Optimal design for the detection of a major gene segregation in crosses between 2 pure lines

Jean Michel J. M. Elsen, Pascale P. Le Roy

Genet Sel Evol (1995) 27, 275-285

Original article

Optimal design for the detection of a major gene segregation in crosses between 2 pure lines

JM Elsen [1], P Le Roy [2]

[1] *Institut national de la recherche agronomique, station d'amélioration génétique des animaux, BP27, 31326 Castanet-Tolosan cedex;*
[2] *Institut national de la recherche agronomique, station de génétique quantitative et appliquée, 78352 Jouy-en-Josas cedex, France*

(Received 15 June 1994; accepted 15 December 1994)

Summary – A simulation method was used to compare different experimental designs for their power to detect a major gene using a maximum likelihood approach. The optimal design is most often the production of F2 as the only segregating genetic type, with a limited effect of the relative numbers of F2s and non-segregating groups (parentals and F1) on the power. Dominant genes were more easily detected than additive ones. A model dealing with the heteroskedasticity of the polygenic component was also studied.

major gene / optimization / maximum likelihood / homozygous line

Résumé – Protocoles optimaux pour la détection d'un gène à effet majeur en ségrégation dans des croisements entre 2 lignées pures. *Différents protocoles expérimentaux ont été comparés par simulation sur leur puissance pour la détection d'un gène à l'aide d'un test du maximum de vraisemblance. Le protocole optimal est le plus souvent celui pour lequel le seul type génétique où le gène est en ségrégation est la F2, avec un faible effet de la proportion de F2 par rapport aux types génétiques sans ségrégation (parentaux et F1). Les gènes dominants sont détectés plus facilement que les gènes additifs. Un modèle considérant l'hétéroscédasticité de la composante polygénique est aussi étudié.*

gène majeur / optimisation / maximum de vraisemblance / lignée homozygote

INTRODUCTION

The genetic maps presently under development will soon be a great help in the detection of quantitative trait loci. Nevertheless, as stated by Goffinet *et al* (1994), evidencing major gene segregation without marker information will remain

important for various reasons: i) genetic maps may not be available for all species; ii) systematic use of molecular markers is very costly; iii) statistical analysis of phenotype distributions is a useful preliminary analysis of available data; and iv) retrospective studies of old experiments without marker information may be valuable.

The basis for population genetics was established by Mendel, who used crosses between pure lines of peas to observe the segregation of genes controlling the colour and appearance of seeds in F2 and backcrosses. Since that time, a number of crosses between homozygous lines and even between heterogeneous subpopulations have been conducted in plants and animals as tests of a major gene segregation between these lines or subpopulations (the parental groups), *eg*, Hanset (1991) and Boujenane *et al* (1991). The subpopulations may often be considered as independent samples (*eg*, Bradford and Famula, 1984; Duchet-Suchaux *et al*, 1992; Loisel *et al*, 1994).

The underlying hypothesis is usually that the parental groups (P1 and P2) are homozygous in opposite states (AA and BB) at a particular locus governing the measured trait. Under this hypothesis, the first cross (F1) is homogeneous with all animals AB; the F2s (crosses between F1 parents) may be AA, AB or BB with probabilities of 1/4, 1/2 and 1/4 respectively; the backcrosses (either BC1, crosses between F1 and P1, or BC2, crosses between F1 and P2) are also heterogeneous AA or AB animals (BC1) and AB or BB animals (BC2) with proportions 1/2, 1/2.

The statistical analysis of the data obtained from these populations was clearly described by Elston and Stewart (1973) and Stewart and Elston (1973). They showed how a maximum likelihood approach could be used to test various genetic hypotheses differing in gene numbers and types (additive/dominant, autosomal/sex-linked). Alternative methods were described by Mode and Gasser (1972) and Weber (1959). The power of this type of experiment has been recently investigated by Janss and Van der Werf (1992), limiting their study to the case of F2 populations.

In this paper, we describe a study of the optimal structure of the population defined by the relative and absolute numbers of subgroups (P1, P2, F1, F2, BC1 and BC2). Different structures were compared using simulations and their power to detect a major gene in a maximum likelihood approach was investigated. Some information about a more robust model is also provided. The use of simulations for the evaluation of the statistical properties of the likelihood ratio test is justified by the non-observation of classical asymptotic distributions in the particular context studied (Goffinet *et al*, 1992; Loisel *et al*, 1994).

METHODS

Model

Two hypotheses were compared. H_0 assumes that the difference between the parental lines P1 and P2 is due to a large number of genes, each with a small effect in controlling the trait measured, and H_1 assumes that beyond this polygenic difference, a major gene is fixed at opposite homozygous states (AA and BB) in the parental lines.

Y_{ij} is the performance of the jth individual of the ith genetic type. Six genetic types are considered (P1, P2, F1, F2, BC1, BC2) with $i = 1$ to 6 respectively. The number of individuals in the ith group is n_i.

Under H_0, the performance Y_{ij} was modeled as:

$$Y_{ij} = \mu + l_i + e_{ij}$$

where μ is the general mean and l_i the genetic type i effect which can be detailed using Dickerson's crossbreeding parameters (Dickerson, 1973). In this study, the only parameters considered were the direct individual additive effects (r and s for the parental populations P1 and P2 respectively) and the direct heterosis effect (h):

$$\begin{aligned} l_1 &= r \\ l_2 &= s \\ l_3 &= (r+s)/2 + h \\ l_4 &= (r+s)/2 + h/2 \\ l_5 &= (3r+s)/4 + h/2 \\ l_6 &= (r+3s)/4 + h/2 \end{aligned}$$

e_{ij} is the residual effect which is normally distributed $\mathcal{N}(0, \sigma_e^2)$.

Under H_1, the performance Y_{ij} is modeled as:

$$Y_{ij} = \mu + l_i + g_k + e_{ij} \quad \text{with probability} \quad p_{ik}$$

where g_k is the major genotype k effect ($k = 1$ for AA, 2 for AB and 3 for BB) and p_{ik} is the probability of the kth genotype in the ith genetic type.

Under the preceding fixed alleles hypothesis:

i	Group	p_{i1}	p_{i2}	p_{i3}
1	P1	1	0	0
2	P2	0	0	1
3	F1	0	1	0
4	F2	1/4	1/2	1/4
5	BC1	1/2	1/2	0
6	BC2	0	1/2	1/2

The case where the within-major-genotype variance varies between groups may be studied simply by replacing σ_e^2 with σ_{ei}^2. In our simulations, this has been explored for a limited range of population structures.

Test statistic

The hypothesis H_0 was tested using the likelihood ratio test $\mathcal{L} = -2\ln(L_0/L_1)$ where:

$$L_0 = \prod_{i=1}^{6}\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}\,\sigma_e} \exp\left\{-\frac{1}{2}\left(\frac{y_{ij}-\mu-l_i}{\sigma_e}\right)^2\right\}$$

$$L_1 = \prod_{i=1}^{6}\prod_{j=1}^{n_i} \left[\sum_{k=1}^{3} p_{ik} \frac{1}{\sqrt{2\pi}\,\sigma_e} \exp\left\{-\frac{1}{2}\left(\frac{y_{ij}-\mu-l_i-g_k}{\sigma_e}\right)^2\right\}\right]$$

It must be emphasized that, in this model, no familial relationships are considered between the measured individuals.

The H_0 hypothesis (no major gene segregating in F2s and/or backcrosses) was rejected if the test statistic $\mathcal{L}$ exceeded a threshold λ. Due to non-observation of regulatory conditions, the asymptotic distribution of $\mathcal{L}$ under H_0 is probably not the classical χ^2 with a number of degrees of freedom equal to the difference between the number of parameters to be estimated under H_1 and H_0 (Goffinet *et al*, 1992; Jans and Van der Werf, 1992). Moreover, for a limited number of individuals, the true asymptotic distribution may not be attained. To cope with these difficulties, empirical rejection thresholds were obtained from simulations.

Cases studied

First, the power was evaluated for different population structures, given a total number of 180 individuals measured. These situations are given in table I. In all cases, P1, P2 and F1 were in equal proportions. In the C1 cases, the backcrosses were not produced and the segregation of the major gene was visible only in the F2. In the C2 cases, the F2 was absent and the 2 backcrosses were present in equal proportions. The C3, C4 and C5 cases described the situations where both F2 and backcrosses were present. The proportion t of individuals belonging to the 'segregating groups' increased between C10 and C19, C20 and C26, and C3 and C5. The proportion of F2s to backcrosses increased between C30 and C35, C40 and C44, and C50 and C54. The major gene was characterized for each of these cases by an effect of 2 residual standard deviations between the means of homozygotes, either additive ($g_1 = 0$, $g_2 = 1$ and $g_3 = 2$, *ie*, $a = (g_3 - g_1)/2 = 1$) or dominant ($g_1 = g_2 = 0$ and $g_3 = 2$, *ie* $d = g_2 - (g_1 + g_3)/2 = -1$).

Secondly, the effects of the whole population size ($\sum_i n_i = 30$ to 480 individuals) and of the major gene effect (4 values for a between 0.25 and $1\sigma_e$, and $d = 0$ or $-a$) were evaluated in the case where half of the population was made up of F2 individuals. The other half was equally divided between P1, P2 and F1 individuals.

Finally, considering these types of major genes, the likelihood was modified to consider the case where the within-group variance differs between the F2 (σ^2_{F2}) and the non-segregating subpopulations (σ^2_{NS}). Simulations were performed considering $\sigma^2_{F2} = 1$ and $\sigma^2_{NS} = \sigma^2_{F2}$, $\sigma^2_{F2}/1.25$ or $\sigma^2_{F2}/1.5$, for the structures C10 to C19 and their equivalent with the total number of measured individuals doubled.

Table I. Population structures studied for 2 standard deviations gene effects in the homoskedastic case (number of individuals per genetic type).

Case	*P1*	*P2*	*F1*	*F2*	*BC1*	*BC2*
C10	45	45	45	45	0	0
C11	40	40	40	60	0	0
C12	35	35	35	75	0	0
C13	30	30	30	90	0	0
C14	25	25	25	105	0	0
C15	20	20	20	120	0	0
C16	15	15	15	135	0	0
C17	10	10	10	150	0	0
C18	5	5	5	165	0	0
C19	0	0	0	180	0	0
C20	36	36	36	0	36	36
C21	32	32	32	0	42	42
C22	28	28	28	0	48	48
C23	24	24	24	0	54	54
C24	20	20	20	0	60	60
C25	16	16	16	0	66	66
C26	12	12	12	0	72	72
C30	30	30	30	20	35	35
C31	30	30	30	30	30	30
C32	30	30	30	40	25	25
C33	30	30	30	50	20	20
C34	30	30	30	60	15	15
C40	20	20	20	40	40	40
C41	20	20	20	50	35	35
C42	20	20	20	60	30	30
C43	20	20	20	70	25	25
C44	20	20	20	80	20	20
C50	10	10	10	50	50	50
C51	10	10	10	60	45	45
C52	10	10	10	70	40	40
C53	10	10	10	80	35	35
C54	10	10	10	90	30	30
C55	10	10	10	100	25	25

Numerical techniques

The results were obtained from simulations. Appropriate subroutines from the NAG library were used for the generation of genotypes and normal values (G05CCF, G05DDF, G05CAF). The maximization of the likelihood was performed using a quasi-Newton algorithm (E04JBF from the NAG Library). Only 1 starting point was tested for each maximization.

The rejection thresholds under H_0 were estimated from the 10% empirical quantiles of the test statistic distribution, for each population structure studied,

defined by the group sizes n_i. The power at the 10% level was simply estimated for each case studied by taking the number of test statistic values that exceeded the corresponding H_0 quantile. Two thousand simulations were performed in each of the H_0 and H_1 cases.

RESULTS AND DISCUSSION

Optimal structure under the homoskedastic model

Figure 1 gives the power of situations C1 and C2 as a function of the ratio t of the segregating population (F2 or the 2 backcrosses) size to the total population size. Whereas the 2 types of designs (F2 or BC alone) give a similar power for a dominant gene, the F2 must be used in the case of an additive gene, with a power varying between 60 and 70% against 30 to 40% for the backcross. In the C1 situations the maximum power is always reached for an equal proportion of segregating ($n_4 = 90$) and non-segregating populations ($n_1 = n_2 = n_3 = 30$), *ie* with a t ratio of 1/2. In contrast, in the C2 situations, this optimal proportion seems to differ according to whether a dominant (where the optimum is about 3 times more in backcross individuals than in non-segregating individuals) or an additive gene (the maximum power being attained with the minimum number of backcross individuals studied) is considered.

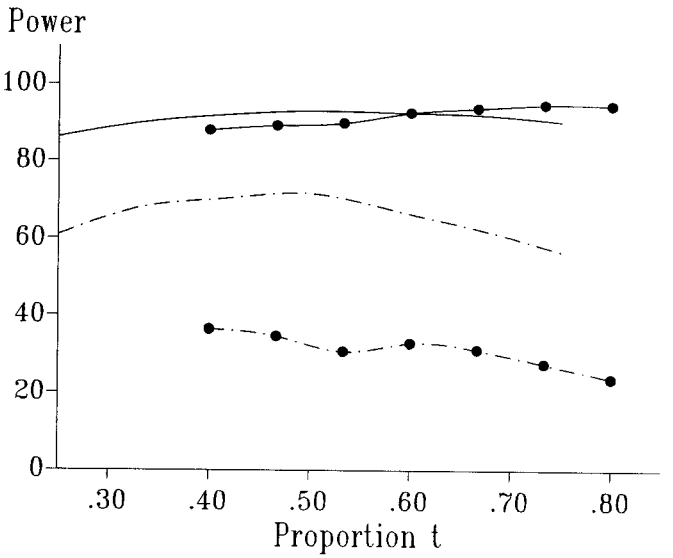

Fig 1. Power of the design with only 1 type of segregating population (F2 or BC) for various proportions t of the segregating group (a total of 180 individuals measured, gene effect $a = 1$ sd). F2 ($d = 0$: — –; or $d = -a$: —); BC ($d = 0$: —•–; or $d = -a$: —•).

Figure 2 describes the case where the F2 and backcross groups were both produced (C3, C4 and C5). The power is given as a function of the ratio u of

the number of F2s to the number of F2 + backcross individuals, for the 3 situations considered with respect to the t parameter: 1/2 (C3 cases, $n_1 = n_2 = n_3 = 30$), 2/3 (C4 cases, $n_1 = n_2 = n_3 = 20$) and 5/6 (C5 cases, $n_1 = n_2 = n_3 = 10$). The power appeared to be very insensitive to the ratio u for a dominant gene and when considering an additive gene with a small number of parental individuals ($t = 5/6$). In situations with an additive gene with a larger proportion of parental individuals ($t = 1/2$ or 2/3), the maximum power was attained by maximising the proportion of F2s.

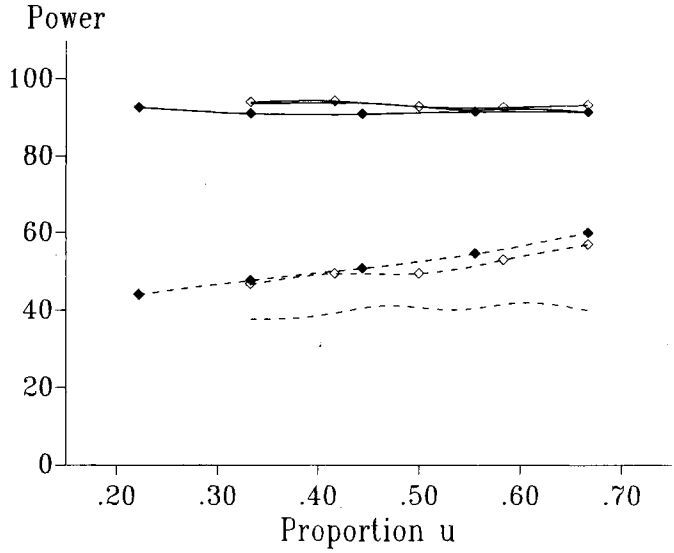

Fig 2. Power of the design with 2 types of segregating population in proportion t for various proportions u of F2s in the segregating group (a total of 180 individuals measured, gene effect $a = 1$ sd). $t = 1/2$ ($d = 0$: -♦-; or $d = -a$: -♦-); $t = 2/3$ ($d = 0$: - ◇ -; or $d = -a$: -◇-); $t = 5/6$ ($d = 0$: - - -; or $d = -a$: —).

Evidence for a major gene comes from the detection of a mixture of subdistributions within the global distribution of either F2 and/or backcrosses. In principle, the test statistic used (the likelihood ratio test) makes use of the whole non-normality of the global distribution. This non-normality is greater when the means of the subdistributions are more extreme. This phenomenon probably explains the lack of power of the backcross cases as compared to the F2 cases when an additive gene was studied. In this situation, the difference between distribution components means of the global F2 distribution was twice as a high as the difference in either the BC1 or the BC2.

When a hypothesis can be made about the type of dominance, before the experiment is designed, then maximum power will be attained by limiting the segregating subpopulation to the single backcross showing segregation. However, the power of such a design will be zero if the true dominance is in the opposite direction. Table II compares the power of this design with the power of an F2 when

a total of 180 individuals were measured, half of which were in the non-segregating (P1, P2 and F1) populations.

All these results may also be directly related to the proportion of the variance of the trait due to the major gene in the segregating groups (table III); this proportion increases with the differences between subdistributions means.

Table II. Power of designs with either an F2 or a BC1 as the only segregating group (180 individuals; half in the F2 or BC1).

Type of gene action			*F2*	*BC1*
0	0	1	34.2	9.5
0	0.5	1	26.8	14.7
0	1	1	34.2	37.9
0	0	2	93.6	7.8
0	1	2	70.0	37.5
0	2	2	93.6	98.0

Table III. Proportion of the variance of the trait due to the major gene in the F2 and BC1 (in $\sigma_e/64$).

Type of gene action			*F2*	*BC1*
0	0	0.5	3	0
0	0.25	0.5	2	1
0	0.5	0.5	3	4
0	0	1	12	0
0	0.5	1	8	4
0	1	1	12	16
0	0	1.5	27	0
0	0.75	1.5	18	9
0	1.5	1.5	27	36
0	0	2	48	0
0	1	2	32	16
0	2	2	48	64

Size of the design

The minimum number of individuals to be measured in order to have a 90% power for the detection of a gene effect $a = 1$ standard deviation is 150 when considering a dominant gene ($d = -a$) and about 500 when considering an additive gene ($d = 0$) (fig 3). Larger populations are required for smaller gene effects. The changes in curve shape with the gene effect a must be emphasized. These curves are nearly linear for power under 70% and, in this linear part, the slope (*ie* the gain in power

per extra individual measured) increases with a. The resulting increase in size of the design required for a 70% power does not appear to be linear in $1/a$.

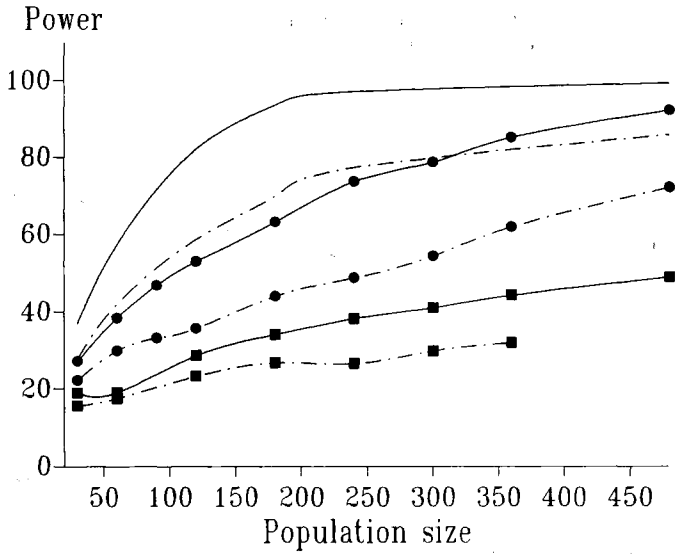

Fig 3. Power of the design for various population sizes and gene effects (P1, P2, F1 and F2 in proportions 1/6, 1/6, 1/6 and 1/2). $a = 0.5$ ($d = 0$: -■-; or $d = -a$: -■-); $a = 0.75$ ($d = 0$: -●-; or $d = -a$: -●-); $a = 1$ ($d = 0$: — -; or $d = -a$: —).

Janss and Van der Werf (1992) considered a 1 standard deviation additive gene effect ($a = 1$) and a 5% significance level and found a 12% power when only F2 individuals were measured (1 000 individuals) but a 100% power when 500 F1s were added to these 1 000 F2s. From our simulations, the further inclusion of parental P1 and P2 performances in the analyses appears to be extremely useful. We confirmed these results at the 10% level with some simulations performed with F2 individuals only. The power of detecting an additive 2 standard deviations gene with 1 000 F2s reached only 24%, a value attained with only 30 individuals when the parental subgroups were included.

Robustness to heteroskedasticity

Janss and Van der Werf (1992) argued that the inclusion of F1 data decreases the robustness of the analysis, a false major gene being easily detected when the F2 group variance is higher than in the F1 population (100% false detection with a 50% variance increase). As described above, this heteroskedasticity can be included in the model without difficulty.

Figure 4 shows the power of such a heteroskedastic model for various population sizes, when the performances are simulated with $\sigma^2_{F2} = 1.5\sigma^2_{NS}$. Additive and dominant genes of a 1 standard deviation effect were considered. The results obtained with $\sigma^2_{F2} = 1.25\sigma^2_{NS}$ and $\sigma^2_{F2} = \sigma^2_{NS}$ were very similar. The detection

power for additive genes was low and nearly independent of the population size and structure. In contrast, in the case of a dominant gene, the power increased strongly with population size and reached its maximum when all individuals belonged to the F2 population, which is the opposite of the homoskedastic case where the non-segregating populations were useful.

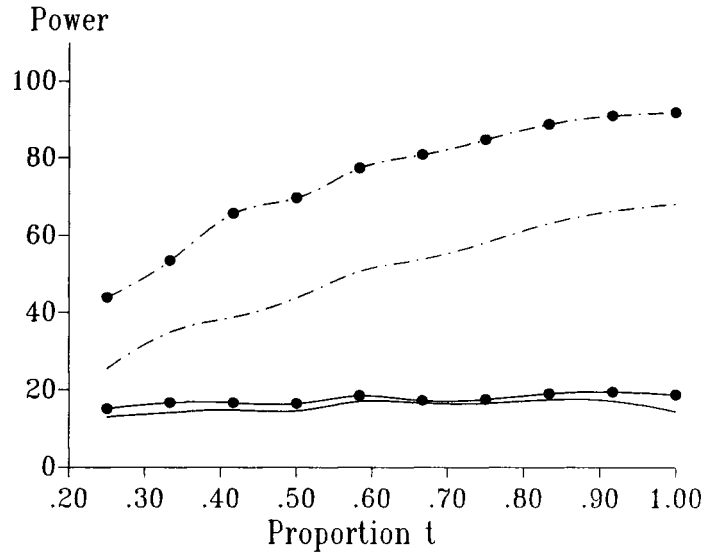

Fig 4. Power of the design with various proportions t of F2 individuals in the heteroskedastic model (a total of 180 or 360 individuals measured, gene effect $a = 1$ sd). $n = 180$ ($d = 0$: —; or $d = -a$: — -); $n = 360$ ($d = 0$: -●-; or $d = -a$: -● -).

This result shows that the information in the non-segregating population derives from the level of the within-group variance. This variance for the F2 can be estimated in the parental and F1 groups in the homoskedastic model, but not in the heteroskedastic model. In the latter, the major gene segregation was only tested through the non-normality of the F2 group, while in the previous model the increase of variance between F1 and F2 also contributed to this testing.

CONCLUSION

In general, the generation of backcrosses does not compete with the production of F2s alone as a segregating population. This is particularly true for an additive gene. The power of the detection test seems to be poorly sensitive to the proportion of F2s in the whole population. The optimum appears to be 50% of F2s with equal proportions of P1, P2 and F1. Large dominant genes are easily detected in such small populations (fewer than 200 individuals for a 2 standard deviations gene effect). Additive genes are less easily detected.

These results were obtained by comparing mixed with polygenic inheritance in the homoskedastic case. To prevent a lack of robustness due to heteroskedasticity,

a model including variance differences between F2s and parental populations may be used. In this case, the major gene is detected through the non-normality of the F2, with a loss of power. Another extreme situation may be found if the differences between genetic types are due only to the segregation at the major locus. Comparing this monogenic hypothesis to the polygenic one causes difficulty since these hypotheses are not nested. This may be solved simulating empirical quantiles as done in this study or using the Akaike (1973) criteria.

REFERENCES

Akaike H (1973) Information theory and an extension of ML principle. *In: 2nd International Symposium on Information theory* (BN Petrov, F Csahi, eds), Ak Kiado, Budapest, Hungary, 267-281

Boujenane I, Bradford GE, Famula TR (1991) Inheritance of litter size and its components in crosses between D'Man and Sardi breeds of sheep. *J Anim Sci* 69, 517-524

Bradford GE, Famula TR (1984) Evidence for a major gene for rapid postweaning growth in mice. *Genet Res Camb* 44, 293-308

Dickerson GE (1973) Inbreeding and heterosis in animals. *In: Proc Anim Breeding Genet Symp in honour of Dr JL Lush*, Am Soc Anim Sci Assoc, Champaign, IL, USA, 54-77

Duchet-Suchaux M, Menanteau P, Le Roux H, Elsen JM, Lechopier P (1992) Genetic control of resistance to enterotoxigenic *Escherichia coli* in infant mice. *Microbiol Pathogenesis* 13, 157-160

Elston RC, Stewart J (1973) The analysis of quantitative traits for simple models from parental, F1 and backcross data. *Genetics* 73, 695-711

Goffinet B, Loisel P, Laurent B (1992) Testing in normal mixture models when the proportions are known. *Biometrika* 79, 842-846

Goffinet B, Bekmann J, Boichard D *et al* (1994) Méthodes mathématiques pour l'étude des gènes contrôlant des caractères quantitatifs. *Genet Sel Evol* 26, 9s-20s

Hanset R (1991) The major gene of muscular hypertrophy in the Belgium Blue cattle breed. *In: Breeding for Disease Resistance in the Farm Animal* (JB Owen, RFE Axford, Wallingford, eds), CAB International, UK, 467-478

Janss LLG, Van der Werf JHJ (1992) Identification of a major gene in F1 and F2 data when alleles are fixed in the parental lines. *Genet Sel Evol* 24, 511-526

Loisel P, Goffinet B, Monod H, Montes de Oca G (1994) Detecting a major gene in an F2 population. *Biometrics* 50, 512-516

Mode CJ, Gasser DL (1972) A distribution free test for major gene differences in quantitative inheritance. *Math Biosci* 14, 143-150

Stewart J, Elston RC (1973) Biometrical genetics with 1 or 2 loci: the inheritance of physiological characters in mice. *Genetics* 73, 675-693

Weber E (1959) The genetical analysis of characters with continuous variability on a mendelian basis. I. Monohybrid segregation. *Genetics* 44, 1131-1139