Note

Sire design power calculation for QTL mapping experiments

Antonello Carta[a*], Jean-Michel Elsen[b]

[a] Istituto Zootecnico e Caseario per la Sardegna,
Bonassai, 07040 Olmedo (SS), Italy
[b] Station d'amélioration génétique des animaux, Inra, BP 27,
31326 Castanet-Tolosan cedex, France

Abstract – Estimates of sire design power for QTL mapping experiments obtained using three different methods of algebraic approximation were analysed by comparing them with the results of data simulations. Even when the binomial probability that any number of sires out of the total number of sires are jointly heterozygous at the marker and the QT loci was taken into consideration, the algebraic approximations overestimated powers. However, they could be used to rank designs differing in the number of sires if the total size of the experiment is given. The results were discussed, focusing on the assumptions made about the number of informative offspring, the balance between the two offspring sub-groups which receive the same marker allele from the sire and the distribution of the statistic. Given that a full algebraic approach would be computationally costly, data simulation can be considered a useful tool in estimating the power of QTL detection sire designs. © Inra/Elsevier, Paris

QTL/ power/ simulation/ protocol design

Résumé – Calcul de la puissance de détection de QTL dans un modèle « père ». Trois méthodes analytiques pour l'estimation de la puissance du protocole fille pour la détection des QTLs à l'aide d'un marqueur flanquant ont été étudiées en comparaison avec des résultats obtenus par simulation. Ces estimations sont surestimées, même quand est prise en compte la distribution de probabilité du nombre de pères double hétérozygotes au marqueur et au QTL. Cependant, elles peuvent être utilisées pour classer des protocoles de façon relative, à taille de population totale fixée. Les résultats sont discutés en référence aux hypothèses sur le nombre de descendants informatifs, la balance entre les descendants selon l'allèle marqueur reçu de leur père, et la nature des distributions. Compte tenu du coût numérique élevé d'un calcul analytique complet, les simulations demeurent un outil efficace pour l'estimation de la puissance de ces protocoles de détection de QTL. © Inra/Elsevier, Paris

QTL/ puissance/ simulation/ planification expérimentale

* Correspondence and reprints
E-mail: izcszoo@tin.it

1. INTRODUCTION

The use of genetic markers to locate genes whose polymorphism partly explains the genetic variability of quantitative traits was proposed by Sax [3] and further detailed by Neimann-Sørensen and Robertson [2] and others. The principle is to identify, in the offspring of an individual, those which received one or other of the two chromosomal fragments surrounding the marker in question. If a quantitative locus is located on this fragment, and if the parent is heterozygous at both the marker and QTL (quantitative trait locus), then a systematic difference is observed between the two sub-groups of progeny. With the development of molecular markers based on DNA variations, the application of these ideas has become feasible on a large scale particularly in livestock populations, where large families are routinely recorded. The design of such experiments has been studied in detail by a number of authors, in particular Soller and Genizi [4] and Weller et al. [6]. In order to optimize these designs, it is necessary to estimate their power. Focusing on simple population structures, Soller and Genizi [4], as well as Weller et al. [6], approached this power estimation considering fully balanced populations, and using approximate distributions of the test statistic. In these early papers, markers were studied one by one, and the test statistics applied were simple ANOVA methods, modelling trait means as linear combinations of sire and marker within sire effects. In their approximation, these authors worked with the asymptotic χ^2 or normal approximation of the F statistic, and considered simply the mean contrast averaging different possibilities for the sire and offspring genotypes at the QT and marker loci. The power of such designs, as well as more complex experiments involving two or three generations and mixing half- and full-sib families, was further studied by van der Beek et al. [5]. In their paper, these authors considered the mixture of sub-populations, as characterized by the number of heterozygous sires at the QTL, rather than the mean.

Alternatively, the estimate of the design power may be obtained by simulating heterogeneous populations and applying studied test statistics to the generated sets of data, without any approximation, but at the expense of more computing time. This approach was followed by Le Roy and Elsen [1] in a study addressing the relative value of ANOVA and maximum-likelihood methods for QTL detection.

The aim of this study is to evaluate the validity of approximate sire design power estimates, by comparing three algebraic methods with simulating data.

2. HYPOTHESES AND COMPARED METHODS

2.1. Hypotheses

Powers were calculated for a single marker analysis. Multiallelic marker loci (with $na = 4$ alleles) were studied. Alleles M_i were assumed to be distributed with frequencies in a geometric series ($f_1 = f$, $f_2 = \alpha f$, $f_3 = \alpha^2 f, \ldots$, with $f = 1/(1+\alpha+\alpha^2 \ldots)$). In this situation, the parameter a was obtained, given the mean heterozygosity of the marker ($E(fhm)$), solving the equation $E(fhm) = 1 - \Sigma_i(\alpha^{i-1})^2/(\Sigma_i\alpha^{i-1})^2$. This marker was supposed to be totally linked to

the QTL. The design was organized with np half-sib families comprising no progenies per sire. mp was the expected number of sires for which a marker contrast can be computed, i.e. the expected number of heterozygous sires at the marker locus, and lp, the expected number of heterozygous sires at both marker and QT loci. mo was the expected effective family size, i.e. the mean number of offspring per sire for which the marker allele received from the sire is identified. This effective family size is linked to the allele frequencies by the relation: $mo = \Sigma_{ij} 2 f_i f_j (1 - 0.5(f_i + f_j))/\Sigma_{ij} 2 f_i f_j$. The first type error α (accepting a linked QTL when it does not exist) was fixed at 1 %.

2.2. Compared methods

The following three approximations were studied.

1) The approximation used by Weller et al. [6]: in this approximation, only mean sire and daughter numbers were considered. The power was given by $P1 = P\big[F(NC(lp), mp, mp(mo - 2)) \geqslant f\big]$, where $F\big(NC(lp), mp, mp(mo - 2)\big)$ is a non-central F variable with a non-centrality parameter $NC(lp)$ and mp and $mp(mo - 2)$ degrees of freedom. The threshold f corresponds to the $(1 - \alpha)$ percentile of the central F distribution. The $NC(lp)$ is computed as: $NC(lp) = lpE^2(MC)/SE^2(MC)$, where $E^2(MC)$ is the square of the expectation of a marker contrast and $SE^2(MC)$ is the square of the standard error of the marker contrast. If a sire is heterozygous at the QTL, then $E^2(MC) = GE^2(1 - r)^2$, where GE^2 is the square of the gene effect and r the recombination rate between the marker locus and the QTL. For a half-sib family SE is calculated as $(4 - h^2)/mo$ where h^2 is the polygenic heritability of the trait (within QTL genotype).

2) The approximation followed by van der Beek et al. [5]: in this approximation, the variability in number of heterozygous sires at the QTL is considered. The power was given by:

$$P2 = \sum_{xp=0}^{mp} \Pr(xp/mp) P\big[F(NC(xp), mp, mp(mo - 2)) \geqslant f\big]$$

where xp is the number of heterozygous sires at the QTL and $\Pr(xp/mp)$ is the binomial probability that xp out of mp (the expected number of heterozygous sires at the marker locus) are heterozygous also at the QTL.

3) An approximation where variation at both the sire marker and the QT loci are considered. The power was given by:

$$P3 = \sum_{yp=0}^{np} \Pr(yp/np) \sum_{xp=0}^{yp} \Pr(xp/yp) P\big[F(NC(xp), yp, yp(mo - 2)) \geqslant f\big]$$

where yp is the number of heterozygous sires at the marker locus and $\Pr(yp/np)$ is the binomial probability that yp out of np sires are heterozygous at the marker locus.

4) In order to test the reliability of the three algebraic methods above, the design power was also estimated by simulating data and applying the standard

F test. For each power calculation 10 000 replicates were used under the null and the alternative hypotheses. The variance ratio for the classic hierarchical ANOVA was calculated as:

$$\frac{\sum\limits_{i=1}^{yp} \left(\bar{z}_{iM1} - \bar{z}_{iM2}\right)^2 / \left(\frac{1}{n_{iM1}} + \frac{1}{n_{iM2}}\right)}{yp}$$

$$x \frac{\sum\limits_{i=1}^{yp}(n_{iM1} + n_{iM2}) - 2yp}{\sum\limits_{i=1}^{yp}\left[\sum\limits_{k=1}^{niM1}\left(z_{iM1k} - \bar{z}_{iM1}\right)^2\right] + \left[\sum\limits_{k=1}^{niM2}\left(z_{iM2k} - \bar{z}_{iM2}\right)^2\right]}$$

where Z_{iM1k} (resp. Z_{iM2k}) are the quantitative performances of the jth daughter of an heterozygous M1M2 sire i, which received marker allele M1 (resp. M2), and n_{iM1} (resp. n_{iM2}) is their number. The power was estimated by the ratio between the number of replicates under the alternative hypothesis whose statistic exceeds a certain threshold and the total number of replicates. The threshold was the $(1 - \alpha)$ percentile of the 10 000 replicates under the null hypothesis. Thus, no assumptions about the distribution of the statistic were made.

3. RESULTS

Table I reports the power estimates of sire designs with a half-sib family structure for a gene effect (GE) of 0.5 or 1 phenotypic standard deviation (σ_p), for various numbers of sires, for two total experiment sizes (*tno* equal to 500 or 1 000 daughters), for a constant polygenic heritability h^2 of 0.25 and assuming a recombination rate (r) of 0. Expected heterozygosities at both loci, marker and QT, are assumed to be 0.5. Four alleles are segregating at the marker locus with frequencies 0.664, 0.229, 0.079 and 0.028. Note that the total heritability (including the variation at the QTL) equals 0.375 if GE = 0.5, 0.75 if GE = 1.0. It is shown that when the gene effect is one half σ_p and the total experiment size is 500 daughters, the three algebraic methods give similar results and, considering that the power is low in this situation, these approximations only slightly overestimated the power as compared to the simulated data. The results for the same GE but with a total experiment size of 1 000 daughters, confirm that no significant differences exist between algebraic methods except when the number of sires is low in which case $P1$ greatly overestimated the power. The overestimation of algebraic methods with respect to simulations is more important here than it is with a total experiment size of 500 daughters.

As regards the GE of 1.0 σ_p, when the total experiment size is 500 daughters algebraic results continued to overestimate power except for $P3$, when the number of sires is equal to 2, in which case $P1$ gives particularly high power compared to the other algebraic and simulation methods. For a total experiment size of 1 000 daughters, $P1$ greatly overestimated power for any considered number of sires, while $P2$ and $P3$ give results more similar to simulated data.

Table I. Powers (*P1*, *P2*, *P3*, *P4*) and expected effective family sizes (*mo*) of experiments with a half-sib family structure for marker and QT loci that both have 0.5 expected heterozygosity, for two total experiment sizes (*tno*), two gene effect values (GE) expressed in phenotypic standard deviation units and various numbers of families (*np*).

tno	GE	np	mo	P1	P2	P3	P4
500	**0.5**	**2**	152	0.36	0.36	0.32	0.25
		4	76	0.27	0.29	0.28	0.18
		6	51	0.21	0.23	0.23	0.14
		8	38	0.18	0.19	0.19	0.10
		10	30	0.16	0.17	0.17	0.10
		12	25	0.14	0.15	0.15	0.08
		14	21	0.12	0.13	0.13	0.07
		16	19	0.11	0.12	0.12	0.07
		18	16	0.10	0.10	0.10	0.06
		20	15	0.10	0.10	0.10	0.06
	1.0	**2**	152	0.97	0.51	0.44	0.44
		4	76	0.94	0.72	0.66	0.56
		6	51	0.90	0.75	0.72	0.54
		8	38	0.87	0.75	0.72	0.49
		10	30	0.84	0.74	0.72	0.49
		12	25	0.80	0.72	0.70	0.43
		14	21	0.76	0.70	0.68	0.40
		16	19	0.74	0.69	0.67	0.40
		18	16	0.70	0.65	0.64	0.37
		20	15	0.69	0.65	0.64	0.36
1 000	**0.5**	**2**	305	0.72	0.49	0.43	0.39
		4	152	0.61	0.55	0.51	0.38
		6	101	0.54	0.50	0.49	0.32
		8	76	0.48	0.47	0.46	0.30
		10	61	0.44	0.43	0.42	0.27
		12	51	0.40	0.40	0.39	0.26
		14	43	0.36	0.37	0.36	0.25
		16	38	0.33	0.34	0.34	0.21
		18	34	0.31	0.32	0.32	0.20
		20	30	0.29	0.30	0.30	0.20
	1.0	**2**	305	1.00	0.51	0.44	0.44
		4	152	1.00	0.75	0.69	0.68
		6	101	1.00	0.87	0.82	0.74
		8	76	1.00	0.91	0.87	0.79
		10	61	1.00	0.92	0.90	0.79
		12	51	1.00	0.93	0.91	0.81
		14	43	0.99	0.93	0.92	0.81
		16	38	0.99	0.94	0.92	0.78
		18	34	0.99	0.94	0.92	0.77
		20	30	0.99	0.94	0.92	0.77

Polygenic $h^2 = 0.25$; $r = 0$.

A. Carta, J.-M. Elsen

Power estimates for a constant total experiment size and number of sires (1 000 and 10, respectively), for two GE values (0.5 and 1 σ_p) with various expected frequencies of heterozygosity at the marker locus ($E(fhm)$) and at the QTL ($E(fhq)$) are shown in *table II*.

Table II. Powers ($P1$, $P2$, $P3$, $P4$) of experiments with ten half-sib families and a total experiment size of 1 000 daughters for different expected frequencies of heterozygosity at the marker ($E(fhm)$) and QTL ($E(fhq)$) loci and for two gene effect values (GE) expressed in phenotypic standard deviation units.

$E(fhm)$	$E(fhq)$	GE	Total h^2	mo	$P1$	$P2$	$P3$	$P4$
0.25	0.10	0.5	0.275	53	0.03	0.04	0.04	0.02
	0.20		0.300		0.06	0.07	0.08	0.02
	0.30		0.325		0.10	0.11	0.12	0.04
	0.40		0.350		0.14	0.14	0.16	0.04
	0.50		0.375		0.19	0.18	0.21	0.06
	0.10	1.0	0.350		0.14	0.16	0.18	0.06
	0.20		0.450		0.36	0.30	0.32	0.11
	0.30		0.550		0.58	0.42	0.45	0.17
	0.40		0.650		0.75	0.54	0.56	0.21
	0.50		0.750		0.86	0.64	0.65	0.26
0.50	0.10	0.5	0.275	61	0.05	0.07	0.07	0.05
	0.20		0.300		0.12	0.15	0.15	0.09
	0.30		0.325		0.21	0.24	0.24	0.17
	0.40		0.350		0.32	0.33	0.33	0.22
	0.50		0.375		0.44	0.43	0.42	0.29
	0.10	1.0	0.350		0.32	0.33	0.32	0.25
	0.20		0.450		0.73	0.56	0.55	0.42
	0.30		0.550		0.93	0.73	0.71	0.60
	0.40		0.650		0.99	0.85	0.82	0.70
	0.50		0.750		1.00	0.92	0.90	0.80
0.75	0.10	0.5	0.275	75	0.08	0.11	0.12	0.13
	0.20		0.300		0.23	0.25	0.26	0.27
	0.30		0.325		0.42	0.39	0.41	0.41
	0.40		0.350		0.60	0.53	0.55	0.56
	0.50		0.375		0.74	0.66	0.68	0.67
	0.10	1.0	0.350		0.60	0.45	0.46	0.47
	0.20		0.450		0.96	0.72	0.73	0.72
	0.30		0.550		1.00	0.87	0.87	0.86
	0.40		0.650		1.00	0.94	0.95	0.94
	0.50		0.750		1.00	0.98	0.98	0.97

Polygenic $h^2 = 0.25$; $r = 0$.

When GE is equal to 0.5 and $E(fhm)$ is low (0.25–0.5) the differences between algebraic methods are negligible and there is evidence that the overestimation of algebraic methods tends to become more important as $E(fhq)$ increases. Algebraic results are more realistic when $E(fhm)$ is 0.75 which corresponds to equal frequencies (0.25) for the four alleles at the marker

locus. The same trends can be pointed out for a GE of 1 σ_p. Nevertheless, in this case $P1$ tends to estimate higher powers than other algebraic methods and the differences between simulations and algebraic methods become very large.

4. DISCUSSION/CONCLUSION

These results showed that important differences exist between power calculated with algebraic approximations and simulating data. Even if the binomial probability that any number of sires out of the total number of sires are jointly heterozygous at both the marker and the QT loci is taken into account, as in the $P3$ method, algebraic approximation cannot always be used to estimate the power of different sire designs for QTL detection when the total experiment size is given. However, even though they overestimate power, $P2$ and $P3$ could be used to rank designs differing in the number of sires when the total size of the experiment is given. On the contrary, it seems to be inadequate not to include the binomial probability and to use the expected number of heterozygous parents also in order to optimize the choice of the number of sires mainly when the total experiment size is given, the gene effect is large and the expected frequencies of heterozygotes at the marker and at the QT loci are close to 0.5.

The same conclusions can be drawn from an analysis carried out considering a diallelic marker locus (unpublished data).

Probably, part of the difference between the algebraic and simulation results can be attributed to assumptions made about the number of informative offspring per sire, the balance between the two offspring sub-groups which receive the same marker allele from the sire, and the distribution of the statistic.

As regards the distribution of the statistic, it should be noted that the use of χ^2 distribution instead of F did not significantly change the algebraic estimates obtained in this work (unpublished data).

All in all, it would be programming and computing costly to consider all eventualities concerning the offspring sub-group sizes using a full algebraic approach. Thus, simulating the data can still be considered in these situations as the most useful tool for estimating the power of QTL detection sire designs.

REFERENCES

[1] Le Roy P., Elsen J.M., Numerical comparison between powers of maximum-likelihood and analysis of variance methods for QTL detection in progeny test designs: the case of monogenic inheritance, Theor. Appl. Genet. 90 (1995) 65–72.

[2] Neimann-Sørensen A., Robertson A., The association between blood groups and several production characteristics in three Danish cattle breeds, Acta Agric. Scand. 11 (1961) 163–196.

[3] Sax K., The association of size differences with seed coat pattern and pigmentation in *Phaesolus vulgarus*, Genetics 8 (1923) 552–560.

[4] Soller M., Genizi A., The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations, Biometrics 34 (1978) 47–55.

[5] van der Beek S., van Arendonk J.A.M., Groen A.F., Power of two- and three-generation QTL mapping experiments in an outbred population containing full-sib or half-sib families, Theor. Appl. Genet. 91 (1995) 1115–1124.

[6] Weller J.L., Kashi Y., Soller M., Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle, J. Dairy Sci. 73 (1990) 2525–2537.