ANIMAL BIOTECHNOLOGY, 10(1&2), 25-35 (1999)

STANDARDIZATION AND CONVERSION OF MARKER POLYMORPHISM MEASURES

by
Y. Da[1], P.M. VanRaden[2], M. Ron[3], J.E. Beever[4], A. A. Paszek[1], J. Song[3], G. R. Wiggans[2], R. Ma[4], J.I. Weller[3], and H.A. Lewin[4]
Program in Comparative Genomics, Department of Veterinary PathoBiology, University of Minnesota, Saint Paul, MN 55108 USA

ABSTRACT

Large scale gene mapping efforts in domestic animals have generated and mapped a large number of genetic markers that are useful for mapping quantitative trait and disease loci and for DNA diagnostic purposes such as parentage testing. Marker polymorphism is an important criterion for selecting genetic markers in planning experiment for mapping quantitative trait loci or for DNA diagnostic purposes. Current formulations of marker polymorphism measures are functions of marker allele frequencies. In this study, two measures of marker polymorphism that are available from gene mapping studies and do not require allele frequencies were proposed and analyzed: the observed polymorphic information content (PIC) and the observed family information content (FIC). The observed FIC was more stable than the observed PIC because the observed FIC is unaffected by the variation in the frequency of heterozygous parents. However, both FIC and PIC are dependent on the gene mapping design. The effective number of alleles is recommended as a tool to standardize marker polymorphism measures so that polymorphism of different markers can be compared on an qual basis, and to obtain a new polymorphism measure (such an exclusion probability) from an existing measure (such as FIC). The usage of the effective number of alleles to standardize FIC, PIC and exclusion probabilities is illustrated using genetic markers in a published linkage map.

(*Keywords*: Genetic markers, polymorphism information content, family information content, effective number of alleles)

INTRODUCTION

The amount of genotype data in domestic animals is growing rapidly after years of large scale gene mapping efforts (Barendse et al., 1997; Crawford et al., 1994; Ma et al., 1996; Rohrer et al. 1996).

[1] Address correspondence to: Dr. Yang Da, 301B Veterinary Science, Dept. of Veterinary PathoBiology, University of Minnesota, 1971 Commonwealth Ave., St. Paul, MN 55108 USA, y-da@tc.umn.edu
[2] Animal Improvement Programs Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, Maryland 20705, USA
[3] Agricultural Research Organization, The Volcani Center, Bet Dagen 50250, Israel
[4] Laboratory of Immunogenetics, Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

Mapped genetic markers have important applications in mapping quantitative trait loci (QTL), parentage testing and linkage studies. Since genotyping of DNA-level genetic markers is relatively expensive, most studies have considered the number of individuals genotyped to be the limiting factor (Weller et al., 1990; Brascamp et al., 1993). Marker polymorphism affects the sample size required for QTL mapping and the number of markers required for parentage testing. Therefore, marker polymorphism is an important criterion for selecting genetic markers. For linkage and QTL studies, marker polymorphism refers to the expected frequency of genotyped individuals that unequivocally identify the allele transmission from the parent to the offspring. Several measures to predict marker polymorphism are available (Botstein et al., 1980; Da & Lewin, 1995; Ron et al., 1993) and the predicted polymorphism was found consistent with the observed polymorphism (Ron et al., 1995). However, these predicted measures require knowledge of allele frequencies that may not have reliable estimates from gene mapping studies. In contrast to the predicted polymorphism, observed marker polymorphism is available from any gene mapping design and does not require knowledge of allele frequencies, but the observed marker polymorphism is dependent on gene mapping designs. For example, marker polymorphism derived from a full-sib design is not comparable to that of a half-sib design. The full-sib design (Barendse et al., 1997; Rohrer et al, 1996) yields more marker meiosis information than the half-sib design (Crawford et al., 1994; Ma et al., 1996), because two parents are genotyped per family in the full-sib design whereas the half-sib design typically genotypes one parent per family (Da and Lewin, 1995). For parentage testing, marker polymorphism is measured by exclusion probabilities (Jamieson 1965, 1994; Garber and Morris 1983; Weir 1990). However, exclusion probabilities also require knowledge of allele frequencies and are not observable from gene mapping studies. The purpose of this study was to provide a general and readily available approach to standardize measures of marker polymorphism for different designs so that marker polymorphism measures can be evaluated and compared accurately, and to derive alternative formulations of polymorphism measures (including exclusion probabilities) using marker information present in gene mapping studies.

Standardization and Conversion of Polymorphism Measures

Marker polymorphism measures: Marker polymorphism for linkage and QTL mapping research can be reported at two levels, the sample level, which is the marker polymorphism for all families including

families with homozygous parents, and the family level, which is the average polymorphism per heterozygous parent. Each level can be reported in two ways, the expected and observed frequency of informative offspring. The sample level polymorphism can be predicted by the expected sample frequency of informative offspring, better known as the polymorphism information content (PIC; Botstein et al., 1980), and the family level polymorphism can be predicted by the frequency of informative offspring per heterozygous parent, to be referred to as family information content (FIC), in parallel with PIC. The expected FIC for genotyping two parents (i_2, Da and Lewin, 1995) and for genotyping one parent (i_1; Ron et al., 1993) are given by the following formulations (for a proof of both formulae, see Table 1 in Da and Lewin, 1995):

$$i_2 = 1 - p_u p_v \quad (1)$$

$$i_1 = 1 - 2(p_u + p_v) \quad (2)$$

where p_u and p_v are allele frequencies of marker alleles u and v of the heterozygous parent.

In gene mapping studies where a mixed genotyping scheme is present, i.e., some offspring have two genotyped parents and some have one genotyped parent, the FIC can be expressed as:

$$i_m = (1 - \alpha)i_1 + \alpha i_2 \quad (3)$$

where α = proportion of offspring with two genotyped parents. Note that $0 \leq \leq 1$, $i_m = i_1$ if $= 0$, $i_m = i_2$ if $= 1$, and $i_1 \leq i_m \leq i_2$.

According to the definition of FIC, the observed FIC for a heterozygous parent (i_{ojk}) is simply the number of informative offspring of the heterozygous parent divided by the number of offspring in the family, i.e., $i_{ojk} = m_{jk}/M_k$, where subscript 'o' denotes 'observed', j = number of genotyped parents (j = 1 or 2), m_{jk} = the number of informative offspring for heterozygous parent k when j parents in the family are genotyped (j = 1 or 2), M_k = the number of offspring in family k. Using informative offspring from all heterozygous families, the FIC for a design is the number of informative offspring divided by the total number of offspring of all heterozygous parents (fathers or mothers), i.e.,

$$FIC_o = i_{oj} = \sum_{k=1}^{s} m_{jk} \; /M = \sum_{k=1}^{s} \frac{M_k}{M} \frac{m_{jk}}{M_k} = \sum_{k=1}^{s} c_k \, i_{ojk} \quad (4)$$

where s = number of heterozygous families, M = total number of offspring in the s heterozygous families, and $c_k = M_k/M$.

Therefore, the observed FIC given by equation (4) is a weighted average of the observed FIC for each family with c_k as the weight. The observed FIC from a mixed genotype scheme (i_{om}) is calculated in the same way as given in equation (4) except that the interpretation and the underlying genotyping scheme are different.

The predicted PIC for genotyping two parents (I_2; Botstein et al., 1980) and for genotyping one parent (I_1; Da and Lewin 1995) are given by:

$$I_2 = 2\sum_{u=1}^{n-1}\sum_{v=u+1}^{n} p_u\, p_v - 2\sum_{u=1}^{n-1}\sum_{v=u+1}^{n} p_u^2\, p_v^2 \qquad (5)$$

$$I_1 = 2\sum_{u=1}^{n-1}\sum_{v=u+1}^{n} p_v - \sum_{u=1}^{n-1}\sum_{v=u+1}^{n} p_v(p_u + p_v) \qquad (6)$$

where n = number of alleles.

By the definition of PIC, the observed PIC for each design is the number of informative offspring divided by the total number of offspring in all families (N), i.e.,

$$PIC_o = I_{oj} = \sum_{k=1}^{s} m_{jk} \;/N = \frac{M}{N}(i_{oj}) \qquad (7)$$

Comparing equation (7) with equation (4), the observed PIC is different from FIC by only a factor of M/N, which is the ratio of the number of offspring in heterozygous families to the total number of offspring and is an estimate of heterozygosity for equal family size. Obviously, M/N is affected by variations in the number of heterozygous parents and in the sizes of families with heterozygous parents. Consequently, the observed PIC is affected by the same factors that affect M/N whereas FIC is not affected by those factors and is more stable than the observed PIC. Because of this advantage, the observed FIC will be selected as a basis to report and standardize information measures, and to derive PICs and exclusion probabilities.

Effective number of alleles Effective number of alleles will be defined as the required number of alleles with equal allele frequency to achieve the observed or predicted polymorphism, which can be PIC, FIC, or heterozygosity. For reasons discussed in the previous section, only the observed FIC will be used to calculate the effective number of alleles. For equal allele frequencies, equations 1-3 reduce to $i_2 = 1 - 1/n^2$, $i_1 = 1 - 1/n$, and $i_m = 1 - (1-\alpha)/n - \alpha/n^2$, respectively. Let n_{e2} = effective number of alleles derived from FIC when two parents are genotyped, n_{e1} = effective number of alleles derived from FIC when one parent is genotyped, n_{em} = effective number of alleles derived from FIC with mixed genotyping. Then, replacing i_2, i_1, and i_m by i_{o2}, i_{o1}, and i_{om} respectively and solving for n yields the respective effective number of alleles.

$$n_{e2} = 1/(1 - i_{o2})^2 \tag{8}$$

$$n_{e1} = 1/(1 - i_{o1}) \tag{9}$$

$$n_{em} = \frac{1}{2}\left[\frac{1 - \alpha}{1 - i_{om}} + \sqrt{\left(\frac{1 - \alpha}{1 - i_{om}}\right)^2 + \frac{4\alpha}{1 - i_{om}}} \right] \tag{10}$$

The effective number of alleles has three applications. First, the effective number of alleles is an indication of the variation in allele frequencies and can detect overestimates of marker polymorphism. If the effective number of alleles is much smaller than the observed number of alleles, then the actual allele frequencies are unequal and have large variations. If the effective number of alleles is larger than the observed number of alleles, then the observed marker polymorphism overestimates the marker polymorphism because it exceeds the maximum expected polymorphism. This detection of overestimate is important because the sample size required for a design based on overestimated polymorphism is likely to be insufficient. In such cases, the effective number of alleles should be adjusted downward. A reasonable but not ideal adjustment is to replace the effective number by the observed number. This adjustment is a reasonable adjustment because it is using the only additional information available, it is not ideal because the resulting estimate is the maximum expected polymorphism and may still be an overestimate. The second application is to standardize

polymorphism measures to a common basis, and the third application is to convert one polymorphism measure into another, as to be described below.

Standardization: Observed marker polymorphism from different genotyping schemes should be standardized to the same genotyping scheme so the polymorphism of different markers can be compared on an equal basis. The standardization can be done using the effective number of alleles. The following formulations standardize an observed FIC from genotyping two parents into an FIC genotyping one parent, or vice versa:

$$i_{2s} = 1 - \frac{1}{n_{e1}^2} = 1 - (1 - i_{o1})^2 \tag{11}$$

The standardization of FIC from a mixed genotyping scheme into i_{o2}, i_{oi} can be done using the following

$$i_{1s} = 1 - \frac{1}{n_{e2}} = 1 - \sqrt{1 - i_{o2}} \tag{12}$$

formulations:

$$i_{2s} = 1 - 1/n_{em}^2 \tag{13}$$

$$i_{1s} = 1 - 1/n_{em} \tag{14}$$

Conversion of polymorphism measures: Using the effective number of alleles as a vehicle, one polymorphism measure can be converted into another. The following formulations provide estimates of heterozygosity (h), the PICs and exclusion probabilities based on the effective number of alleles:

$$h = 1 - 1/n_{ek} \tag{15}$$

$$I_{2s} = (1 - \frac{1}{n_{ek}})(1 - \frac{1}{n_{ek}^2}) \tag{16}$$

$$I_{1s} = (1 - \frac{1}{n_{ek}})^2 \tag{17}$$

$$E_1 = \frac{(n_{ek} - 1)(n_{ek}^3 - n_{ek}^2 - 2n_{ek} + 3)}{n_{ek}^4} \tag{18}$$

$$E_0 = \frac{(n_{ek} - 1)(n_{ek}^2 - 3n_{ek} + 3)}{n_{ek}^3} \tag{19}$$

where I_{2s} and I_{1s} are derived from equations 5-6 respectively assuming equal allele frequency, n_{ek} = the effective number of alleles defined by equations 8-10 (k = 1, 2, or m), E_1 = exclusion probability with one confirmed parent (Jamieson, 1965, 1995; Garber and Morris, 1983; Weir, 1990), and E_0 = exclusion probability without a confirmed parent (Garber and Morris 1983). Equations (15-19) can be considered as standardized polymorphism measures because the same measures of different markers are comparable. The difference between I_{2s} and I_{1s} given by equations (16-17) and I_2 and I_1 given by equations (5-6) is that I_{2s} and I_{1s} do not require allele frequencies whereas I_2 and I_1 is dependent on allele frequencies.

Note that marker heterozygosity (h) given by equation 15 is the same as equations 12 and 14. Therefore, the standardized i_1 can be interpreted as an estimate of marker heterozygosity. This interpretation, however, must be limited to the case when both measures are expressed in terms of the effective number of alleles. For example, the predicted h and i_1 should not be interpreted as the same quantity because they have different formulations and measure different information.

EXAMPLE AND DISCUSSION

To illustrate the application of the approach suggested in this study, we analyze an example of six genetic markers from a male-specific linkage map using cattle half-sib families (Ma et al., 1996). In Table 1, two parents were genotyped (α = 1.0) for markers FGR and LMP2, one parent was genotyped (α = 0) for MGTG4B and URB026, while F13A (α = 0.84) and BoLA-A (α = 0.75) had a mixed genotyping scheme. The observed FIC (i_o) and PIC (I_o) values were calculated according to "rules to identify informative genotypes" in Da and Lewin (1995). Because of the different genotyping schemes, the observed FIC and PIC for different markers are not comparable and they do not rank the marker polymorphism correctly. According to the observed FIC, FGR is slightly more informative than URB026 (0.72 versus 0.71). Intuitively, this may not be correct because FGR has only two alleles whereas URB026 has five alleles. If the observed FICs are standardized to the same genotyping scheme, then the results become reasonable. For genotyping one parent, the standardized FIC for FGR becomes only 0.46, compared to 0.71 for URB026. If two parents are genotyped, the standardized FIC for URB026 becomes 0.91, higher than the observed FIC of 0.72 for FGR. Therefore, the standardized FICs show that URB026

Table 1. Marker polymorphism measures for genotyping two parents (α=1), one parent (α =1), and mixed genotyping scheme ($0<\alpha<1$): observed number of alleles (n), effective number of alleles (n_e), observed family information content (i_o), standardized FIC for genotyping two parents (i_{2s}) or one parent (i_{1s}), observed PIC (I_o), estimated PIC for genotyping two parents (I_{2s}) or one parent (I_{1s}), and estimated exclusion probability with one confirmed parent (E_1) or no confirmed parent (E_0)[a]

Locus		n	n_e	FIC			PIC			Exclusion probability	
				i_o	i_{2s}	i_{1s}[d]	I_o	I_{2s}	I_{1s}	E_1	E_0
FGR	1.0	2	1.88	0.72	0.72	0.46	0.25	0.33	0.21	0.17	0.12
LMP2	1.0	2	1.69	0.65	0.65	0.40	0.13	0.26	0.16	0.13	0.11
F13A	0.84	2	2.37	0.78	0.82/0.75	0.58/0.50	0.09	0.38	0.25	0.19	0.13
BoLA-A	0.75	28	5.74	0.93	0.97	0.83	0.32	0.77	0.67	0.79	0.47
MGTG4B	0	11	2.72	0.63	0.86	0.63	0.56	0.55	0.40	0.32	0.19
URB026	0	5	3.40	0.71	0.91	0.71	0.51	0.65	0.50	0.43	0.27

[a] i_o was calculated using by equation (7), n_e by equations 8-10, i_{2s} by equation 11 or 13, i_{1s} (= standardized heterozygosity) by equation 12 or 14, I_o by equation (7), I_{2s}, I_{1s}, E_1 and E_0 by equations 16-19 respectively. Data were adopted from Ma et al., 1996.

[b] Calculated using the effective number of alleles.

[c] Calculated using the observed number of alleles, a downward adjustment for the overestimate of effective number of alleles.

[d] Can be interpreted as an estimate of the marker heterozygosity.

in fact should be more informative than FGR by 0.25 or 0.19 for genotyping one or two parents respectively. The observed PIC has the same problem for markers BoLA-A and MGTG4B and the standardized PIC would solve that problem. The comparison between the observed FIC (i_o) and PIC (l_o) for markers BoLA-A and MGTG4B shows that the observed FIC being more stable than the observed PIC. The observed FIC (0.93) ranked BoLA-A as the most informative whereas the observed PIC (0.32) ranked BoLA-A as the third informative among the six markers. The ranking by the observed PIC is unusual because a marker with 28 alleles generally should be more informative than a marker with five alleles. Close examination of the data showed that BoLA-A had only four (out of nine) heterozygous sires, and three of the four families were among the smallest families. Therefore, the observed PIC for BoLA-A is an underestimate due to the smaller than expected number of heterozygous sires and smaller than average family sizes of the heterozygous families. Marker MGTG4B, on the other hand, seems to have an overestimate of PIC because its observed PIC ranks it as the most informative of the six markers even though it has only 2.72 effective alleles. Data examination showed that MGTG4B had eight (out of nine) heterozygous sires and seven of the eight families were the largest families. The standardized PICs provided reasonable rankings, with BoLA-A ranked number 1 (instead of number 3), and MGTG4B ranked number 3 (instead of number 1). Note that the standardized measures in Table 1 and n_e are independent of genotyping schemes and all yield the same. The effective number of alleles is an indication of the distribution of allele frequencies and can detect overestimates of marker polymorphism. Large discrepancy between the effective and observed numbers of alleles indicates large variations in allele frequencies. For example, the number of observed alleles for BoLA-A is 28 but the effective number of alleles is only 5.74, indicating that a few alleles were predominant and the rest of the alleles each had a small frequency. Similar discussion applies to MGTG4B, which has 11 observed alleles but only 2.72 effective alleles. In contrast, marker F13A has an effective number of alleles that is larger than the observed number of alleles. This indicates that the observed marker polymorphism for marker F13A is more than expected. With the downward adjustment suggested earlier, the FIC of F13A for genotyping two parents should be 0.75 instead of 0.82. Without the comparison between the effective and observed numbers of alleles, the fact that the marker polymorphism for F13A exceeded the expected maximum would not have been detected. The last two columns of Table 1 show the estimation of exclusion

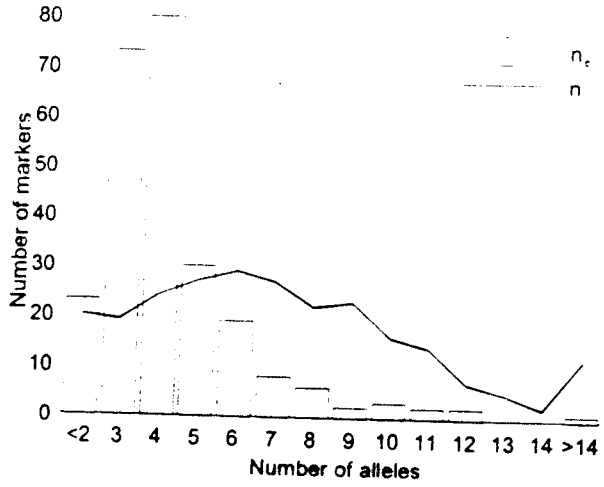

Figure 1. Comparison of marker polymorphism based on effective and observed numbers of alleles. The vertical bar shows the number of markers classified by the effective number of alleles (n_e) and the solid line shows the number of markers classified by the observed number of alleles (n). For the effective number of alleles, "3" means $2 \leq n_e \leq 3$, and the same interpretation applies to categories 4-14 on the X-axis. For the observed number of alleles, the category "<2" should be read as "2". Data shown were 247 microsatellite markers reported in Ma et al., 1996.

probabilities using the effective numbers of alleles. The results show that exclusion probabilities generally are lower than the standardized PICs. Figure 1 shows the comparison between the effective and observed numbers of alleles for the 247 microsatellite markers reported in Ma et al. (1996). The effective number (n_e) shows that most microsatellite markers (61.9%) have n_e values in the range of 2-4 whereas the observed number of alleles had a rather flat distribution which is less related to the marker polymorphism. The effective number of alleles (derived from FIC) is more useful than the observed number of alleles for measuring marker polymorphism, and is a convenient tool to standardize and convert marker polymorphism measures. In contrast, the observed number of alleles is not an appropriate measure of marker polymorphism, but is an indicator of polymorphism potentials.

REFERENCES

Barendse, W., Vaiman, D., Kemp, S.J., Sugimoto, Y., Armitage, S.M., Williams, J.L., Sun, H.S., Eggen, A., Agaba, M., Aleyasin, S.A., Band, M., Bishop, M.D., Buitkamp, J., Byrne, K., Collins, F., Cooper, L., Coppettiers, W., Denys, B., Drinkwater, R.D., Easterday, K., Elduque,

C., Ennis, S., Erhardt, G., Ferretti, L., Flavin, N., Gao, Q., Georges, M., Gurung, R., Harlizius, B., Hawkins, G., Hetzel, J., Hirano, T., Hulme, D., Jorgensen, C., Kessler, M., Kirkpatrick, B.W., Konfortov, B., Kostia, S., Kuhn, C., Lenstra, J.A., Leveziel, H., Lewin, H.A. Leyhe, B., Lil, L., Martin Burriel, I., McGraw, R.A., Miller, J.R., Moody, D.E., Moore, S.S., Nakane, S., Nijman, I.J., Olsaker, I., Pomp, D., Rando, A., Ron, M., Shalom, A., Teale, A.J., Thieven, U., Urquhart, B.G.D., Vage, D.-I., Van de Weghe, A., Varvio, S., Velmala, R., Vilkki, J., Weikard, R., Woodside, C., Womack, J.E., Zanotti, M., Zaragoza, P. (1997) A medium-density genetic linkage map of the bovine genome. Mammalian Genome. 8(1):21-28.

Botstein D., White R.L., Skolnick M., Davis R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**, 314-331.

Brascamp E.W., van Arendonk J.A.M. & Groen A.F. (1993) Economic appraisal of the utilization of genetic markers in dairy cattle breeding. *Journal of Dairy Science* **76**, 1204-13.

Crawford, A.M., Montgomery G.W., Pierson C.A.,Brown T., Dobbs K.G., Sunden S.L.F., Henry, H.M., Ede A.J., Swarbrick P.A., Berryman T, Penty J.M, and Hill D.F. (1994) Sheep linkage mapping: nineteen linkage groups derived from the analysis of paternal half-sib families. *Genetics* **137**, 573-79.

Da Y. & Lewin H.A. (1995) Linkage information content and efficiency of fullsib and halfsib designs for gene mapping. *Theoretical and Applied Genetics* **90**, 699-706.

Garber R.A. & Morris J.W. (1983) General equations for the average power of exclusion for genetic systems of n codominant alleles in one-parent and no-parent cases of disputed parentage. In: *Inclusion Probabilities in Parentage Testing*, pp. 277-80.

Jamieson A. (1965) The genetics of transferrin in cattle. *Heredity* **20**, 419-41.

Jamieson A. (1994) The effectiveness of using co-dominant polymorphic allelic series for (1) checking pedigrees and (2) distinguishing full-sib pair members. *Animal Genetics* **25**, Supplement 1, 37-44.

Ma R.Z. Beever J.E., Da Y., Green C.A., Russ I., Park C., Heyen D.W., Everts R.E., Fisher S.R., Overton K.M., Teale A.J., Kemp S.J., Hines H.C., Guérin G. & Lewin H.A. (1996) A male linkage map of the cattle (Bos taurus) Genome. *Journal of Heredity* **87**, 261-271.

Rohrer, G.A., L.J. Alexander, Z. Hu et al., (1996). A comprehensive map of the porcine genome. Genome Research. 6:371-391.

Ron, M., Lewin H.A., Da Y., Band M., Yanai A, Blank Y, Feldmesser E. and Weller J.I. (1995). Prediction of informativeness for microsatellite markers among progeny of sires used for detection of economic trait loci in dairy cattle. *Animal Genetics* **26**, 439-441.

Ron M., Band M., Wyler A., Weller J.I. (1993) Unequivocal determination of sire allele origin for multiallelic microsatellites when only the sire and progeny are genotyped. *Animal. Genetics* **24**, 171-176.

Weir B.S. (1990) *Genetic Data Analysis*. Sinauer Associates Inc., Sunderland, MA, USA, p. 87.

Weller J.I., Kashi Y. & Soller M. (1990) Power of 'daughter' and 'granddaughter' designs for genetic mapping of quantitative traits in dairy cattle using genetic markers. *Journal of Dairy Science* **73**, 2525-37.