Crop Breeding and Applied Biotechnology 9: 308-312, 2009
Brazilian Society of Plant Breeding. Printed in Brazil

Efficiency of the multilocus analysis for the construction of genetic maps

Leonardo Lopes Bhering[1*], Cosme Damião Cruz[2], Edmar Soares de Vasconcelos[2], Márcio Fernando Ribeiro de Resende Júnior[2], Willian Silva Barros[2], and Tatiana Barbosa Rosado[2]

Received 24 November 2008

Accepted 05 August 2009

ABSTRACT- *The use of genetic maps is a useful tool in genetic research. The association between map distance and recombination frequency is expressed by a genetic mapping function. However, several of these functions do not presuppose the joint recombination percentage. In other words, they are not multilocus probabilities. This work aimed to compare, through simulations, the efficiency in the use of different mapping functions with and without multilocus analysis as a tool in the construction of genetic maps. A genome constituted of three linkage groups (50, 100 and 200 cM) was simulated for a comparative study. Four mapping populations were simulated, F_2, with 50, 100, 200 and 400 individuals, with 10 replicas each. It was verified, after the analyses, that the multilocus analysis was not efficient to rescue the size of the connection groups, concluding that the non use of the multilocus analysis would be viable.*

Key words: Statistical genomics, simulation, mapping.

INTRODUCTION

Genetic maps are useful tools in various fields of genetic research. The map distance between two markers is defined as a significant number of recombination events in each meiosis, expressed in centimorgans (cM). The relationship between map distance and recombination frequency is expressed by a mapping function. Different mapping functions correspond to different degrees of interference in crossing over (Stam 1993). However, various mapping functions fail to yield proper joint recombination probabilities for more than three loci, which mean they are not multilocus probabilities (Goldstein et al. 1995).

Models for multilocus analysis are different from models for two loci analysis because it considers all the information of the intervals between pairs of loci in a linkage group (Schuster and Cruz 2004).

According to Liu (1998), there are a few obstacles to the implementation of the multilocus analysis in genetic linkage studies and in the construction of maps. Some of those obstacles are the complexity of the multilocus likelihood function, the large number of parameters when high levels of interference occurs and the loss of part of the data necessary to construct a multilocus model. In addition there is still demand for intensive computational time for the construction of genetic maps in this multilocus model.

When the multilocus analysis is used, the appropriateness of the mapping functions depends on whether what is assumed about the existence and magnitude of interference is true or false. The choice

[1] Embrapa Agroenergia, Av. W3 Norte, (Pq EB), s/n, 70770-901, Brasília, DF, Brazil. *E-mail: leonardo.bhering@embrapa.br
[2] Universidade Federal de Viçosa (UFV), Departamento de Biologia Geral, 36570-000, Viçosa, MG, Brazil

of the mapping function is very important to the result obtained since the adoption of one, between the different mapping functions, depends on assumptions about the distribution of the crossing over, degree of interference and length of the chromosome segment in question (Schuster and Cruz 2004).

The objective of this study is to compare, through simulations, the efficiency in the use of different mapping functions with and without multilocus analysis as a tool in the construction of genetic maps.

MATERIAL AND METHODS

A genome with three linkage groups was simulated. The first linkage group had 50 centimorgans (cM), the second and third had 100 cM and 200 cM respectively. Each group had 11 molecular markers, co-dominant, and equally spaced, so that each linkage group had the saturation of 5, 10 and 20 cM respectively. Homozygous and contrasting parents were used. Parent 1 was dominant and parent 2 was recessive. We simulated four F 2 mapping populations with 10 replicas each. The size of each population was 50, 100, 200 and 400 individuals respectively.

After these simulations the distance matrices were generated using the different mapping functions and multilocus analysis, resulting in five treatments. The first treatment had the distance unit expressed in recombination frequency. This was used as a control treatment since it did not use any multilocus analysis. The second treatment had the distance expressed in Haldane function and multilocus analysis based on the maximum likelihood method. The third treatment had the distance expressed in Haldane´s function and analysis based on multilocus method of least squares. Treatment four had the distance expressed in Kosambi´s function and multilocus analysis based on the maximum likelihood method. Treatment five had the distance expressed in Kosambi´s function and analysis based on multilocus method of least squares.

All simulations and comparisons of the genomes were accomplished using the software GQMOL (Cruz and Schuster 2001)

After estimating the distances between the marks an analysis of variance (Anova) was performed for the total length of each linkage group. To compare the means of each treatment studied, a Tukey test was performed using the software Genes (Cruz 2006).

RESULTS AND DISCUSSION

In the population with 50 individuals it was observed, through the analysis of variance, statistical significance only between treatments in the first linkage group with length of 50 cM (GL1). Treatments for the second linkage group (100cM) did not show statistical significance (Table 1). In this population, the treatments for the third linkage group (200cM) were not analyzed because during the mapping procedure, in 70% of the times (7 replicas), the markers located in this linkage group were mapped in more than one group. This event shows that population's sizes below 50 individuals are not sufficient to recover the information during the mapping procedure when a low level of saturation, such as 20cM, is used. In this case, larger populations or higher saturation levels should be used (Bhering and Cruz 2008).

The treatments where multilocus analysis was performed based on least square (treatments 3 and 5) showed worse distance estimations in relation to values obtained by the method of maximum likelihood (treatments 2 and 4), although this was not a statistically significant difference (Table 2). In this case the closer to 50 cM indicates that is the best methodology to recover the information originally set out in simulations.

Table 1. Analysis of variance of the two linkage groups for the population of 50 individuals GL1(50cM) and GL2 (100cM)

		Lenght (GL1)		Lenght (GL2)	
Source of variation	**df**	**MS**	**F**	**MS**	**F**
Treatments	4	207.4724**	4.6085	1527.1603	0.0785
Residue	45	45.0197		1945.0596	
Total	49				
Mean		47.1706		102.2828	
CV(%)		14.2242		43.1185	

** Significant at a statistical significance level of 1% by F test

Table 2. Tukey test performed on the population of 50 individuals at the significance level of 5%

Treatment	Length (GL1)	
3	50.667	a*
1	50.44	ab
2	50.376	ab
4	42.395	ab
5	41.975	b
DMS	8.5296	

* Averages followed by the same letter do not differ statistically at 5% level

It was also observed that the distances estimations (42,395 and 41,975) were more distant from the expected size of the linkage group when the Kosambi function was used (treatments 4 and 5 respectively).

In the population of 100 individuals it was observed that the linkage groups 1 and 2 (GL1 and GL2) had significant differences at 1%, between their treatments. Only the GL3 was not statistically different for the treatments tested (Table 3). All the treatments succeeded in mapping the simulation information showing that by increasing the population size of the original information can be correctly estimated.

Table 4 shows the result obtained by the Tukey test performed for the first and second linkage group (GL1 and GL2) since the third linkage group was not statistically significant. The treatments 1, 2 and 3 did not differ among themselves, and have their values closer to what was set initially (50 cM for GL1 and 100 cM for GL2). Once again it was observed that when the Kosambi function is used, the distances are more divergent than desirable. It was also observed that distance estimates for treatment 1 (where the multilocus analysis was not used) were closer to the values initially set (50 cM for GL1 and 100 cM for GL2) although this result was not statistically different.

The treatments of population of 200 individuals showed significant differences for all 3 sizes of the linkage group tested as can be seen in the Table 5 of analysis of variance.

The Tukey test was performed on the population of 200 individuals (Table 6). It was observed that for the GL1, treatments 1, 2 and 3 did not differ among themselves, and the distance values were the closest to the original 50 cM. For GL2, Treatment 1, which did not use multilocus analysis, showed the closest estimation of the expected distance of 100 cM. Similar results were observed for GL3, but in this case treatments 1, 2 and 3 did not differ statistically. This indicates that the multilocus analysis was not an effective methodology for this type of analysis.

Table 3. Analysis of variance of the three linkage groups for the population of 100 individuals

	Length (GL 1)			Length (GL 2)		Length (GL3)	
Source of variation	df	MS	F	MS	F	MS	F
Treatments	4	234.864**	5.802	1472.068**	21.506	28515.367	2.061
Residue	45	40.474		68.447		13830.765	
Total	49						
Mean		47.865		98.377		261.575	
CV(%)		13.291		8.409		44.960	

** Significant at a statistical significance level of 1% by F test

Table 4. Tukey test performed on the population of 100 individuals at a significance level of 5%

Treatment	Length (GL1)		Treatment	Length (GL2)	
2	52.928	a *	3	109.95	a *
3	51.17	a	2	109.668	a
1	49.83	ab	1	100.86	a
4	42.975	b	4	85.937	b
5	42.423	b	5	85.474	b
DMS	8.0875	DMS	10.5173		

* Averages followed by the same letter do not differ statistically at 5% level

Table 5. Analysis of variance of the three linkage groups for the population size of 200

		Length (GL 1)		Length (GL 2)		Length (GL 3)	
Source of variation	**df**	**MS**	**F**	**MS**	**F**	**MS**	**F**
Treatments	4	230.5899**	9.512	1607.9652**	31.407	17556.8225**	9.193
Residue	45	24.2221		51.1974		1909.7841	
Total	49						
Mean		47.7306		104.7128		253.6142	
CV(%)		10.3112		6.8332		17.2313	

** Significant at a statistical significance level of 1% by F test

Table 6. Tukey test performed on the population of 200 individuals at the significance level of 5%

Treatment	Length (GL1)		Treatment	Length (GL2)		Treatment	Length (GL3)	
1	51.56	a	3	118.114	a	3	311.159	a
2	51.365	a	2	116.912	a	2	273.571	ab
3	50.754	a	1	104.19	b	5	253.961	bc
4	42.827	b	4	92.343	c	4	226.58	bc
5	42.147	b	5	92.005	c	1	202.8	c
DMS	6.2565		DMS	9.096		DMS	55.5544	

The Haldane´s function is more effective for smaller linkage groups (50 cM) than the Kosambi´s function. When the size of the linkage group is increased, distances in Kosambi´s function becomes more effective. Considering the same unit of distance (Haldane or Kosambi) there was not any statistical difference between the two methods of multilocus analysis (Maximum Likelihood and Least Squares).

The analysis of variance of the linkage group length for the population size of 400 is shown in Table 7. Note that it was observed, for all three linkage groups, significant differences between the treatments. The tukey test, shown in table 8, demonstrated similar results than the ones obtained in table 6 (population of 200 individuals). Treatments 1, 2 and 3, did not differ among themselves. Treatment 1, without multilocus analysis statistically differed from the others For the linkage groups 2 and 3 (GL2 and GL3). The linkage group´s lengths of treatment 1 were the closest among all treatments to the desired size (100 cM gor GL2 and 200 cM for GL3).

In possession of all the results, it could be observed that multilocus analysis was not efficient to recover the linkage group´s size of this simulation. In all cases the treatment which the distance estimation was expressed in recombination frequency (Morgan), obtained the best distance estimate, since their values were always the closest to the values initially set.

Table 7. Analysis of variance of the three linkage groups for the population size of 400

		Length (GL 1)		Length (GL 2)		Length (GL 3)	
Source of variation	**df**	**MS**	**F**	**MS**	**F**	**MS**	**F**
Treatments	4	226.138**	17.152	1690.948**	51.151	15024.993**	8.486
Residue	45	13.184		33.057		1770.502	
Total	49						
Mean		50.410		103.056		245.251	
CV(%)		7.202		5.579		17.156	

** Significant at a statistical significance level of 1% by F test.

Another important fact that we can infer is that Haldane's function was more efficient when smaller linkage groups (50 cM) were used. In both cases significant differences between the two multilocus methodologies tested could not be observed (Maximum Likelihood and Least Squares). Therefore the superiority of either strategy for multilocus analysis cannot be inferred.

When the population's size was increased, larger differences between treatments were observed. In the population of 50 individuals only linkage group 1 showed statistic difference and GL3 was unable to reproduce the original linkage group. All the other populations' sizes reproduced the original linkage group and all the groups were statistically significant in the sizes of 200 and 400.

With the results obtained it could be observed that:

The multilocus analysis was not efficient in any case for the construction of genetic maps and was not useful to improve the efficiency of the distance estimates.

It was not possible to infer the efficiency of the two multilocus methodologies tested since we could not detect statistically significant differences between them.

Distances expressed in Kosambi´s function were more efficient thant Haldane´s function for larger linkage groups.

The simulation study was better accomplished with larger population sizes since in those sizes the differences between treatments could be observed.

Eficiência da análise multiloco para a construção de mapas genéticos

RESUMO - *Na pesquisa genética, a utilização de mapas genéticos se torna uma ferramenta muito útil, sendo que a relação entre distância de mapa e freqüência de recombinação é expressa por uma função de mapeamento genético. Porém, várias destas funções não pressupõem a porcentagem de recombinação conjunta, ou seja, não são probabilidades de multilocos. O objetivo deste trabalho foi comparar, através de simulações, a eficiência na utilização de diferentes funções de mapeamento com e sem análise multiloco como ferramenta na construção de mapas genéticos. Para estudo comparativo foi simulado um genoma constituído de três grupos de ligação (50, 100 e 200 cM). Foram simuladas quatro populações de mapeamento, F2, com 50, 100, 200 e 400 indivíduos, e com 10 réplicas cada. Após análises dos resultados verificou-se que a análise multiloco não foi eficiente para resgatar o tamanho dos grupos de ligação, concluindo que a não utilização da análise multiloco seria mais viável.*

Palavras-chave: Estatística genômica, simulação, mapeamento.

REFERENCES

Bhering LL and Cruz CD (2008) Tamanho de população ideal para mapeamento genético em famílias de irmãos completos. **Pesquisa Agropecuária Brasileira 43**: 379-385.

Cruz CD and Schuster I (2001) **Programa GQMOL: Programa para análise de genética quantitativa e molecular v:** 2006.9.1.

Cruz CD (2006) **Programa Genes: Estatística experimental e matrizes**. Editora UFV, Viçosa, 285p.

Goldstein DR, Zhao H and Speed TP (1995) Relative efficiencies of c2 models of recombination for exclusion mapping and gene ordering. **Genomics 27**: 265-273.

Liu (1998) **Statistical genomics: linkage, mapping, and QTL analysis**. CRC Press, Boca Raton, 611p.

Schuster I and Cruz CD (2004) **Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. Editora UFV, Viçosa, 568p.

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. **The Plant Journal 3**: 793-744.