INVESTIGATION OF THE POWER AND DESIGN OF USING SNP CHIPS FOR GENOME-WIDE ASSOCIATION STUDIES IN FARM ANIMALS: AN EXAMPLE

B. Auvray and K.G. Dodds
AgResearch Limited, Applied Biotechnologies Group, Invermay Agricultural Centre, Puddle Alley, Private Bag 50034, Mosgiel, New Zealand

SUMMARY
This paper presents a quick, easy to implement and versatile way of using stochastic simulations to investigate the power and design of using SNP chips for genome-wide association studies in farm animals. It illustrates the methodology by discussing a small example where 3 experimental designs are considered: genotyping 96 animals from each of two groups with distinct genetic backgrounds or genotyping 192 animals of one group or the other. Given the resource at hand, genotyping 192 animals from the larger and less related group is identified as the best design.

INTRODUCTION
Genome-wide single nucleotide polymorphism (SNP) chip technologies offer opportunities for gene discovery in farm animals. However, the use of such chips requires careful planning due to technological or cost related constraints or limitations inherent to association studies and other linkage disequilibrium (LD)-based methods. These limitations depend on many factors specific to each experiment. It is therefore very desirable to optimize the design of an experiment in order to maximize power given the resource at hand, before conducting any experiment using SNP chips.

Several studies have focused on computing power for association studies (Kang et al. 2004; Guedj et al. 2006), but these do not address the problem of potential confounding factors such as population relatedness. The use of stochastic simulations is a simple, yet powerful and versatile way to handle this problem. Indeed, replicating in silico an experiment using varying assumptions about the data and different methods of analysis permit a reasonable estimation of the power of the experiment given these factors and thus the identification of a suitable experimental design.

As an example, this paper will compare the power of 3 different designs for a genome-wide SNP association study.

MATERIALS AND METHODS
Description of the example. A real animal resource composed of two groups A and B with distinct genetic backgrounds (distinct breeds) has been chosen for this example. Group A is composed of 1426 animals (74 sires and 318 cows with DNA samples and 1034 ancestors). Group B is composed of 1952 animals (59 sires and 380 cows with DNA samples and 1513 ancestors). Polygenic breeding value (BV) estimates for a trait of interest with a heritability of ~.3 are available for all animals. Both groups show evidence of a large bi-allelic quantitative trait locus (QTL) with an (allelic substitution) effect of .6 phenotypic standard deviations ($\sigma_p$) for the trait of interest segregating in the same chromosomal region but at different frequencies in each group (minor allele frequency (MAF) of ~.4 in group A and ~.2 in group B), and a SNP chip experiment using a 25K SNP chip is to be designed to fine map this QTL, while at the same time screening the genome for other potential QTL. Limited
funds are available and a choice has to be made whether to (1) genotype 96 animals from each group, (2) genotype 192 animals from group A, (3) genotype 192 animals from group B.

**Simulation procedure.** The methodology employed to simulate populations is as follows:

- Twenty ancestral chromosomal segments with 41 evenly spaced SNP loci are created randomly, with all 41 loci being polymorphic (this amounts to ensuring that all the SNP MAF are >.05 at the beginning of the simulation). These 20 segments are paired to create 10 ancestral animals. Chromosomal segments are chosen to be 4.8 cM long so that SNPs are .12 cM apart (and 25,000 SNPs cover a region 30 M long). The middle SNP (SNP 21) is chosen to be a QTL. The 4.8 cM length has been determined as being sufficient to include most SNPs in LD with the QTL.

- The 10 ancestral animals are mated to create the next generation of 100 animals, which is in turn used to create the following generation of 100 animals and so on, with no generation overlapping. For group A, this process is carried out for 10 generations, while it is run for 100 generations for group B. The difference in numbers of generations between groups A and B is due to the need to simulate different LD patterns between the two groups. A slight selection pressure for the QTL is applied during the whole process, in order to ensure that the QTL frequency after the 10 or 100 generations is close to a target QTL frequency.

- The 100 animals simulated in the final generation of this process are used as founders (unknown parents) of the real pedigrees of groups A or B, and gene-dropping is used to simulate the genotypes of the 392 or 439 animals that can potentially be genotyped. The phenotypes of the animals are simulated by adding to the BV of each animal a random residual term and a QTL effect, depending on the genotype of the animal for SNP 21 and a chosen QTL effect size.

**Simulation implementation.** For each group, the following factors are made variable:

- The target QTL MAF is set to each of .1, .2, .3, .4 and .5 and simulations are grouped depending on the actual QTL MAF after the simulation: [0,.1), [.1,.2), [.2,.3), [.3,.4) and [.4,.5].

- The QTL effect is set to 0, .175, .349 and .524 σp.

- 96 or 192 animals are “genotyped”, i.e. are actually used in the analysis of the simulated chip. In each case, the sires are always selected, and the sample is topped up using cows from the tails of the phenotypic distribution.

Due to time constraints, only 181 and 91 repetitions of each combination of factors were carried out for group A and B respectively. As at some times the QTL is lost due to random drift (especially when targeting low QTL MAF) and actual QTL MAF differ from targeted MAF, the real number of repetitions used for each combination of factors varies from 134 to 244 for group A and from 25 to 45 in group B.

**Analysis of the simulated chip and power estimation.** For the 96 or 192 animals genotyped and for each of the 40 SNPs other than the QTL, the following model is fitted: \( y = x b + Z u + e \), where \( y \) is a vector of simulated phenotypes, \( b \) is the fixed regression coefficient of the phenotype on the number of copies of the first SNP allele, \( u \) is a vector of random genetic additive effects, \( e \) is a vector of random residuals, \( x \) is a vector of number of copies of the first SNP allele and \( Z \) is an identity incidence matrix. The statistic \( W = b^\prime \sqrt{C_{bb}} b \) is then calculated, where \( C_{bb} \) is the element relating to \( b \) in the inverse of the left hand side of the mixed model equations used to solve the model above. An empirical 5% \( \alpha \) significance threshold, dependent on the group considered and the number of animals genotyped is computed using all simulations for the group and number of animals genotyped
with a null QTL effect. This empirical 5% $\alpha$ significance threshold is then used to estimate two other significance thresholds adjusted for multiple testing: a “candidate region-wise” significance threshold, adjusted for the estimated effective number of independent tests carried out on the 4.8 cM region, and a “genome-wise” significant threshold, adjusted for the estimated effective number of independent tests that would be carried if a full 25K chip was used on a 30 M genome. For each group, the “candidate-wise” and the “genome-wise” powers are estimated for each combination of factors as the number of times at least one SNP is found significant at the appropriate significance level divided by the number of simulations with this combination of factors.

RESULTS AND DISCUSSION

Power results. Figures 1 and 2 show respectively the power values for group A and B, for 96 and 192 animals and “candidate-wise” and “genome-wise” significance thresholds, for the 5 ranges of QTL MAF.

The power is generally lower for group A. The reasons for this finding are not clear. It could be due to the average relationship between the animals in this group (.071 between sires and .043 between all animals with DNA samples) being higher than between animals in group B (.067 between the sires, .026 between all animals with DNA samples). It could also be due to the sampling strategy, imposing the selection of all sires and so leaving less extreme cows available for genotyping in group A. A third explanation would be the difference in LD patterns between the 2 groups, with group A exhibiting more useful LD ($r^2$) over medium distances (.2cM to 1cM) but less over small (<.2 cM) or longer (>1cM) distances. More analyses would be required to validate those hypotheses.

Overall, the power levels obtained when genotyping 96 animals are low for QTL with an effect smaller than .5 $\sigma_p$, even with favourable (high) QTL MAF and looking at the “candidate-wise” power curves. This indicates that using 96 animals from group A and 96 animals from group B to detect something even in candidate regions would be
difficult if the effect of the QTL has been slightly overestimated and the chances of finding new QTL of small size ($\leq 0.2 \sigma_p$) that would not be present in both groups are extremely low. Another interesting point to note is the huge effect that the QTL MAF has on the power, as it can range from .2 to ~1 when the QTL MAF goes from .1 to .5.

Given the results, it seems that the best strategy would be to concentrate all resources into genotyping 192 animals from group B as opposed to genotyping 96 animals of each group, or 192 of the group A. Nevertheless, although higher than for group A, power levels obtained for group B and 192 animals are also low and not many new small QTL would be found using such an experimental design.

**Conclusion.** This example presents an easy and quick way to implement stochastic simulations to investigate the power and design of using SNP chips for genome-wide association studies. A number of general conclusions can be derived from this small study. Firstly and evidently, it doesn’t look like 192 animals (let alone 96) would normally be enough for a comprehensive genome-wide association study using SNP chips. Secondly, besides the obvious effects of the QTL effect size and the number of animals genotyped, the large effect that the QTL MAF has on the power has been clearly demonstrated. Lastly and more importantly, it is shown that power curves can vary widely between 2 seemingly similar animal resources, perhaps due to different responses to a particular experimental design because of different population histories or pedigree structures. This highlights the need to take these into account when designing a genome-wide association study using animal resources made up of related animals. In conclusion, this study reaffirms the need for careful planning of genome-wide association studies using SNP chips.

**REFERENCES**