Selection of SNP Panels for Parentage Testing

*C. Gondro**, E.M. Strucken*, H.K. Lee†, K.D. Song† and S.H. Lee‡.
*University of New England, Armidale, Australia, †Hankyong National University, Anseong, Korea, ‡National Institute of Animal Science, Pyeonchang, Korea

ABSTRACT: Correct parentage assignment is a fundamental requirement for a successful breeding program so that production performances can be linked back to the correct families to improve estimates of breeding values. In this study, we evaluated the influence of SNP panel sizes for parentage testing in two species (cattle and sheep) with very different genetic structures. Results suggest that small parentage panels will not port well across breeds and should be designed specifically for a particular breed. If cross breed panels are needed around 450 SNP should be used. Finally, we developed an evolutionary algorithm based on differential evolution to optimize selection of SNP marker panels for parentage assignment. Results show that the algorithm is more efficient at selecting markers for the panel than rule based marker selection.
Keywords: parentage testing; SNP; differential evolution

Introduction

Correct parentage assignment is a fundamental requirement for a successful breeding program so that production performances can be linked back to the correct families to improve estimates of breeding values. However, in commercial breeding programs pedigree problems can occur due to missing data, human error or even wilful forgery. In any of these cases, a DNA-based parentage test can clarify the ancestry and help improve the breeding program. Single nucleotide polymorphisms (SNP) are rapidly replacing microsatellites as the marker of choice for parentage testing in livestock due to their ease of automation, lower genotyping cost per marker and standardization between different laboratories (Gudex et al. 2014).

From an information content perspective SNP are only bi-allelic and more of them are needed to obtain the same level of information contained in the highly polymorphic microsatellites. Previous studies have reported on the conversion rate needed to migrate from microsatellites to SNP to maintain the same performance in parentage tests. As a general approximation, between 40 and 100 SNP are equivalent to between 14 and 20 microsatellites (Fisher et al. 2009). The International Society for Animal Genetics (ISAG) assembled not only a microsatellite panel but also, more recently, an SNP panel for parentage testing of *Bos taurus* cattle that should be used internationally to make results comparable between laboratories. Whilst ISAG recommended 12-14 microsatellites markers, an SNP panel should include at least 100 markers. In 2012 ISAG released the current panel which consists of 100 *core* SNP, mostly derived from European breeds plus an additional 100 SNP which also included *Bos indicus* cattle and should improve parentage assignments in zebu and composite breeds (CMMPT 2012).

A parent and its offspring should not have any Mendellian inconsistencies but these do occur due to genotyping errors. In practice, when working with 100 SNP marker panels, one genotype mismatch is usually adopted as an acceptable error rate in true parent-offspring relations. Efficacy of parentage testing panels are mostly discussed in terms of the *power of exclusion*, which means the probability that two randomly chosen individuals are correctly identified as unrelated (Weller et al. 2006) or on the rate of correctly assigned parentage (Fisher et al. 2009; Gudex et al. 2014). However, rates of wrongly assigned or wrongly excluded parentage are infrequently discussed.

In this study, we evaluated the influence of SNP panel sizes for parentage testing in two species with very different genetic structures. We used Korean Hanwoo cattle as representative for a pure-bred heavily selected population with a small effective population size (Ne ~100) and a mixed sheep population (Merino, White Suffolk, Border Leicester, Poll Dorset, Texel and various crosses) as a reference for a large outbred population. To compare the efficacy of different panels, we used a *separation value* which is a useful and simple metric to design parentage panels or to evaluate their efficacy for parentage testing; it simultaneously teases apart true/false positives/negatives and provides some comparative measure of panel reliability (Strucken et al. 2014). The separation value is calculated by first building a square matrix with the number of Mendellian inconsistencies (number of opposing homozygotes) between all pairs of individuals. The value itself is then simply the difference between the minimum number of Mendellian inconsistencies found across all false parent-offspring pairs and the maximum number in the true parent-offspring pairs. The larger the separation value, the better the panel is at resolving parentage assignments and, if the value becomes zero or negative, a perfect separation between true and false parent-offspring relations is impossible. To compare panels of different sizes, the separation value is divided by the number of SNP used. Note however that the metric is subject to sample size bias (with larger samples separation values tend to shrink).

Finally, we developed an evolutionary algorithm based on differential evolution (Storn and Price 1997) to optimize selection of SNP marker panels for parentage assignment.

Materials and Methods

Data. Cattle data consisted of 290 half-sibs from 36 sires genotyped on the 700k Illumina BeadChip array. All animals were pure-bred Korean Hanwoo. Genotypes were split into a *discovery* (20 sires, 152 offspring) and a *validation* (16 sires, 138 offspring) dataset. The sheep population was a mix of various breeds of pure and

crossbred animals genotyped on the 50k Illumina array and consisted of 2,441 half-sibs from 37 sires. The data was split into discovery (20/1,119) and validation (17/1,322). Basic quality control filtering of genotypes was performed on the data. Unmapped SNP and SNP on mitochondria or sex chromosomes were also excluded.

Random marker panels. In both, cattle and sheep, marker panels were randomly selected with varying numbers of SNP. Panels of size 10 to 1,000, in increments of 10, were generated. Between 1,000 and 10,000 the increment was 100; between 10,000 and 100,000 the increment was 1,000 and only for cattle, above 100,000 the increment was 10,000. This resulted in 344 different panel sizes in cattle and 228 in sheep. For each panel size 100 random repeats were generated. For each panel separation values, as defined above, were then calculated such that

$$sv = (min(FR) - max(TR))/nsnp$$

where *FR* is the number of opposing homozygotes in false parent-offspring relations; *TR* is the number of opposing homozygotes in true parent-offspring relations and *nsnp* is the number of SNP in the panel.

Evolutionary algorithm. An algorithm based on Differential Evolution (DE) was developed to optimize the marker panel. To select SNP for the panel, random keys were used. A random key is an evolvable vector of real values (one for each SNP) which are sorted in the objective function and the ranking of the key is used to rank the SNP. The concept is that SNP better for parentage testing evolve to higher values in the key and the rest to lower values; once the keys are sorted they reflect the relative value of a given SNP. An additional parameter to be optimized is the number of SNP in the panel – a *cutoff value*. Basically the DE evolves the cutoff value, sorts the SNP based on their key values and uses the top ranked ones up to the number defined by the cutoff parameter. More in-depth details on the algorithm are given in Gondro and Kwan (2012). The fitness function used was

$$fitness = (min(FR) - max(TR))/nsnp^2$$

with an additional penalty to increase heterozygosity of selected SNP by penalizing the solution as a proportion of their deviation from an average allelic frequency of 0.5, such that

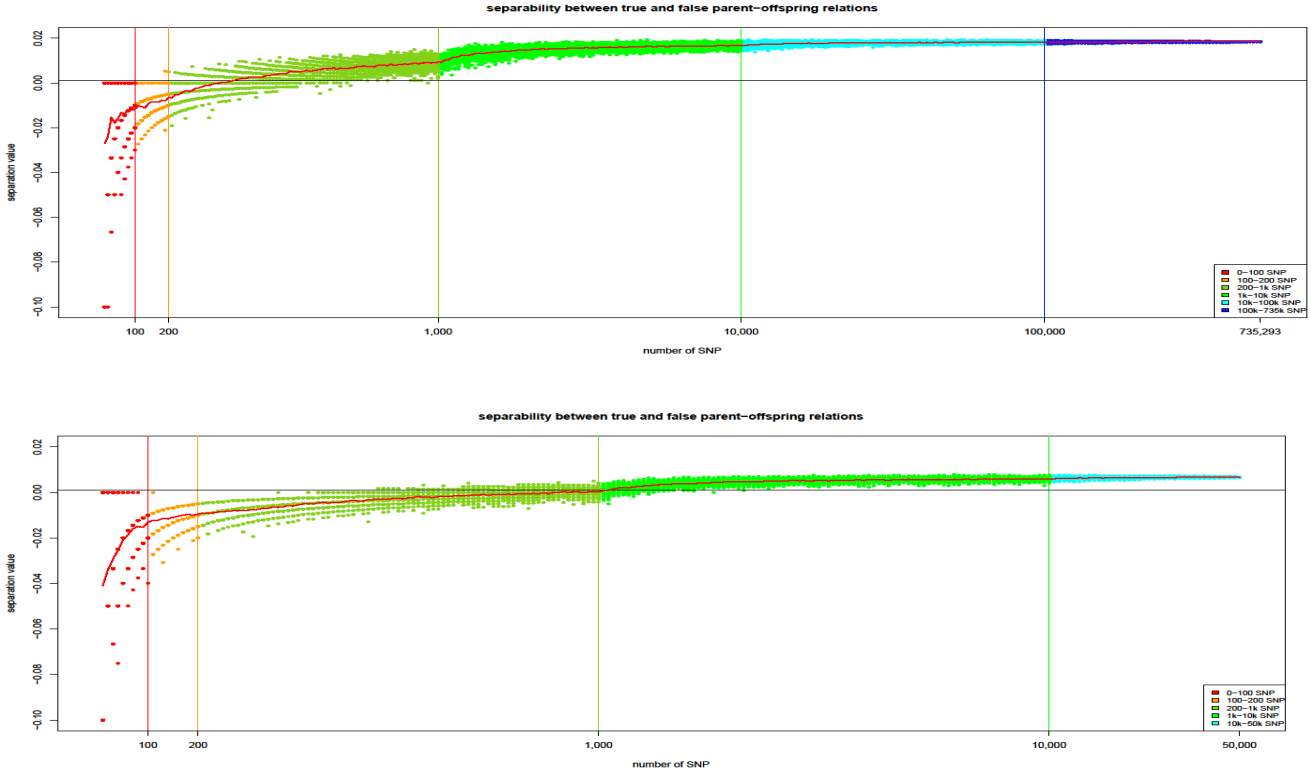
$$fitness = fitness - fitness * (\frac{\sum_1^n abs(SNPfreq_n - 0.5)}{\sum_1^n 0.5})$$

The discovery populations were used to evolve marker panels and the validation set to check accuracy of solutions.

Results and Discussion

Number of markers in panel. Figure 1 shows the separation values in cattle (top pane) and sheep (bottom pane). In both populations, with up to 200 SNP panels the separation values were almost exclusively negative or zero. In cattle, the average values across 100 repeats were -0.0117 (100 SNP) and -0.0065 (200 SNP). For comparison purposes, the ISAG panels had separation values of -0.011 (100 SNP) and 0 (200 SNP). A Hanwoo specific panel recently developed had a separation value of 0.01 (200

Figure 1: Separation values between true and false parent-offspring relations for panels with different numbers of SNP. Points shown are for 100 random repeats for each panel size. Red line shows the average values. Top pane: cattle; bottom pane: sheep. Points above the horizontal black line have positive separation values; i.e. there is no ambiguity in parentage assignment.



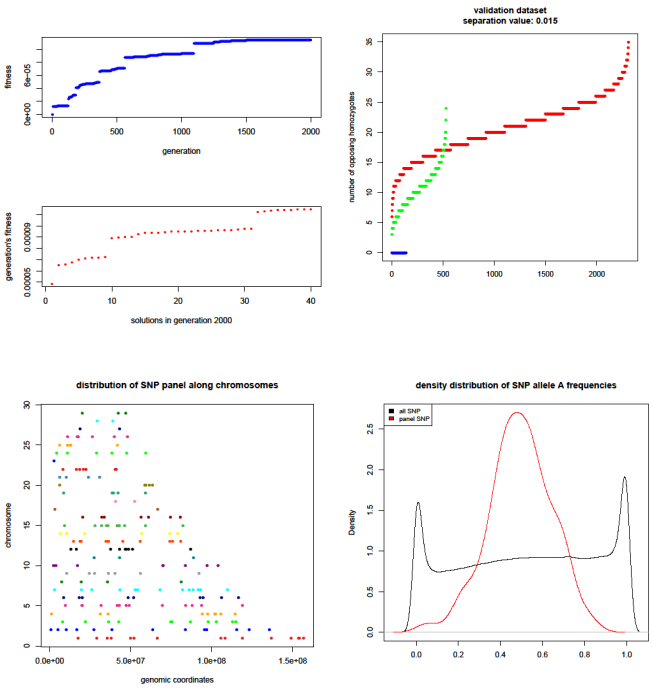
SNP). At around 200 markers positive separation values start to show up (~1-2%) and steadily increase from there (~50% with 400 SNP). With 600 or more SNP, almost any random set is 100% accurate. Sheep followed a similar pattern to cattle; with up to 200 SNP panels the separation values were all negative or zero (averages -0.0132 for 100 SNP and -0.0096 for 200 SNP). The genetically more diverse sheep population needs more SNP to resolve parentage than the single breed Hanwoo. This is probably due to the varied genetic background which makes it difficult to identify a small set of SNP that work well across the full spectrum of diversity. This is seen throughout as separation values are lower than in cattle, even using all 50k SNP. However, at ~2,000 or more SNP any random set is equally adequate for parentage testing but still three times the number needed for the Hanwoo cattle. Results suggest that small panels not designed for a particular breed will be little better than a random marker set. Likewise, to achieve generality across breeds a larger marker panel will be needed. Note that there is a sample size bias against sheep which makes the values not directly comparable; however using equal numbers from a single sheep breed and mixed breeds, the same trend was still evident (data not shown).

Optimization of parentage assignment panel. The evolutionary algorithm was tested with constrained number of SNP (i.e. solutions should have a fixed number: 100, 200 and 400) or unconstrained where the number of SNP was also an optimization parameter. Figure 2 illustrates results in Hanwoo for 200 SNP. The evolved panel had a separation value of 0.015 (with all 700k SNP the value is 0.0185) and was better than any random panel (maximum value in 100 repeats was 0.005), the ISAG panel (0) and a rule based panel recently developed for the breed (0.01). The figure also shows that there is good heterozygosity for selected SNP and they are well spread across the genome. Solutions (cattle and sheep) with positive separation values could be obtained with 100 SNP but they were negative on the validation data, suggesting lack of generality with this number of SNP. In sheep, only solutions with 400 SNP would yield positive separation values in the validation dataset. Unconstrained solutions evolved panels of sizes 180 – 287 SNP in cattle and 432 – 1351 in sheep (10 repeats) again suggesting that mixed sheep breeds need larger panel sizes. Using only pure bred Merino sheep, evolved panel sizes were similar to Hanwoo (data not shown).

Conclusion

Results suggest that small parentage panels (100/200 SNP) will not port well across all breeds and should be designed specifically for a breed or maybe even for populations within a breed. If cross breed panels are needed, at least in sheep, around 450 SNP should be used. This assumes that close to 100% population-wide accuracy is desired; practical applications are usually less demanding since they tend to validate/refute *nominated* parent-offspring pairs. We have also developed an evolutionary algorithm to build marker panels for parentage testing and have shown that it performs better than rule based marker selection.

Figure 2: Results from the evolutionary algorithm for Hanwoo cattle constrained to 200 SNP. Clockwise; evolution of better panels by the evolutionary algorithm during 2,000 generations; positive *separability* (0.015) in the validation dataset between true parent-offspring relations (blue) and false (red/green); SNP selected according to genomic coordinates (X base pair position, Y chromosome); density distribution plot of selected SNP (red) and all SNP (black).



Acknowledgements

This work was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ008196), Rural Development Administration, Republic of Korea and an Australian Research Council Discovery Project DP130100542.

Literature Cited

CMMPT (2012). Cattle Molecular Markers and Parentage Testing Workshop. In: ISAG Conference, Cairns.

Fisher P.J., et al. (2009). J. Dairy Sci. 92:369-74.

Gondro, G., P. Kwan (2012). 351-377 in Multidisciplinary Computational Intelligence Techniques: Applications in Business, Engineering and Medicine. IGI Global.

Gudex B., et al. (2014). Anim. Genet. 45(1):142-143.

Storn, R. and K. Price (1997) J. Global Optim. 11(4):341-359.

Strucken, E.M., et al. (2014). Anim. Genet. (in press).

Weller J.I., E. Seroussi and M. Ron (2006) Anim. Genet. 37:387-9.