Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Use of stochastic simulations to investigate the power and design of a whole genome association study using single nucleotide polymorphism arrays in farm animals*

AUVRAY Benoît†, DODDS Ken G.

(*Applied Biotechnologies Group, AgResearch Limited, Invermay Research Centre, Private Bag 50034, Mosgiel 9053, New Zealand*)

†E-mail: benoit.auvray@agresearch.co.nz

Received Sept. 19, 2007; revision accepted Sept. 20, 2007

Abstract: This paper presents a quick, easy to implement and versatile way of using stochastic simulations to investigate the power and design of using single nucleotide polymorphism (SNP) arrays for genome-wide association studies in farm animals. It illustrates the methodology by discussing a small example where 6 experimental designs are considered to analyse the same resource consisting of 6 006 animals with pedigree and phenotypic records: (1) genotyping the 30 most widely used sires in the population and all of their progeny (515 animals in total), (2) genotyping the 100 most widely used sires in the population and all of their progeny (1 102 animals in total), genotyping respectively (3) 515 and (4) 1 102 animals selected randomly or genotyping respectively (5) 515 and (6) 1 102 animals from the tails of the phenotypic distribution. Given the resource at hand, designs where the extreme animals are genotyped perform the best, followed by designs selecting animals at random. Designs where sires and their progeny are genotyped perform the worst, as even genotyping the 100 most widely used sires and their progeny is not as powerful of genotyping 515 extreme animals.

Key words: Simulation, Association study, Single nucleotide polymorphism (SNP), Power, Quantitative trait loci (QTL)
doi:10.1631/jzus.2007.B0802 **Document code:** A **CLC number:** Q78; TP31

INTRODUCTION

Genome-wide single nucleotide polymorphism (SNP) array technologies offer opportunities for gene discovery in farm animals (Barendse *et al*., 2007). However, the use of such arrays requires careful planning due to technological or cost-related constraints or limitations inherent to association studies and other linkage disequilibrium (LD)-based methods (Lawrence *et al*., 2005; Clark *et al*., 2005). 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 arrays.

Several studies have focused on computing power for association studies (Kang *et al*., 2004; Guedj *et al*., 2007), 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 permits 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 compares the power of six different designs for a genome-wide SNP association study.

MATERIALS AND METHODS

Description of the example

A real animal resource composed of 6 006 ani-

* Project supported by Ovita Limited, Dunedin, New Zealand

mals born between 1980 and 2005 has been chosen for this example. Polygenic breeding value estimates (EBV) for a trait of interest with a heritability of ~0.2 are available for all animals. A genome-wide association study using a 60K SNP array is to be designed, and the 6 following designs are envisaged: (1) genotyping the 30 most widely used sires and all of their progeny, which represents 515 animals in total ("sires-515"), (2) genotyping the 100 most widely used sires in the population and all of their progeny, which represents 1 102 animals in total ("sires-1102"), (3) genotyping 515 animals randomly selected ("rand-515"), (4) genotyping 1 102 animals randomly selected ("rand-1102"), (5) genotyping 515 animals from the tails (i.e. the most extreme) of the phenotypic distribution ("extr-515") or (6) genotyping 1 102 animals from the tails of this distribution ("extr-1102").

Simulation procedure

The methodology employed to simulate populations is as follows:

1. 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 minor allele frequencies (MAFs) are ≥0.05 at the beginning of the simulation). These 20 segments are paired to create 10 ancestral animals. Chromosomal segments are chosen to be 2 cM long so that SNPs are 0.05 cM apart (and 60 000 SNPs cover a region 30 M long). The middle SNP (SNP 21) is chosen to be a quantitative trait locus (QTL). The 2 cM length has been determined as being sufficient to include most SNPs in LD with the QTL.

2. 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. This process is carried out for 10 generations. A slight selection pressure for the QTL is applied during the whole process, in order to ensure that the QTL frequency after the 10 generations is close to a target QTL frequency.

3. The 100 animals simulated in the final generation of this process are used as founders (unknown parents) of the real pedigree, and gene-dropping (MacCluer *et al*., 1986) is used to simulate the genotypes of the 6 006 animals that can potentially be genotyped. The phenotype of an animal i is simulated as:

$$y_i=x_iq+u_i+e_i,$$

where y_i is the simulated phenotype, x_i is the number of copies of allele 1 at the QTL, q is the QTL (allelic substitution) effect used in the simulation, u_i is the EBV of animal i and e_i is a random residual drawn from $N(0,(1-h^2)\sigma_u^2/h^2)$, with the additive genetic variance (σ_u^2) equal to 0.543 and the heritability (h^2) equal to 0.2.

Simulation implementation

For each group, the following factors are made variable:

1. The target QTL MAF is set to 0.1 and 0.4.

2. The QTL effect is set to 0, 0.25 phenotypic standard deviations (σ_p) and $0.50\sigma_p$.

3. Different animals are "genotyped", i.e., are actually used in the analysis of the simulated array, depending on the design chosen.

Two hundred replicates of each combination of factors were carried out. Due to the stochastic nature of the simulation, the actual QTL MAF for each replicate differs from the targeted MAF, and at some times the QTL may even be lost due to random drift (especially when targeting the low QTL MAF).

Analysis of the simulated array and power estimation

For the animals genotyped and for each of the 40 SNPs other than the QTL, the following mixed model is fitted:

$$\boldsymbol{y}=\boldsymbol{x}b+\boldsymbol{Z}\boldsymbol{u}+\boldsymbol{e},$$

with

$$var\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{pmatrix}=\begin{pmatrix}\sigma_u^2\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{0} & \sigma_e^2\boldsymbol{I}\end{pmatrix},$$

where $\boldsymbol{y}$ 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, $\boldsymbol{u}$ is a vector of random genetic additive effects, $\boldsymbol{e}$ is a vector of random residuals, $\boldsymbol{x}$ is a vector of number of copies of the first SNP allele, $\boldsymbol{Z}$ is an identity incidence matrix, $\boldsymbol{I}$ is an identity matrix and $\boldsymbol{A}$ is the additive relationship matrix computed from the pedigree. The statistic

$W=|b|/\sqrt{\boldsymbol{C}^{bb}}$ is then calculated, where $\boldsymbol{C}^{bb}$ is the element relating to b in the inverse of the left hand side of the mixed model equations (Henderson, 1984) used to solve the model above.

For each design, empirical "genome-wise" 5% α significance thresholds are computed as follows:

1. Assuming that genomes consist of 1500 independent chromosomal segments of 2 cM, each with 40 SNPs, for a total of 30 M and 60000 SNPs, a "genome-wise" Bonferroni corrected 5% α ($\alpha_{\text{genome-wise}}$) is computed as (Shaffer, 1995):

$$\alpha_{\text{genome-wise}}=1-(1-5\%)^{(1/1500)}=3.42\times10^{-5}.$$

2. The null distribution of the maximum W across the 40 SNPs is obtained by using each replicate of the design where a null QTL effect has been simulated, and the "genome-wise" threshold T is taken to be the $1-\alpha_{\text{genome-wise}}$ quantile of that distribution. As only 400 replicates have been carried out at most, the threshold is interpolated, and numerically very close to the maximum W.

Then, for each design, the function *loess* from the base package of the R software (R Development Core Team, 2007) is used (with the default parameters) to model the power by fitting locally second degree polynomials by least squares as:

$$\boldsymbol{p}=f(\boldsymbol{MAF},\boldsymbol{q}),$$

where $\boldsymbol{p}$ is a vector of indicators equal to 1 when $W\geq T$ and 0 otherwise, $\boldsymbol{MAF}$ is a vector of minor allele frequencies and $\boldsymbol{q}$ is a vector of QTL effects.

RESULTS AND DISCUSSION

The results can be observed from various angles. Fig.1 displays for each of the 6 experimental designs the power surfaces as a function of the QTL MAF and effect. Fig.2 shows for each design the power curves as a function of the QTL effect size, for 3 QTL MAF (0.1, 0.25 and 0.4), while Fig.3 presents the power curves as a function of the QTL MAF, for 3 QTL effect sizes ($0.1\sigma_p$, $0.25\sigma_p$ and $0.4\sigma_p$).

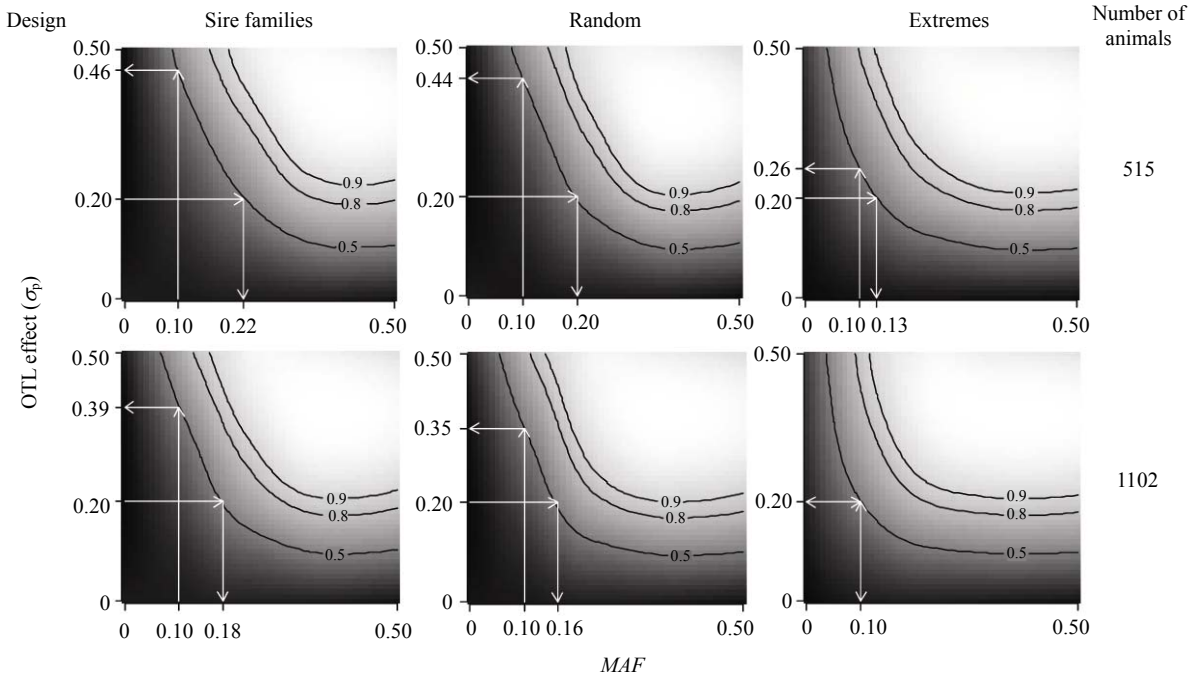

Fig.1 Power surface as a function of the QTL MAF and effect for each of the 6 designs

The rows represent the 3 sampling methods ("sires", "rand" or "extr") and the columns the number of animals genotyped (515 or 1102). The colour black corresponds to power=0 and the colour white to power=1. Contour lines of equal power for power=0.5, 0.8 and 0.9 are plotted. The white arrows exemplify the difference between designs by highlighting respectively the QTL effect size needed to reach a statistical power of 0.5 when chasing a QTL with a MAF of 0.1 and the QTL MAF needed to reach a statistical power of 0.5 when chasing a QTL with an effect of $0.2\sigma_p$. The behaviour of the contour lines going back up for large MAF is an artefact of the power modelling procedure

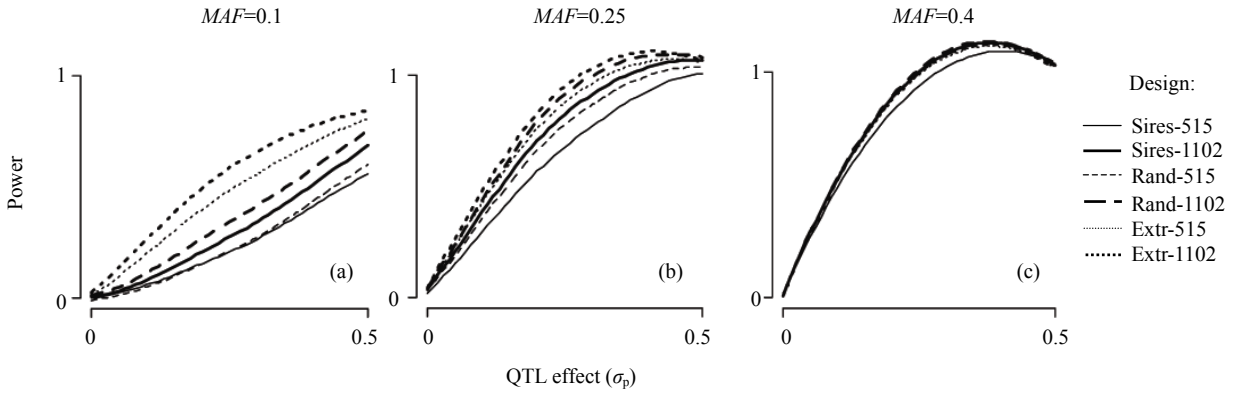

Fig.2 Power as a function of the QTL effect for 3 QTL MAF for each of the 6 designs. (a) *MAF*=0.1; (b) *MAF*=0.25; (c) *MAF*=0.4

Power values >1 or decreasing as the QTL effect sizes increase (observed for large QTL effects and MAF) are artefacts of the power modelling procedure

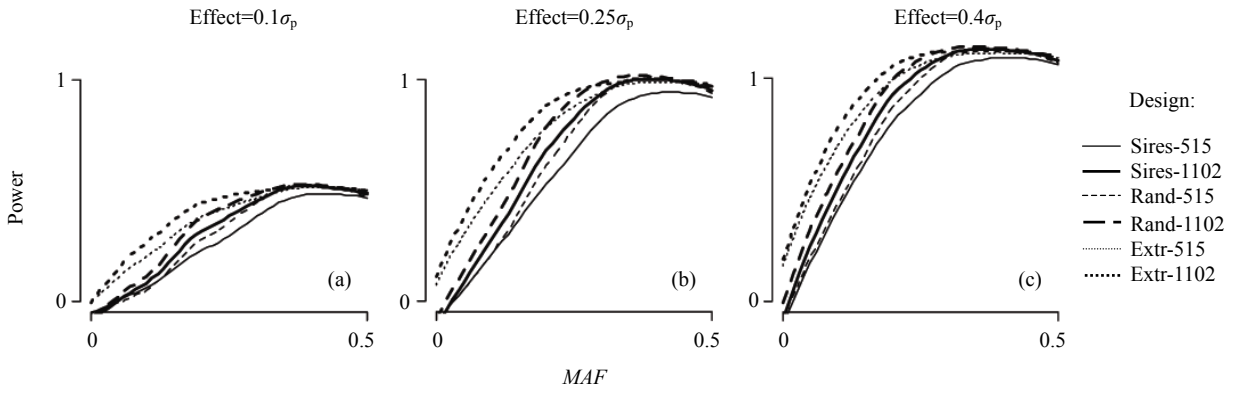

Fig.3 Power as a function of the QTL MAF for 3 QTL effect sizes for each of the 6 designs. (a) Effect=0.1σ_p; (b) Effect=0.25σ_p; (c) Effect=0.4σ_p

Power values >1 or decreasing as the QTL MAF increases (observed for large QTL effects and MAF) as well as power values <0 are artefacts of the power modelling procedure

Obviously, the power is lower when fewer animals are genotyped. Maybe not surprisingly as the increased power of selective genotyping designs has been largely documented [see for instance in (van Gestel *et al*., 2000)], the power is also lower for the designs genotyping the sires compared to those genotyping the extreme animals, even when comparing designs "sires-1102" and "extr-515", with the first having more than twice the number of animals genotyped. This difference can be very large when the QTL MAF is small (<0.2). The power curves obtained for the two "extr" designs are very similar.

The designs "rand" perform slightly better than the designs "sires". The reasons for this are unclear. The reduction of power does not seem to be due to a difference between the phenotypic distributions of both designs as these are very similar. The LD patterns are also comparable between the 2 designs, with for instance the distribution of the r^2 between adjacent SNPs having similar distributions with mean ~0.08 and variance ~0.03 for both designs. Despite having similar LD, the 2 designs have very different average additive relationship between animals, much higher (~4 times higher when genotyping 1 102 and up to ~7 times higher when genotyping 515) in the "sires" groups than in the "rand" groups. This could lead to more confounding between polygenic and QTL effects in the "sires", thus causing a reduction of the power by decreasing the significance of the SNP effects.

Overall, the power levels obtained with design "sires-515" are too low to be practically useful when

the QTL MAF is small (≤0.2) or the QTL effect is small to moderate (≤0.25σ_p), while the power levels for design "sires-1102" are not acceptable when both QTL MAF and QTL effect are small to moderate. For a QTL effect of ~0.1σ_p or under, all 6 designs have low power.

Given the results, it seems that the best strategy would be to concentrate all resources into genotyping the tails of the phenotypic distribution. It also seems that more than doubling the number of animals from 515 to 1102 is unnecessary if chasing QTL with effect sizes over 0.25σ_p. Nevertheless, power levels obtained for low QTL MAF (~0.1) and QTL effect size of 0.25σ_p or under are relatively low even for the design "extr-1102" and suggest that not many new small or rare 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 arrays for genome-wide association studies.

A number of general conclusions can be derived from this small study. Firstly, genotyping the extreme animals is more powerful a design than selecting animals based on relative contribution in a pedigree or at random. However, the increase of power observed with selective genotyping is of course mitigated by the fact that genotypes can be used to map QTL for different traits, thus reducing the applicability of these designs, but also by the potential greater value (for an animal industry, for instance) of having the genotypes of more widely used animals, compared to rarely used ones or random animals. 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. In conclusion, this study reaffirms the need for careful planning of genome-wide association studies using SNP arrays.

References

Barendse, W., Reverter, A., Bunch, R.J., Harrison, B.E., Barris, W., Thomas, M.B., 2007. A validated whole-genome association study of efficient food conversion in cattle. *Genetics*, **176**(3):1893-1905. [doi:10.1534/genetics.107.072637]

Clark, A.G., Boerwinkle, E., Hixson, J., Sing, C.F., 2005. Determinants of the success of whole-genome association testing. *Genome Res.*, **15**(11):1463-1467. [doi:10.1101/gr.4244005]

Guedj, M., Della-Chiesa, E., Picard, F., Nuel, G., 2007. Computing power in case-control association studies through the use of quadratic approximations: application to meta-statistics. *Ann. Hum. Genet.*, **71**(2):262-270. [doi:10.1111/j.1469-1809.2006.00316.x]

Henderson, C.R., 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Canada.

Kang, S.J., Gordon, D., Finch, S.J., 2004. What SNP genotyping errors are most costly for genetic association studies? *Genet. Epidemiol.*, **26**(2):132-141. [doi:10.1002/gepi.10301]

Lawrence, R.W., Evans, D.M., Cardon, L.R., 2005. Prospects and pitfalls in whole genome association studies. *Philosophical Transactions of the Royal Society B Biological Sciences*, **360**(1460):1589-1595. [doi:10.1098/rstb.2005.1689]

MacCluer, J.W., Vandeberg, J.L., Read, B., Ryder, O.A., 1986. Pedigree analysis by computer simulation. *Zoo Biol.*, **5**(2):147-160. [doi:10.1002/zoo.1430050209]

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, Http://www.R-project.org

Shaffer, J.P., 1995. Multiple hypothesis testing. *Ann. Rev. Psychol.*, **46**:561-584. [doi:10.1146/annurev.ps.46.020195.003021]

van Gestel, S., Houwing-Duistermaat, J.J., Adolfsson, R., van Duijn, C.M., van Broekhoven, C., 2000. Power of selective genotyping in genetic association analyses of quantitative traits. *Behav. Genet.*, **30**(2):141-146. [doi:10.1023/A:1001907321955]