Journal of Animal and Veterinary Advances 11 (7): 933-938, 2012
ISSN: 1680-5593

AGPBSim: Gene Pyramiding Breeding in Animals from Perspective of Synthetic Biology

[1]Lingyang Xu, [1,2]Fuping Zhao, [1]Hangxing Ren, [1]Jian Lu, [1]Li Zhang,
[1]Caihong Wei and [1]Lixin Du
[1]Institute of Animal Science, Chinese Academy of Agricultural Sciences,
National Center for Molecular Genetics and Breeding of Animal, 100193 Beijing, China
[2]Department of Biostatistics, Tulane University, New Orleans, 70112 Louisiana, USA

Abstract: AGPBSim (Animal Gene-Pyramiding Breeding Simulation) is an individual-based, genetical information integrated simulation program. It was developed to investigate gene-pyramiding breeding given base population sizes, initial allele frequencies and selection strategies. The process of gene-pyramiding breeding was measured using population hamming distance, superior genotype frequency and average phenotypic values. AGPBSim is high flexible at various levels: four cross schemes and three selection strategies and trait architecture using various genotype-phenotype models were integrated in gene pyramiding breeding simulation. The GUI design of AGPBSim can facilitate design of gene-pyramiding breeding strategies by performing virtual breeding simulation on this platform.

Key words: Breeding, allele frequencies, population, architecture, gene-pyramiding, platform

INTRODUCTION

Gene pyramiding aims to design a superior trait through combining favorite target alleles into a single target genotype. Many quantitative trait loci and linked markers have been identified as the fast development of molecular dissection of complex traits. AGPBSim (Animal Gene-Pyramiding Breeding Simulation) was developed to integrate valuable molecular information in breeding program using marker-assisted gene-pyramiding design. This strategy can be considered as the using synthetic biology in the design of synthetic components (QTLs) of a desired specificity (traits) (De Lorenzo *et al*., 2006). This study aims to develop computational models to analyze, simulate and predict the behavior of artificial and synthetic systems (Animal breeding systems).

Marker-assisted gene pyramiding is an important branch of marker-assisted selection. The theoretical study of marker-assisted gene pyramiding study has just begun (Servin *et al*., 2004; Zhao *et al*., 2009). Gene pyramiding also belongs to the field of breeding by design which was proposed by Peleman and van der Voort (2003) with the goal of controlling all allelic variations for all genes of agronomic importance. AGPBSim is the first attempt to integrate the concept of gene-pyramiding design from an engineering perspective using evolution computation in soft computing techniques (Holland, 1992; Zadeh, 1994). Theoretically, breeding by design can be regarded as an optimization process and evolutionary computation is a subfield of artificial intelligence that involves combinatorial optimization problems. Researchers considered the artificial breeding process as an optimized program by building a model that links the genotype to phenotype. To maximize the phenotypic value, researchers selected the optimal genotype combinations in the breeding processes.

AGPBSim regards gene pyramiding breeding as an optimization process, individual carried various genotype is measured by the genotypic score and phenotypic value considering various selected strategies. Generation selection and genetic operator promoted the individual carried the optimal genotype combination and individual with optimal genotype combination responds to the optimal trait (Xu *et al*., 2011). AGPBSim implements two types of and selection strategies four types of cross schemes, including two population cross, three population cross, four population symmetry and cascading cross.

Various programs of gene-pyramiding design breeding, including different cross schemes and selection strategies under different initial favorite allele frequencies and base population sizes can be compared via the

Corresponding Author: Lixin Du, National Center for Molecular Genetics and Breeding of Animal, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, No. 2 Yuanmingyuan West Road, Haidian District, 100193 Beijing, P.R. China

program of AGPBSim. In addition, AGPBSim supplies the users unfamiliar with the command line through a simple and intuitive Graphical Interface (GUI). The simulation results such as population hamming distance, superior genotype frequency and phenotypic value can be computed and saved as the text files. Simulation results of visualization in AGPBSim can communicate information clearly and effectively through graphical means to users.

FEATURES

AGPBSim consists of four simulation components, two population cross, three population cross, four population symmetry and cascading cross. Each component corresponds to one type of cross breeding program with different target genes pyramiding numbers which can be easily used and extended according different reality.

Population: The individual's genotype at one locus was coded by 0 or 1. Initial base population were represented by N x M Matrix (N denoted the number of individuals in the base population and M/2 denoted the number of loci). The base population was initialized with random or fixed favorite alleles frequencies. In each generation, individuals in population were evaluated by genotypic score and phenotypic values under different selection strategies.

Discrete recombination and population hamming distance: Discrete recombination was used to combine (mates) two individuals (parents) to produce new offspring which was inspired by evolution computation (Goldberg, 1989; Holland, 1992). New offspring were produced by the crossover of two selected parents. Discrete recombination use crossover mask to indicate which parent will supply bits (allele) to the offspring a crossover mask was as the same length as the individual structure which was randomly generated by 0 or 1 with equal probability.

Crossover mask 1 indicates the allele of offspring at this locus inherited from parent 1, crossover mask 0 indicates the allele of offspring at this locus inherited from parent 2.

Discrete recombination at each locus produced offspring with new genotype combination. Offspring 1 was produced by mast 1 and offspring 2 was produced by mast 2, the allele inherited from parent 1 was marked with underline as follow:

Parent 1	0 1 1 1 0 0 1 1
Parent 2	1 0 1 0 1 1 0 0
Mask 1	0 1 1 0 0 0 1 1
↓	
Offspring 1	1 1 1 0 1 1 1 1
Mask 2	1 0 0 1 1 1 0 0
↓	
Offspring 2	0 0 1 1 0 0 0 0

Population Hamming Distance (PHD) derived from information theory denotes the total number of different alleles compared the population in each generation with ideal population (Pilcher *et al*., 2008). The PHD is zero indicating the fixation of favorite alleles at all loci. XOR is a type of logical disjunction on two populations (Eq. 1). For example of (Eq. 2), the population hamming distance is 19:

PHD = XOR (POP (t), POP (ideal)) = 19

PHD = XOR (population (I), population (ideal)) (1)

$$POP(t)=\begin{pmatrix}10&10&11&00\\01&11&00&01\\00&01&11&10\\01&11&10&00\\11&10&10&01\end{pmatrix}\quad POF(ideal)=\begin{pmatrix}11&11&11&11\\11&11&11&11\\11&11&11&11\\11&11&11&11\\11&11&11&11\end{pmatrix} \tag{2}$$

SELECTION

Three types of genotype-phenotype models were used in the studies (Fig. 1). Further described as follows: The Model I can be written as:

$$p_i = \mu_0 + \sum_{j=i}^{m} g_j x_{ij} + \varepsilon_i \tag{3}$$

Where:

p_i = The phenotypic value for the i individual (i = 1, 2,…,n)
μ_0 = The population mean
g_j = The gene effect at jth locus (j = 1, 2 ,…, m)
x_{ij} = A dummy variable indicating genotype
ϵ_i = Random residual effect, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

The values of genotypes were defined in terms of the midpoint (m), additive (a) and dominance (d) genetic parameters. The numerical coding of three genotypes 11, 10, 00 were 5, 4 and 1, respectively in the Model III. For an analysis of genotypes in a single environment, heritability on an individual basis will be estimated using the following equation:

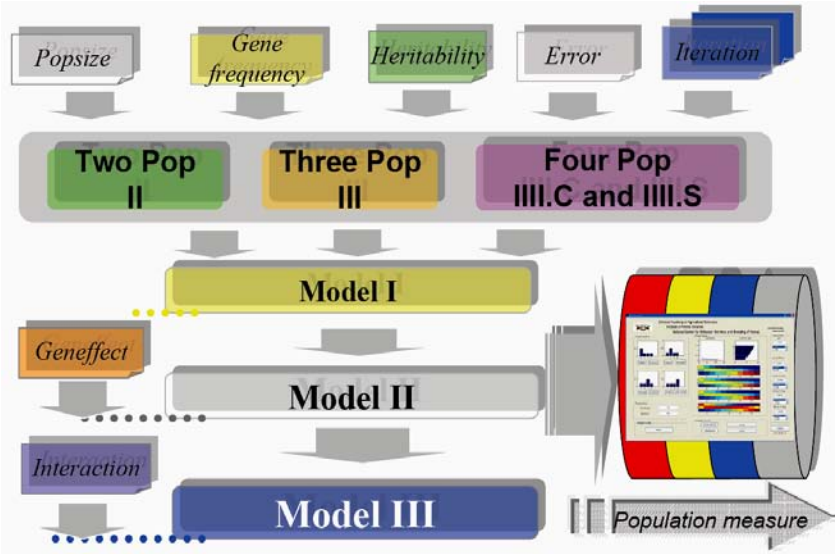

Fig. 1: Three types of genotype-phenotype models, model I is simple model with nothing integrated, model II, only gene effects were integrated and model III, both gene effects and gene interaction effects were integrated

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2} \quad (4)$$

From the defined heritability an estimate of σ_ε^2 is obtained by calculating σ_ε^2 and re-arranging Eq. 4 to:

$$\sigma_\varepsilon^2 = \frac{\sigma_g^2}{h^2} - \sigma_g^2 \quad (5)$$

For the genotype-phenotype Model II, only gene effects were integrated and Model III, both gene effects and gene interaction effects were integrated as following:

$$p_i = \mu_0 + \sum_{j=1}^{m} g_j x_{ij} + gim_{C(m, x_{ij})} + \varepsilon_i \quad (6)$$

Where the variable p_i, μ_0, x_{ij}, ϵ_i were denoted as the same as Model I. The difference between Model I and II are the values of genotypes that could be variously defined in Model 2 and g_j is the gene effect at jth locus (j = 1, 2, …, m).

Moreover, $gim_{(1\text{-}1,1\text{-}2,1\text{-}3,1\text{-}4)} \sim N(0, 1)$ is the genotype interaction effect, similar to ploygenic effect but represent the actual information integration in Model II. The value denotes genotype interaction effects using four-dimensional (4D) matrices and was sampled from $gim_{(1\text{-}1,1\text{-}2,1\text{-}3,1\text{-}4)}$ (supplementary file).

Gene pyramiding in cross program: Researchers designed four types of cross programs which were represented by II, III, III.C and IIII.S (Fig. 2). II represented pyramiding two target genes from popA and popB. The popA and popB were crossed to produce population popAB. The top 500 individuals based on phenotypic values were selected for the next generation and each pair of parents was assumed to produce four offspring with the sex ratio 1:1. Then the selection parents were randomly intercrossed to produce the subsequent generations until two target genes were pyramided into an ideal genotype. III represents pyramiding three target genes from popA, popB and popC, The popA and popB were hybrided to produce the hybrid population popAB, The initial population size of popC was set as 2×N then the top 500 of popAB and popC were hybrided to produce population popABC. Other breeding parameters and select strategies are as the same as schemes II. IIII represents pyramiding four target genes from popA, popB, popC and popD, other breeding parameters and select strategies are as the same as schemes II and III. For four population cascading cross (IIII.C), the base population size of popA, popB, popC and popD were N, N, 2×N and 4×N, PopA and popB were crossed to produce popAB, the top 500 of popAB and popC were crossed to produce population popABC than the top 500 of popABC crossed with popD to produce population popABCD. For symmetric cross scheme (IIII.S), the base population size of popA, popB, popC and popD were N, N, N and N, respectively, PopA and popB were crossed to produce popAB and popC and popD were crossed to produce popCD then the top 500 of popAB and the top 500 of popCD were crossed to produce popABCD in the next generation.

Iteration: Researchers performed Monte Carlo simulation for each gene pyramiding breeding programs.

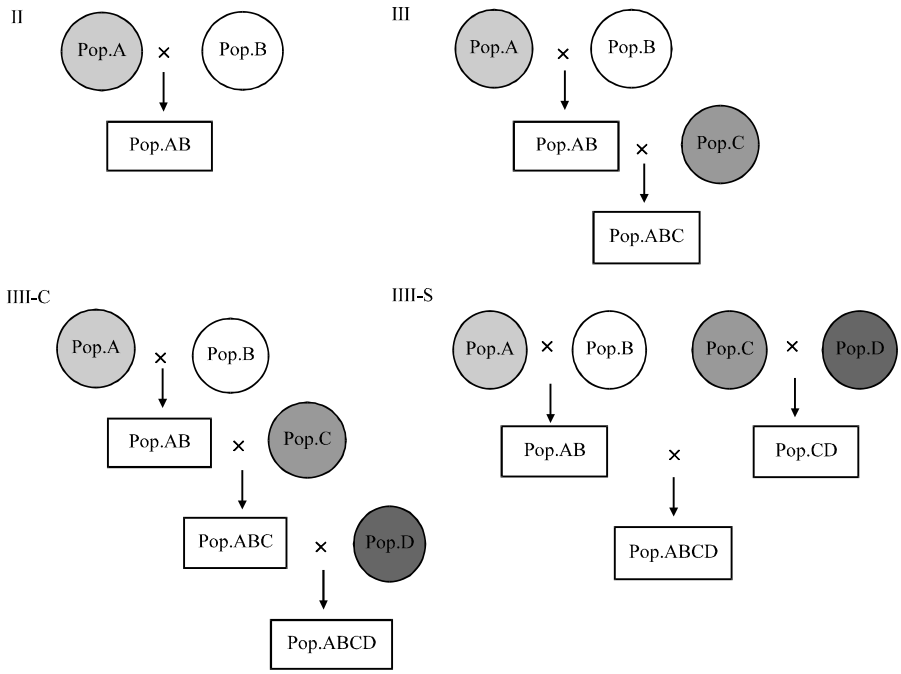

Fig. 2: Four types of cross programs which were represented by II, III, IIII.C and IIII.S

The computer programs and GUI design were implemented via Matlab and run on Microsoft Windows XP.

AGPBSim GUI design: AGPBSim GUI consists of one main interface and four sub-interfaces responding to four functional modules: twopop, threepop, fourpop-s and fourpop-c.

Input: AGPBSim can simulate the random population with various population sizes and alleles frequency levels and the frequencies of favorite genotype can be visualized by histograms and AGPBSim can also load data of population genotypes by text files which collect from the actual experiments. Trait heritability can be given input parameters range from 0-1 thus allowing for a wide variety of quantitative trait configuration.

Output: The gene pyramiding generation, population hamming distance and the superior genotype frequency and text files are computed and saved throughout simulation performed. The average population hamming distance and phenotypic values are calculated and presented into the edit text on interface. For iterative simulation, interface can list average phenotypic value and superior genotype frequencies at target loci to facilitate the comparison of trends over generations.

Table 1: Comparison of average phenotypic progress using phenotypic selection and genotypic selection

Cross scheme	Generation (t)	Phenotype selection 0.2	0.4	0.6	Genotype selection
II-A	7[1]	0.34[2]	0.72[3]	0.88[4]	0.87[5]
II-B	6	0.34	0.67	0.82	0.94
II-C	5	0.31	0.58	0.73	0.81
III-A	9	0.43	0.92	1.12	1.17
III-B	8	0.51	0.98	1.14	1.15
III-C	9	0.42	0.88	1.07	1.10
III-D	9	0.44	0.92	1.10	1.13
IIII-C-A	11	0.46	1.04	1.27	1.32
IIII-C-B	10	0.49	1.04	1.27	1.32
IIII-C-C	10	0.46	1.03	1.29	1.37
IIII-C-D	11	0.49	1.03	1.24	1.27
IIII-C-E	11	0.46	0.99	1.20	1.23
IIII-S-A	11	0.49	1.06	1.28	1.32
IIII-S-B	9	0.52	1.10	1.36	1.46
IIII-S-C	10	0.50	1.07	1.07	1.38
IIII-S-D	10	0.50	1.31	1.32	1.38

[1]The generation gene pyramided at using genotypic selection; [2]The average phenotypic progress over t generations using phenotypic selection with trait heritability 0.2; [3]The average phenotypic progress over t generations using phenotypic selection with trait heritability 0.4; [4]The average phenotypic progress over t generations using phenotypic selection with trait heritability 0.6; [5]The average phenotypic progress over t-generations using genotypic selection; [6]The average phenotypic progress calculated by [p (t) - p (1)]/t. p (t) denotes the average phenotype value at the generation St and p (1) denotes the average phenotype value at the generation 1

AGPBSim generates the panel of heat maps represents three types of genotype frequencies over nine generations as well as text files used to save the data of genotype frequencies.

A case study using AGPBSim: The average phenotypic progress for cross programs II, III, IIII-C and IIII-C were investigated using AGPBSim as a case study. Table 1 shows the average phenotypic progress using genotypic selection and phenotypic selection. The population size is 500, the detail information about allele frequencies for each cross programs were provided in additional table file.

Researchers first used genotypic selection to get the gene pyramiding generation G (t). Then at the generation t, researchers investigated the average phenotypic progress using phenotypic selection given different trait heritability (only model II is considered here), the average phenotypic progress was calculated by [p (t)-p (1)]/t, p (t) denotes the average phenotype value at the generation t and p (1) denotes the average phenotype value at the generation 1.

In the cross programs II, III and IIII, genotypic selection strategy is superior to phenotypic selection in accelerating gene pyramiding. The results show the trait with lower heritability is more appropriate for using genotypic selection to pyramid target genes. The phenotypic selection strategy for heritability 0.6 is as same as for genotypic selection strategy.

Compared the scheme IIII-C with IIII-S, the results of G (t) and average phenotypic progress show that IIII-S is superior to IIII-C.

The simulation also investigate cross order that influence on the schemes in cascading cross via calculating the value of average phenotypic progress and the scheme IIII.C-C is slightly superior to IIII.C-D and IIII.C-E.

DISCUSSION

AGPBSim regards gene pyramiding breeding as an optimization process inspired by the science of evolutionary computation (Goldberg, 1989; Holland, 1992), this is first attempt to use the optimization ideas to deal with the information integration in breeding practice. Gene pyramiding breed integrated with genetic information would facilitate the decision-making of breeders in breeding practice. Some bodies consider traditional mass selection strategies also result in gene pyramiding. Phenotypic selection strategy is used to investigate target gene which controlling quantitative trait and more over, researchers can compare the gene process of gene pyramiding using genotypic selection and phenotypic selection.

Initial favorite allele frequencies greatly affect the process of gene pyramiding breeding using phenotypic selection and another important factor is the trait heritability.

From the Table 1, researchers can conclude that for trait with high heritability, gene pyramiding breeding using phenotypic selection strategy needs less generation and more generation was needed when considering the low heritability trait. In order to achieve gene pyramiding successfully, breeder should select large size base population with high favorite allele frequencies. In phenotypic selection, researchers set trait heritability to 1 which is equivalent to genotypic selection derived from the formula 4 (Table 2-4).

Table 2: Allele frequencies in first/second loci in population A and B for II

Cross scheme	Population size	A1/A2[1]	B1/B2[2]
II-A	500	0.50/0.00	0.00/0.50
II-B	500	0.25/0.00	0.00/0.25
II-C	500	0.50/0.25	0.25/0.50

[1]Allele frequencies in first/second loci in population A; [2]Allele frequencies in first/second loci in population B

Table 3: Allele frequencies in first/second/third loci in population A, B and C for III

Cross schemes	Population size	A1/A2/A3[1]	B1/B2/B3[2]	C1/C2/C3[3]
III-A	500	0.50/0.00/0.00	0.00/0.50/0.00	0.00/0.00/0.50
III-B	500	0.00/0.00/0.50	0.00/0.25/0.00	0.00/0.00/0.25
III-C	500	0.50/0.25/0.00	0.00/0.50/0.25	0.25/0.00/0.50
III-D	500	0.25/0.00/0.00	0.00/0.25/0.00	0.00/0.00/0.50

[1]Allele frequencies in first/second/third loci in population A; [2]Allele frequencies in first/second/third loci in population B; [3]Allele frequencies in first/second/third loci in population C

Table 4: Allele frequencies in first/second/third/fourth loci in population A, B, C and D for IIII.S and IIII.C

Cross scheme	Population size	A1/A2/A3/A4[1]	B1/B2/B3/B4[2]	C1/C2/C3/C4[3]	D1/D2/D3/D4[4]
IIII.S-A	500	0.50/0.00/0.00/0.00	0.00/0.50/0.00/0.00	0.00/0.00/0.50/0.00	0.00/0.00/0.00/0.50
IIII.S-B	500	0.25/0.00/0.00/0.00	0.00/0.25/0.00/0.00	0.00/0.00/0.25/0.00	0.00/0.00/0.00/0.25
IIII.S-C	500	0.50/0.00/0.00/0.00	0.00/0.50/0.00/0.00	0.00/0.00/0.25/0.00	0.00/0.00/0.00/0.25
IIII.S-D	500	0.50/0.00/0.00/0.00	0.00/0.25/0.00/0.00	0.00/0.00/0.50/0.00	0.00/0.00/0.00/0.25
IIII.C-A	500	0.50/0.00/0.00/0.00	0.00/0.50/0.00/0.00	0.00/0.00/0.50/0.00	0.00/0.00/0.00/0.50
IIII.C-B	500	0.25/0.00/0.00/0.00	0.00/0.25/0.00/0.00	0.00/0.00/0.25/0.00	0.00/0.00/0.00/0.25
IIII.C-C	500	0.50/0.00/0.00/0.00	0.00/0.50/0.00/0.00	0.00/0.00/0.25/0.00	0.00/0.00/0.00/0.25
IIII.C-D	500	0.50/0.00/0.00/0.00	0.00/0.25/0.00/0.00	0.00/0.00/0.50/0.00	0.00/0.00/0.00/0.25
IIII.C-E	500	0.25/0.00/0.00/0.00	0.00/0.25/0.00/0.00	0.00/0.00/0.50/0.00	0.00/0.00/0.00/0.50

[1]Allele frequencies in first/second/third/fourth loci in population A; Âllele frequencies in first/second/third/fourth loci in population B; Âllele frequencies in first/second/third/fourth loci in population C; [4]Allele frequencies in first/second/third/fourth loci in population D

CONCLUSION

In this study, the results indicate that using genotypic selection is more superior for gene pyramiding than phenotypic selection. Design of cross scheme should concern the initial favorite allele frequency, cross order and the trait heritability. Trait heritability is the main factor that affecting the effective gene pyramiding breeding for the quantitative traits. When the genotypic value is preset, trait heritability would have a direct impact on the average phenotypic value predicted by the model and would finally affect the process of gene pyramiding. As to the trait with larger heritability, the dominant components in the model are the gene effects so, gene pyramiding breeding would be a process of select individual with the optimized genotype combination over generations.

AGPBSim was developed as an simulation platform for the gene-pyramiding breeding. On this platform, different level of population sizes, initial gene frequencies and flexible selection strategies can be designed and the progress of gene-pyramiding breeding can be predicted. In recent years, theoretical and experimental studies on system biology would provide a new perspective for understanding complex traits (Benfey and Mitchell-Olds, 2008; Sauer *et al*., 2007; Sieberts and Schadt, 2007). As the information from the analysis of complex phenotypes becomes more and more precise, the relationship between gene networks at a micro-level would also become more and more clear. Further development of accurate and practical models is necessary to link the genotype and phenotype in order to increase the accuracy of model prediction. Evolutionary computation technology will help exploit useful information and guide the precise optimal design of breeding by gene pyramiding.

Optimal models integrating useful genetic information would be developed in future studies as the development of system biology and high-throughout array technology. Moreover, different cross schemes and selection strategies can be designed and compared based on this gene pyramiding simulation platform.

ACKNOWLEDGEMENTS

The studies were supported by the National Natural Science Foundation of China (No. 30972094). Genetically Modified Organisms Breeding Major Projects (No. 2009ZX08008-003B) and National Modern Agricultural Industry Technology Fund for Scientists in Sheep Industry System.

REFERENCES

Benfey, P.N. and T. Mitchell-Olds, 2008. From genotype to phenotype: Systems biology meets natural variation. Science, 320: 495-497.

De Lorenzo, V., L. Serrano and A. Valencia, 2006. Synthetic biology: Challenges ahead. Bioinformatics, 22: 127-128.

Goldberg, D.E., 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison Wesley Publishing Co., Reading, Massachutes.

Holland, J.H., 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. The MIT Press, UK., ISBN-13: 9780262581110, pp: 211.

Peleman, J.D. and J.R. van der Voort, 2003. Breeding by design. Trends Plant Sci., 8: 330-334.

Pilcher, C.D., J.K. Wong and S.K. Pillai, 2008. Inferring HIV transmission dynamics from phylogenetic sequence relationships. PLoS Med., 5: e69-e69.

Sauer, U., M. Heinemann and N. Zamboni, 2007. Genetics. Getting closer to the whole picture. Science, 316: 550-551.

Servin, B., O.C. Martin, M. Mezard and F. Hospital, 2004. Toward a theory of marker-assisted gene pyramiding. Genetics, 168: 513-523.

Sieberts, S.K. and E.E. Schadt, 2007. Moving toward a system genetics view of disease. Mamm. Genome, 18: 389-401.

Xu, L.Y., H.X. Ren, X.H. Sheng, L. Zhang, C.H. Wei and L.X. Du, 2011. Selection for gene pyramiding design in admixed population. J. Anim. Vet. Adv., 10: 2421-2433.

Zadeh, L.A., 1994. Fuzzy logic, neural networks and soft computing. Commun. ACM, 37: 77-84.

Zhao, F.P., L. Jiang, H.J. Gao, X.D. Ding and Q. Zhang, 2009. Design and comparison of gene-pyramiding schemes in animals. Animal, 3: 1075-1084.