or functions of genotypes of single animals or several animals (e.g. population allele frequencies). Values in vector b are regression coefficients regressing y on the columns in X. Minimising equation 1 with respect to equation 2 and 3 will yield a vector b of which values will not only explain the genotype in y as a linear function of genotypes in X, coefficients also have the straight forward interpretation what proportion of y is explained by each column in X. If X were containing genotypes of putative sires and dams only, values in b are not guaranteed to give reasonable results. Thus, it is highly advisable to always add the vector of population allele frequencies to X.

Data

The test data set comprised 4612 genotypes of Australian Angus beef cattle where each genotype contained 47702 SNPs extracted from Illumina 50K Bead Chip genotypes obtained during the Australian Beef Cooperative Research Center (www.beefcrc.com, Beef CRC) project and from cooperating breeders.

Genotypes

Genotypes of all animals were used in three different test runs: 1) all 40k SNPs were used, 2) a subset of 500 SNPs randomly selected from the full 40k genotypes was used, 3) a subset of 100 SNPs randomly selected from the full 40k genotypes was used, but the sample space was narrowed to those SNPs with a minor allele frequency larger than 0.3.

Animal assignment to the equation

From the set of 4612 animals those 90 individuals were selected which had a genotyped sire and a genotyped dam in the data set. These 90 animals will be called "explained animals" in the remainder of the article. The genotype of each of these animals formed the y vector the above equation. Matrix X always contained 13 columns, but they were filled in three different ways: 1) three columns for the known sire, known dam and the population allele frequency vector, and the remaining columns for a set of 10 randomly selected animals, 2) two columns for the known dam and the population allele frequency vector and the remaining columns for a set of 11 randomly selected animals, and 3) one column for the population allele frequency vector and the remaining columns for a set of 12 randomly selected animals. Animals forming X are called "explanatory animals" in the remainder of the article. Note that the population allele frequency vector was calculated excluding animals in X and y, the randomly selected animals excluded parents, offspring, full sibs and half sibs, and the random animals were re-sampled for every of the 90 explained animals.

Parentage assignment

Two different methods can be used to assign parentage to individuals in X. The first method (ranking method) ranks the coefficients in b after excluding the coefficient for the population allele frequency. Parents are those having the greatest one (two), coefficients. A second method requires setting a minimum threshold for coefficients in b and every animal which has a coefficient below this threshold is ruled out of being a potential parent (threshold method). Since the ranking method would result in parentage assignment even when a true parent is not among the animals forming X, the threshold method was regarded as more appropriate because in an application to real data animals in X may not contain any true parent at all. In addition, if the threshold is set appropriately (>1/3) the optimisation constraint will force the number of animals with a coefficient in b greater than the threshold to be ≤2, thus avoiding "parentage over-assignment".

Assignment statistics

Power of assignment (Pa) was calculated as:

$$\frac{numberofcorrectparentassigmnets}{numberofparents}$$

where the denominator was 180 if both parents were among the animals in X, and 90 if only the dam was among the animals in X, and power of exclusion (Pe) was calculated as:

$$1-\frac{numberofwrongparentassignments}{maximumnumberofparentassignments}$$

where the denominator was always 180.

Software

CGR was implemented in a FORTRAN wrapper executable which called the NLopt library Johnson (2014). The optimisation solver used the augmented Lagrangian algorithm as global solver and the method of moving asymptote as a local solver. All computations

were carried out on an desktop computer with an Intel(R) Core(TM) i7-3770 processor and 32GB of memory.

Results

Parentage assignment

Pa and Pe for different sets of explanatory animals and genotypes are summarised in Table 1. When the whole 40k SNP genotypes were used, Pa and Pe were both 1 (see Table 1). Thus parentage was assigned correctly for all 90 test animals if the pool of explanatory animals contained a true parent. If the pool did not contain any true parent, parentage was ruled out for all animals in the pool because explained animals were described best by the population allele frequency vector. When 500 randomly selected SNPs were used as genotypes, Pa and Pe were still very high with 1 and 0.99 respectively. When at least one parent was among the explanatory animals parentage was always assigned correctly (Pa=1, Pe=1). When no true parent was among the explanatory animals, parentage was correctly ruled out in 179 of 180 cases (Pe=0.99). Decreasing the SNP density further to 100 randomly selected SNPs from those having a minor allele frequency >0.3 made the correct assignment more difficult. If both true parents were among the explanatory animals, parentage was correctly assigned in 173 of 180 cases (Pa=0.96), but all random animals were rejected as parents (Pe=1). If only the true dam was among the explanatory animals parentage was correctly assigned in 89 of 90 cases (Pa=0.99), but all random animals were rejected as parents (Pe=1). If no true parents were among the explanatory animals, parentage was correctly rejected in 178 of 180 cases (Pe=0.99). A statistic for coefficients in b for this SNP set is given in Table 2.

Table 1. Power of assignment (Pa) and power of exclusion (Pe) for different genotypes and sets of explanatory animals.

sets	SNP genotypes					
	40k		500		100	
	Pa	Pe	Pa	Pe	Pa	Pe
both	1	1	1	1	0.96	1
dam	1	1	1	1	0.99	1
none	-	1	-	0.99	-	0.99

power of assignment: probability of assigning the right parent if at least one parent is among the explanatory variables, power of exclusion: probability of rejecting the wrong parent in favour of the right parent or the vector of population allele frequencies, both: both true parents were among the explanatory animals, dam: only the true dam was among the explanatory animals, none: none of the true parents was among the explanatory animals

Table 2. Regression coefficient statistics for different test sets calculated from the 90 test animal evaluations when using 100 SNPs randomly selected from the 40k SNPs if the minor allele frequency >0.3.

coefficient	mean	s	min	max
both parents in the test sets				
sire	0.449	0.055	0.297	0.580
dam	0.441	0.059	0.245	0.615
ran	0.011	0.024	0.000	0.198
mean	0.002	0.012	0.000	0.092
dam in the test sets				
dam	0.512	0.088	0.322	0.770
ran	0.035	0.055	0.000	0.275
mean	0.104	0.124	0.000	0.530
no parents in the test sets				
ran	0.043	0.064	0.000	0.522
mean	0.490	0.206	0.000	0.861

sire: statistics for the coefficients regressing the focused animal on the genotype of the true sire. dam: statistics for the coefficients regressing the focused animal on the genotype of the true dam. mean: statistics for the coefficients regressing the focused animal on the vector of population allele frequencies. ran: statistics for the coefficients regressing the focused animal on the randomly selected animals. The number of random animals was 10 when both parents were among the explanatory animals, 11 when only the dam was used, and 12 when no parents were used as explanatory animals.

Computational demand

Beside reading data, solving time for equation 1 for a single animal was 0.1 real time seconds when using all 40k SNPs. For both the other SNP sets processing time decreased to 0.003 real time seconds per animal.

Discussion

Result show that CGR delivers highly accurate results even when the number of SNPs is as low as 100. This also holds when comparing with results given by Boichard *et al.* (2014), who found the specificity of their algorithm (which is similar to Pe) dropping to ~0.5 if the number of SNPs was ≤100. However, these differences maybe due to different data sets. The core strengths of CGR is the same as of the likelihood based method suggested by the latter authors: the capacity of ruling out parentage if non of the true parents if among the putative parents. CGR may also account for genotyping errors by either excluding the affected loci or replacing the affected loci by an expected value or assigning weights to SNPs reflecting genotype certainty. When genotypes become dense, the assumption about the variance of y may not hold due to arising linkage disequilibrium (LD), but LD will also affect the likelihood formulation of likelihood based methods. However, practical parentage verification aims to minimise the number of used SNPs making inference problems due to neglected LD rather unlikely. CGR relies on the existence of population information which is condensed into the vector of population allele frequencies. One may argue that counting opposing homozygous loci can exists as a stand alone algorithm because it relies on genotyping errors rather than allele frequencies. However, this only holds if the true parent is among the putative parents for sure, which is rather unlikely in practical applications. A way to make counting opposing homozygous loci more versatile is to compare the result of a current pair of animals to the same parameter at many levels of relationships, which in turn requires population information as well, but results may still be biased due to a sample dependant

genotyping error rate. Relaxing constraint 3 may allow CGR to run without a vector of population allele frequencies, but regression coefficients would have to be tested against an empirical distribution which can only be generated from a sufficient number of genotypes.

Conclusion

CGR is a fast, efficient, accurate and easy to implement algorithm to assign parentage on the base of SNP genotypes in samples which contain at least one true parent, or to reject parentage if the samples do not contain a true parent at all. CGR scales automatically to any size of genotypes and has proven to provide accurate results with genotypes comprising only 100 randomly selected SNPs.

Acknowledgements

The author thanks various Australian Angus breeders for supplying genotypes. This work was funded by Meat and Livestock Australia (Project B.BFG.0050).

References

Baruch, E. & Weller, J. 2008. Estimation of the number of snp genetic markers required for parentage verification. *Anim. Genet. 39:5*, 474–479.

Boichard, D., Barbotte, L. & Genestout, L. 2014. Accurassign, software for accurate maximum-likelihood parentage assignment. *In Proc. 10th. WCGALP*, Vancouver, Canada, August p. np.

Ferdosi, M.H. & Boerner, V. 2014. A fast method for evaluating opposing homozygosity in large snp data sets. *Livest. Sci. 166*, 35–37.

Hayes, B. 2011. Technical note: Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. *J. Dairy Sci. 94:4*, 2114–2117.

Heaton, M.P., Leymaster, K.A., Kalbfleisch, T.S., Kijas, J.W., Clarke, S.M., McEwan, J., Maddox, J.F., Basnayake, V., Petrik, D.T., Simpson, B., Smith, T.P.L., Chitko-McKown, C.G. & The International Sheep Genomics Consortium. 2014. Snps for parentage testing and traceability in globally diverse breeds of sheep. *PLoS One, 9:4, e94851.*

Johnson, S.G. 2014. The nlopt nonlinear-optimization package. 2011. *URL http://ab-initio. mit. edu/nlopt*, .

Jones, A.G., Small, C.M., Paczolt, K.A. & Ratterman, N.L. 2010. A practical guide to methods of parentage analysis. *Mol. Ecol. Resour. 101*, 6–30.

Kalinowski, S.T., Taper, M.L. & Marshall, T.C. 2007. Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Mol. Ecol. 16:5*, 1099–1106.

Mäntysaari, E. 2014. Challenges in industry application of genomic prediction experiences from dairy cattle. *In Proc. 10th. WCGALP*, Vancouver, Canada, August pp. 17–22.

Marshall, T., Slate, J., Kruuk, L. & Pemberton, J. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol. 7:5*, 639–655.

McClure, M., McCarthy, J., Flynn, P., Weld, R., Keane, M., O'Connell, K., Mullen, M., Waters, S., Kearney, J. & Kowalski, Z. 2015. Snp selection for nationwide parentage verification and identification in beef and dairy cattle. *ICAR Technical Series 19*, 175–181.

McClure, M., Sonstegard, T.S., Wiggans, G. & Van Tassell, C.P. 2012. Imputation of microsatellite alleles from dense snp genotypes for parental verification. *Front. Genet. 3*, 140.

Strucken, E.M., Gudex, B., Ferdosi, M.H., Lee, H.K., Song, K.D., Gibson, J.P., Kelly, M., Piper, E.K., Porto-Neto, L.R., Lee, S.H. & Gondro, C. 2014. Performance of different snp panels for parentage testing in two east asian cattle breeds. *Anim. Genet. 45:4*, 572–575.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Schenkel, F.S. & Van Tassell, C.P. 2009. Selection of single nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the united states and canada. *J. Dairy Sci., 92:7*, 3431–3436.