Bioinformatics and math statistics

UDC 575.17:575.118.5:575.162:57.087.1

doi: 10.15389/agrobiology.2015.5.571rus

doi: 10.15389/agrobiology.2015.5.571eng

EVALUATION OF THE MEASURE OF POLYMORPHISM INFORMATION OF GENETIC DIVERSITY

Yu.V. CHESNOKOV, A.M. ARTEMYEVA

Federal Research Center the N.I. Vavilov All-Russian Institute of Plant Genetic Resources, Federal Agency of Scientific Organizations, 42-44, ul. Bol'shaya Morskaya, St. Petersburg, 190000 Russia, e-mail yu.chesnokov@vir.nw.ru
Acknowledgements:

Supported in part by Russian Foundation for Basic Research (grant № 13-04-00128-a)

Received February 26, 2015

Abstract

Gene identification and mapping are one of the main goals of plant and animal genetics. Upon verifying genetic linkage it is usually found which marker loci (markers) possess alleles co-segregated with the alleles of the desired locus. Marker utility for these purposes depends on the number of alleles, which the marker possesses, and their relative rates. There are two indexes, or measures, usually used for the polymorphism degree evaluation. They are the heterozygosity (H) for which the evaluation method and variability formula are well known (M. Nei et al., 1974, 1979), and polymorphism information content (PIC) (D. Botstein et al., 1980). Based on published data, we described the statistical approaches which are used for analysis of polymorphism information. Herein, the value of polymorphism information content, heterozygosity and some associated values detected upon evaluation of genetic diversity on interspecific and intraspecific population levels are considered. PIC shows how the marker can indicate the population polymorphism depending on the number and frequency of the alleles (D. Botstein et al., 1980). So the PIC reflects a discriminating ability of the marker and, in fact, depends on the number of known alleles and their frequency distribution, thus being equal to genetic diversity. PIC maximal value for dominant markers is 0.5. Note, that for the markers with equal distribution in the population the PIC values are higher. They are much higher for markers with multiple alleles, and, however, also depend on the frequency distribution of the alleles. Using 135 SSR (simple sequence repeats) and 123 S-SAP (sequence specific amplified polymorphism) primers, we found 135 SSR and 123 S-SAP polymorphic markers among 96 *Brassica rapa* L. samples from the VIR (N.I. Vavilov Institute of Plant Genetic Resources) core collection. The PIC values for both markers, SSR and S-SAP markers were 0.316, 0.257 and 0.379 (50 % higher on average), respectively. Expected heterozygosity (H_E) is usually used to describe the genetic diversity because it is less sensitive to the sample size compared to observed heterozygosity (H_O). The crossings in the population are occasional, if H_O and H_E are similar (i.e., no reliable differences found). They are related as $H_O < H_E$ in inbred population, and as $H_O > H_E$ in case of occasional crossing prevailing compared with inbreeding. Effective multiplex ratio (EMR) is calculated as total number of polymorphic loci per primer multiplied by the rate of polymorphic loci from their total number (W. Powell et al., 1996; J. Nagaraju et al., 2001). Marker index (MI) is a statistical parameter used to estimate total utility of the marker system; the higher MI, the better method is used (W. Powell et al., 1996; J. Nagaraju et al., 2001). Resolving power (Rp) is a parameter characterizing ability of the primer/marker combination to detect differences between large numbers of genotypes (J.E. Gilbert et al., 1999; A. Prevost et al., 1999). The information about some software which can be used for calculation of polymorphism information content value and heterozygosity is also summarized. The formula for effective multiplex ratio, marker index calculation, and resolving power calculation are shown.

Keywords: heterozygosity, polymorphism information content value, effective multiplex ratio, marker index, resolving power, software.

Identification and mapping of genes responsible for the expression of the traits that are of the interest to the researcher are one of the main goals of plant and animal genetics. There is a large number of marker loci visualized using various marker systems (briefly referred to as markers) the positions of which and their order within a chromosome are well known. Upon verifying genetic

linkage it is usually found which marker loci (markers) possess alleles co-segregated with the alleles of the desired locus. Marker utility for these purposes depends on the number of alleles, which the marker possesses, and their relative rates. Qualitatively, a marker is characterized as polymorphic if it contains at least two alleles, and the most common allele has a frequency in a population of at least 99 %. The polymorphism degree evaluation is usually measured by two indexes (measures). One of them is known as heterozygosity (H) for which the evaluation method and variability formula are well known [1, 2]. The other measurement unit in the polymorphism information content (PIC) [3].

Molecular markers have become an effective tool and a means by which both intra- and inter-species genetic diversity is evaluated and characterized. Marker systems are distinguished by the extent (i.e., magnitude) of their informativeness, which in turn depends on the degree of polymorphism. The concept of polymorphism is used to determine the genetic variability in the population, which in recent decades has become the subject of intense study by various disciplines (genetics, ecology, botany, zoology and some others). Examples of this are numerous and obvious [4-10]. However, when planning the use of molecular markers for any research or for the practical use in breeding programs, the questions arise inevitably, and the researchers often have to look for the answers. How difficult is it to find polymorphic loci that are suitable for the planned work? How many markers will be required to be used? How polymorphic should any selected marker be? All these questions can be answered by estimating the measure of the marker informativeness. The two main parameters determined for this purpose are heterozygosity (H) and the polymorphism information content (PIC). In addition to these, there are some associated indicators with the help of which the effectiveness of the chosen system of «primer-marker» and/or of the chosen methodological approach can be determined.

In this paper, we summarized the published data on the statistical approaches which are used for the analysis of polymorphism information, heterozygosity and some associated values determined in the assessment of genetic diversity in both interspecies and intraspecies population levels.

Heterozygosity (H). Locus heterozygosity which is defined as the probability of an individual's heterozygosity in the population in this locus [11] can be calculated by the formula:

$$H = 1 - \sum_{i=1}^l P_i^2, \quad (1)$$

where P_i is the frequency of the i -th allele among the total number of alleles l . In other words, heterozygosity may be considered as an average portion of loci with two different alleles at one locus in a single individual. Usually, this applies to the whole population or some part of it and is divided into the observed and expected heterozygosity. The expected heterozygosity (H_E), or genetic diversity according to M. Nei [1], is the expected probability of an individual's heterozygosity for the relevant locus in multilocus systems (for all analyzed loci). In other words, it is the determined fraction of all individuals that were heterozygous for any randomly selected locus. It is often calculated based on determining the square root of the frequency of the null allele (recessive) as follows similar to equation (1): $H_E = 1 - \sum_i^n p_i^2$, where p_i is the frequency of the i -th allele, n_i is the total number of alleles at all loci. The observed heterozygosity (H_O) is the part of heterozygous genes in the population. It is calculated for each locus as the total number of heterozygotes divided by the sample size. The values of H_E and H_O range from 0 (no heterozygosity) to substantially 1 (a large number of alleles

with equal frequency). The expected heterozygosity is usually defined when describing genetic diversity, because it is less sensitive to the sample size than the observed heterozygosity. When H_O and H_E are similar (not significantly different), the crossing in the population is almost accidental. When $H_O < H_E$, it is an inbred population. When $H_O > H_E$, the random mating system dominates in-breeding in the population.

Polymorphism information content (PIC). The measure or value of the polymorphism information content (PIC) is determined by the ability of a marker to establish polymorphism in the population depending on the number of alleles detected and on their distribution frequency [3]. Thus, PIC identifies the discriminatory ability of the marker, actually depends on the number of known (established) alleles and their distribution frequency, and thus it is equivalent to the gene diversity. In its simplest form, the value of PIC can be calculated like heterozygosity, see equation (1):

$$PIC_j = 1 - \sum_{i=1}^n P_i^2, \quad (2)$$

where i is i -th allele of the j -th marker, n is the number of the j -th marker's alleles, P is allele frequency. Sample calculations for the value of PIC and multiallelic markers are shown in Table 1. At the same time, for dominant markers, equation (2) can be represented as follows [12]:

$$PIC = 1 - (k \sum_{i=1}^{k-1} P_i^2) - k^{-1} \sum_{i=1}^{k-1} \sum_{j=1}^k 2 P_i^2 P_j^2, \quad (3)$$

where k is the number of alleles, P_i and P_j are frequency of the i -th and j -th alleles in the population, respectively. For dominant markers, the PIC value is calculated as described [13]:

$$PIC = 1 - [f^2 + (1 - f)^2], \quad (4)$$

where f is the marker frequency in the data set. For the dominant markers, the maximum PIC value is 0.5. Note, that for the markers with equal distribution in the population the PIC values are higher. The values are much higher for the markers with multiple alleles, however, the PIC value also depends on the distribution frequency of the alleles (see Table 1).

1. Sample calculations for PIC for biallelic and multiallelic markers

Allele frequency	Formula for calculating from the equation (2)	PIC value
Biallelic marker		
$P_1 = 0.5; P_2 = 0.5$	$1 - (0.5^2 + 0.5^2)$	0.50
$P_1 = 0.4; P_2 = 0.6$	$1 - (0.4^2 + 0.6^2)$	0.48
$P_1 = 0.3; P_2 = 0.7$	$1 - (0.3^2 + 0.7^2)$	0.42
$P_1 = 0.2; P_2 = 0.8$	$1 - (0.2^2 + 0.8^2)$	0.32
$P_1 = 0.1; P_2 = 0.9$	$1 - (0.1^2 + 0.9^2)$	0.18
Multi allelic marker		
$P_1 = 0.33; P_2 = 0.33; P_3 = 0.33$	$1 - (0.332 + 0.332 + 0.332)$	0.67
$P_1 = 0.4; P_2 = 0.3; P_3 = 0.3$	$1 - (0.42 + 0.32 + 0.32)$	0.66
$P_1 = 0.4; P_2 = 0.4; P_3 = 0.2$	$1 - (0.42 + 0.42 + 0.22)$	0.64
$P_1 = 0.5; P_2 = 0.3; P_3 = 0.2$	$1 - (0.52 + 0.32 + 0.22)$	0.62
$P_1 = 0.5; P_2 = 0.4; P_3 = 0.1$	$1 - (0.52 + 0.42 + 0.12)$	0.58
$P_1 = 0.6; P_2 = 0.2; P_3 = 0.2$	$1 - (0.62 + 0.22 + 0.22)$	0.56
$P_1 = 0.6; P_2 = 0.3; P_3 = 0.1$	$1 - (0.62 + 0.32 + 0.12)$	0.54
$P_1 = 0.7; P_2 = 0.2; P_3 = 0.1$	$1 - (0.72 + 0.22 + 0.12)$	0.46
$P_1 = 0.8; P_2 = 0.1; P_3 = 0.1$	$1 - (0.82 + 0.12 + 0.12)$	0.35
Note. PIC — polymorphism information content.		

In our research, in evaluating the core collection of *Brassica rapa* L. (96 samples) stored in the VIR, using 21 pairs of SSR (simple sequence repeats), and 12 pairs of S-SAP (sequence specific amplified polymorphism) of primers, we found 135 SSR and 123 S-SAP polymorphic markers that were used

to construct the phylogenetic tree [14]. PIC values were calculated for any marker (Table 2).

2. Identified SSR and S-SAP alleles molecular markers and their PIC among the VIR *Brassica rapa* L. core collection (N.I. Vavilov All-Russian Institute of Plant Genetic Resources)

Number of alleles	Proportion of alleles, %	PIC value
SSR markers		
39	28.9	0-0.1
20	14.8	0.1-0.2
16	11.9	0.2-0.3
20	14.8	0.3-0.4
40	29.6	0.4-0.5
Total	100	
135		
S-SAP markers		
4	3.2	0-0.1
9	7.3	0.1-0.2
18	14.6	0.2-0.3
18	14.6	0.3-0.4
74	60.0	0.4-0.5
Total	100	
123		

Note. SSR — simple sequence repeats, S-SAP — sequence specific amplified polymorphism. PIC — polymorphism information content.

Their frequency distribution for each type of markers is shown in the graphs (Fig.). The PIC value for both markers was 0.316, while it was 0.257 for microsatellite markers and 0.379 for S-SAP markers, i.e. 50 % higher on average, respectively. Thus, in our studies on the evaluation of genetic polymorphisms in the *B. rapa* collection using two markers, the most informativeness was noted in S-SAP markers.

At the same time, when determining the genetic diversity in rice (*Oryza sativa* L.), the average PIC value

was 2 times higher in the case of using SSR markers (0.66) than with the use of RFLP (restriction fragment length polymorphism) markers (0.36) [15].

Distribution of PIC (polymorphism information content) values for SSR (simple sequence repeats, A) and S-SAP (sequence specific amplified polymorphism, B) molecular markers in the evaluation of the *Brassica rapa* L. core collection stored in VIR (N.I. Vavilov All-Russian Institute of Plant Genetic Resources).

In the study of molecular genetic diversity in sweet corn (*Zea mays* L.) using RAPD (random amplified polymorphic DNA) and SSR markers, the results indicating a higher similarity between the studied populations than within them, were obtained [16]. The authors note that the RAPD markers had lower average PIC values (0.17) than the SSR markers (0.57). The absolute PIC values for RAPD and SSR in the above case could not be compared due to the maximum PIC value of 0.5 and 1.0 for RAPD and SSR loci, respectively. RAPD and ISSR markers were also used to assess the genetic variability and relationships among Tunisian local varieties of barley (*Hordeum vulgare* L.) [17]. Despite the high level of polymorphism revealed for both RAPD and SSR markers and the mean PIC values of 0.477 and 0.533 for the RAPD and SSR markers, respectively, the authors conclude that the SSR markers are better suited to assess the genetic diversity of barley than RAPD markers, as SSR markers have greater polymorphism (90.7 %) compared with the RAPD markers (74.0 %).

And finally, another example of PIC value estimation to which we would like to draw your attention. The researchers from Argentina and the USA have

analyzed the genetic diversity of the Argentinean varieties of wheat (*Triticum aestivum* L.), created in the period from 1932 to 1995 [18]. Using the SSR and AFLP (amplified fragment length polymorphism) markers, they found that there were no significant differences in the genetic diversity between the group of varieties created until 1960 and the groups produced in each of the next three decades. The mean diversity estimated with the SSR markers was virtually identical for all the four time periods. The genetic diversity identified through the AFLP markers confirmed the absence of genetic diversity reduction over time. However, significant differences ($P = 0.01$) were found between the varieties of soft wheat created in the 1970s ($PIC = 0.28$) and 1980s ($PIC = 0.34$). The overall results obtained with PIC indicate that the Argentinean varieties of soft wheat were maintained almost at the same level of genetic diversity for over 60 years, and their differences were largely due to the implemented breeding programs, but not to the degree of genetic diversity of the derived varieties. Thus, the measure of polymorphism information content is an important component in the planning of breeding programs and one of the key information and statistical indicators in their implementation.

Software for calculation of H and PIC . To correctly plan a genetic research and evaluation of the results, calculation of heterozygosity (H) and polymorphism information content (PIC) values is often required to describe the marker informativeness, but until recently there have been no simple and public calculators for such computations. To simplify the work on the marker research, in 2012, a group of Hungarian scientists proposed their interactive online $PICcalc$ program (<http://w3.georgikon.hu/pic/english/default.aspx>) [19]. The program makes it possible to calculate the values of H and PIC over allele frequencies with the manual introduction of the values or using a special file containing a binary data matrix. The additional options make it possible to calculate the values for the certain number of loci using a simple text file which ensures estimation of H and PIC for the primer or primer sets used for the analysis of different genetic marker systems associated with binary data. For multilocus markers such as AFLP, ISSR (inter simple sequence repeats) or RAPD, it is theoretically assumed that the fragments of equal length are amplified at the corresponding chromosome loci and that they represent a single dominant locus with two possible alleles (amplicon presence/absence). The maximum value for H and PIC for dominant markers in this case is equal to 0.5, as only two alleles per locus are permitted for this type of markers and both values are influenced by the number and frequency of alleles [13, 20, 21]. Given this characteristic of dominant markers, the program specifically provides for the possibility of calculating H and PIC for them [19].

Earlier, a group of American scientists also offered a computer program (<http://darwin.cwru.edu/pic>) makes it possible to calculate the H and PIC parameters [22]. The main difference with the above-mentioned Internet resource is the uniformly distributed minimum deviation of the unbiased PIC estimation in accordance with the exact dispersion value. To estimate this, the authors derived the formula for the calculation of any polynomial in the set of variables distributed multinomially.

Associated values. Effective multiplex ratio (EMR) is calculated as total number of polymorphic loci (per primer) multiplied by the proportion of polymorphic loci per their total number [23, 24]:

$$EMR = n_p(n_p/n), \quad (5)$$

where n_p is the number of polymorphic loci, and n is the total loci number. The higher the value of EMR, the more efficient the «primer—marker system» is.

Marker index (MI) is a statistical parameter used to estimate the total utility of the marker system. Marker index is the product of the polymorphism information content value (or expected heterozygosity, H_E) and effective multiplex ratio [23, 24]:

$$MI = PIC \times EMR, \quad (6)$$

The higher MI, the better the method is. Resolving power (R_p) is a parameter used to characterize the ability of the primer/marker combination to detect the differences between a large numbers of genotypes [25, 26]:

$$R_p = \sum I_b, \quad (7)$$

where $I_b = 1 - (2 \times 0.5 - p)$ is an amplicon informativeness, p is a proportion of individuals with identified amplicon I .

Thus, there are several approaches that are designed to measure the polymorphism information content and the associated values. DNA markers are now recognized as a rather convenient and high-quality tool for assessing genetic diversity at the molecular level. However, before using a particular marker system, it is necessary to evaluate the technical equipment of a laboratory, the need for the use of the selected marker system and its compliance with the current tasks, staff professionalism, as well as the upcoming maintenance cost and the available means of support services. The required software should be selected based on the calculation of its suitability for solving problems faced by the researcher, including the problems of population genetics when it comes to assessing the information polymorphism. The morphological parameters are important for the interpretation of the results. The estimation of statistically significant association and correlation between morphological and molecular genetic parameters is the key factor in the final decision. And of course, we cannot ignore the biological characteristics of the species studied in the evaluation of genetic parameters, as one and the same parameter can be different in different species not only in phylogeny but also in ontogeny. The latter is particularly important for the effect of the «genotype—environment» interaction.

REFERENCES

1. Nei M., Roychoudhury A.K. Sampling variances of heterozygosity and genetic distance. *Genetics*, 1974, 76: 379-390.
2. Nei M., Li W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *PNAS USA*, 1979, 76: 5269-5273 (doi: 10.1073/pnas.76.10.5269).
3. Botstein D., White R.L., Skalnick M.H., Davies R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphism. *Am. J. Hum. Genet.*, 1980, 32: 314-331.
4. Artem'eva A.M., Chesnokov Yu.V., Kloke E. *Doklady RASKHN*, 2008, 5: 14-18.
5. Artem'eva A.M., Chesnokov Yu.V., Kloke E. *Informatsionnyi Vestnik VOGiS*, 2008, 12(4): 608-619.
6. Artem'eva A.M., Kloke E., Chesnokov Yu.V. *Informatsionnyi vestnik VOGiS*, 2009, 13(4): 759-771.
7. Vishnyakova M.A., Burlyaeva M.O., Alpat'eva N.V., Chesnokov Yu.V. *Informatsionnyi vestnik VOGiS*, 2008, 12(4): 595-607.
8. Artem'eva A.M., Chesnokov Yu.V. *Vavilovskii zhurnal genetiki i selektsii*, 2012, 16(4/2): 1047-1060.
9. Ivanov A.A., Burenin V.I., Chesnokov Yu.V. *Doklady Rossel'khozakademii*, 2012, 3: 20-22.
10. Chesnokov Yu.V., Burenin V.I., Ivanov A.A. *Sel'skokhozyaistvennaya biologiya [Agricultural Biology]*, 2013, 3: 28-36 (doi: 10.15389/agrobiolgy.2013.3.28rus, doi: 10.15389/agrobiolgy.2013.3.28eng).
11. Liu B.H. *Statistical genomics: linkage, mapping and QTL analysis*. CRC Press, Boca Raton, 1998.
12. Anderson J.A., Churchill G.A., Autrique J.E., Tanksley S.D., Sorrells M.E. Optimizing parental selection for genetic linkage maps. *Genome*, 1993, 36(1): 181-186 (doi: 10.1139/g93-024).

13. De Riek J., Calsyn E., Everaert I., Van Bockstaele E., De Loose M. AFLP-based alternatives for the assessment of distinctness, uniformity and stability of sugar beet varieties. *Theor. Appl. Genet.*, 2001, 103: 1254-1265 (doi: 10.1007/s001220100710).
14. Artem'eva A.M., Budan Kh., Kloke E., Chesnokov Yu.V. *Vavilovskii zhurnal genetiki i selektsii*, 2011, 15(2): 398-411.
15. Xu Y., Beachell H., McCouch S.R. A marker-based approach to broadening the genetic base of rice (*Oryza sativa* L.) in the US. *Crop Sci.*, 2004, 44: 1947-1959 (doi: 10.2135/cropsci2004.1947).
16. Bered F., Freitas Terra T., Spellmeier M., Neto J.F.B Genetic variation among and within sweet corn populations detected by RAPD and SSR markers. *Crop Breed. Appl. Biotechnol.*, 2005, 5: 418-425.
17. Karim K., Rawda A., Hatem C.-M. Genetic diversity in barley genetic diversity in local Tunisian barley based on RAPD and SSR analysis. *Biological Diversity and Conservation*, 2009, 2/1: 27-35 (doi: 10.1590/S0100-204X2015000200008).
18. Manifesto M.M., Schlatter A.R., Hopp H.E., Suarez E.Y., Dubcovsky J. Quantitative evaluation of genetic diversity in wheat germplasm using molecular markers. *Crop Sci.*, 2001, 41: 682-690 (doi: 10.2135/cropsci2001.413682x).
19. Nagy S., Poczaï P., Cernak I., Gorji A.M., Hegedus G., Tallér J. PICcalc: An online program to calculate polymorphic information content for molecular genetic studies. *Biochem. Genet.*, 2012, 50: 670-672 (doi: 10.1007/s10528-012-9509-1).
20. Bolaric S., Barth S., Melchinger A.E., Posselt U.K. Genetic diversity in European perennial ryegrass cultivars investigated with RAPD markers. *Plant Breed.*, 2005, 124: 161-166 (doi: 10.1111/j.1439-0523.2004.01032.x).
21. Henry R.J. *Practical applications of plant molecular biology*. Chapman and Hall, London, 1997.
22. Shete S., Tiwari H., Elston R.C. On estimating the heterozygosity and polymorphism information content value. *Theor. Popul. Biol.*, 2000, 57: 265-271 (doi: 10.1006/tpbi.2000.1452).
23. Powell W., Morgante M., Andre C., Hanafey M., Vogel J., Tingey S., Rafalski A. The utility of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol. Breed.*, 1996, 2: 225-238 (doi: 10.1007/BF00564200).
24. Nagaraju J., Damodar Reddy K., Nagaraja G.M., Sethuraman B.N. Comparison of multilocus RFLPs and PCR-based marker systems for genetic analysis of the silkworm, *Bombyx mori*. *Heredity*, 2001, 86: 588-597 (doi: 10.1046/j.1365-2540.2001.00861.x).
25. Gilbert J.E., Lewis R.V., Wilkinson M.J., Caligari P.D.S. Developing an appropriate strategy to assess genetic variability in plant germplasm collections. *Theor. Appl. Genet.*, 1999, 98: 1125-1131 (doi: 10.1007/s001220051176).
26. Prevost A., Wilkinson M.J. A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars. *Theor. Appl. Genet.*, 1999, 98: 107-112 (doi: 10.1007/s001220051046).