MATHEMATICA APPLICANDA
Vol. 43(1) 2015, p. 19–35
doi: 10.14708/ma.v43i1.629

Piotr Szulc* (Wrocław)

Localization of genes

Abstract Development of genetics in recent years has led to a situation in which we are able to look at the DNA chains with high precision and collect vast amounts of information. In addition, it turned out that the relationships between genes and traits are more complex than previously thought. These two things caused the need for close collaboration between geneticists and mathematicians whose task is to develop special methods, coping with specific and difficult genetic problems. The article includes an overview of both classic and the latest approaches to the problem of localizing genes that indicate places in the DNA chain, which significantly influence the traits of interest to us. Because of not the best communication between mathematicians and geneticists, knowledge of methods other than the classic among the latter group is still small.

2010 Mathematics Subject Classification: Primary: 62J05; Secondary: 92D20.

Key words and phrases: statistical genetics, quantitative trait loci, model selection, sparse linear regression, Bayesian Information Criterion.

1. DNA as the carrier of genetic information. Probably nobody has to be convinced about the huge diversity of living organisms on our planet. However, each form of life has a common structure made up of nucleotides (i.e. deoxyribose, a phosphate group and a nitrogenous base) called DNA. When we look closely at this molecule, we see that its exact composition depends on the species with which we are dealing; what is more, it is a kind of guide of how an organism is to be built. For this reason, we may be tempted to treat it as a measure of similarity between species. It is believed that the DNA of chimpanzee in 98% does not differ from the human. And can we find some similarities between man and something as different as yeast? It turns out that we share with them a quarter of genes.

1.1. Genes. What are genes? There is no simple answer to this question, at least at the present level of development of science. This is due to the fact that when this term was created, not much was known about DNA. A gene was understood as a theoretical unit of inheritance, that is something

* This publication is co-financed by the European Union as part of the European Social Fund within the project *Center for Applications of Mathematics* (project no. UDA–POKL.04.02.00–00–108/11–00).

Figure 1: Gen i SNP

that significantly affects the phenotype (set of features) of an individual and is passed down from generation to generation. Only later we tried to find a material object, which would correspond to the abstract entity. In textbooks we will find the answer to those searches: a gene is a piece of DNA, determining formation of one molecule of protein or RNA. In recent years, however, our confidence in understanding what we are dealing with has decreased. The gene seems to be something more complex, and therefore its definitions as well. We will hear voices that maybe it is even worth to give up this idea [8].

In this paper we will understand a gene as a segment of DNA which has a meaning (affecting a trait more or less indirectly), and which is present in at least two versions, so-called alleles. Depending on whether we have a gene in version A or a, it may result, for example, in a higher or lower risk of developing a disease.

1.2. Inter-individual differences in DNA. From this point we will be interested in inter-individual differences in the DNA. We focus on one genre and look for places that make two carrots or two people differ from each other. Such differences are smaller; DNA of two random people will most likely be the same in 99.9%. This per mille is however enough to find many differences between us (it is worth noting that also environment has impact on our features and it is actually not known what the proportions are).

At this point we have to make some distinction between finding genes in humans and other species. To do this, let us have a closer look at DNA structure. What we are most interested in are the nitrogenous bases. Usually they come in four versions: adenine, cytosine, guanine and thymine. Two DNA chains are different due to the fact that in the same location there are various nitrogen bases. In animals and plants we are generally looking for longer segments of DNA, which can occur in different versions, while in humans we most often consider each of nucleotides of an individual, and those in at least one percent of persons are different than the rest, so-called single nucleotide polymorphisms (Single Nucleotide Polymorphism, SNP). The Figure 1 presents schematically how a gene and SNP usually look.

1.3. Why look for genes? At the end of this paragraph we will answer the question, how the information about which places in the DNA are responsible for what could be useful. In humans, we can better understand the cause of the disease and thus develop a more effective medicament. We are also able to much more quickly assess risk and start the treatment earlier. In animals, such as cows, if we discover which genes are responsible for milk production, for example, we can interbreed only the appropriate individuals. Information about the location of a gene is also useful in the cultivation of fruit. If we want to grow in our orchard only sweet fruit, instead of for decades to cross different varieties, looking for the optimal characteristics, we can immediately use these with appropriate parameters [16].

2. General model. We would now like to go into mathematics and translate information about genotypes of an individual. We have identified alleles by A and a, which may seem unreasonable, because what symbol you could choose for the third allele? It turns out that this situation, i.e., the occurrence of a third or subsequent versions, is so unlikely that in general most often this opportunity is not included. This is due to the fact that a mutation in a DNA is rare, so next one in the same place hardly occurs. We could, therefore, encode the genetic information by only two numbers, except for the fact that DNA is in chromosomes which occur in pairs. In the corresponding chromosomes we do not have the same strings as one strand is inherited from a mother and the other from a father. Thus, in a given place within the DNA we have three choices: AA, aA (or AA, but the order is not important), or AA.

To locate the gene responsible for a trait of our interest, it would be best to know all the alleles. Unfortunately, usually we do not have such information and for the location we use so-called molecular markers. These are fragments of DNA which genotype we are able to determine experimentally. If the marker is located near a gene affecting the trait it is likely to be correlated with it and we will find it with the help of this particular marker. In natural populations this correlation is too low and finding the gene is not easy. Therefore we usually use experimental populations in which we intersect individuals closely related, so the correlation structure is much closer to our needs. Additionally, we are able to cross individuals (backcrossing) in such a way that in both chromosomes they have the same allele, that is, aa or AA, thanks to which later analysis becomes simpler [10, 11]. In humans inbreeding is not possible, therefore, due to low correlation we need a huge number of tightly packed markers [3].

In summary, for each individual we can indicate a sequence of genotypes (e.g. encoded as -1, 0 and 1) and the value of the trait of interest. Individual genotypes will be qualitative explanatory variables and the trait will be a dependent variable.

3. Tests in single markers. Our task is to identify which of the genes significantly influence the trait under consideration. And it is worth noting that, indeed, we will focus on locating them and the kind of dependence not necessarily concerns us. At the beginning let us try to approach this problem in the simplest possible way.

If we examine a quantitative trait, we can – by a suitable test – verify null hypothesis that the average value of a trait does not depend on the genotype of the marker. When its distribution does not differ significantly from normal, we often use the classical Student's t-test (if we consider only two versions of genotype) or F test for analysis of variance. If the distribution of a trait is not normal, we can apply the appropriate transformation, or instead of values of a trait consider ranks.

3.1. Linear regression. It is common practice in testing the significance of a given marker to use a linear regression model. We are trying to fit a model

$$Y_i = \beta_0 + \beta_j X_{ij} + \varepsilon_i \ , \quad i = 1, \ldots, n, \tag{1}$$

where ε_i is a random variable with the normal distribution, mean 0 and variance σ^2, while X_{ij} is the genotype of j-th marker. When it has only two values, for example aa i AA, commonly the following encoding is used:

$$X_{ij} = \begin{cases} -1/2, & aa \\ 1/2, & AA \end{cases} \tag{2}$$

The problem occurs when we consider three versions of genotypes, since then the relationships between numbers are important. Therefore, it is best to introduce an additional variable that will solve this problem. The following encoding is used most often:

$$X_{ij} = \begin{cases} 1, & aa \\ 0, & aA \\ -1, & AA \end{cases} \tag{3}$$

and

$$Z_{ij} = \begin{cases} -1/2, & aa \text{ or } AA \\ 1/2, & aA \end{cases} \tag{4}$$

More on encoding can be read at work [15]. The considered model is now in the form of

$$Y_i = \beta_0 + \beta_j X_{ij} + \gamma_j Z_{ij} + \varepsilon_i \ . \tag{5}$$

Using regression models, the null hypothesis presented earlier is now $\beta_j = 0$, or $\beta_j = \gamma_j = 0$. In order to verify this hypothesis we can apply in both cases the F-Snedecor test, in which we examine the ratio of the squares of residuals to the sum of squares explained by the model or the likelihood ratio test. When in the model we only have the X_{ij}, we can also use the Student's t-test, in which the value of the estimator $\hat{\beta}_j$ is divided by its standard deviation. We will not go into detail about these tests, because they are classic approach to study the significance of the regression coefficients. It can be also show that for the models considered by us, F-Snedecor test is equivalent to test F for analysis of variance (and the Student's t-test for the model with two genotypes is equivalent to F-Snedecor test).

3.2. The problem with multiple testing. When we use tests in individual markers, regardless of whether they are classic tests or linear regression approach, we face the problem of multiple testing: if we carry out a single test at the significance level α, then we have no guarantee that we will maintain this level performing more tests. For example, if we have a thousand markers, then performing tests at the level of 0.05 (and assuming that the marker genotypes are independent), we can expect about 50 false discoveries. This is not acceptable and therefore we apply corrections for multiple testing to control the probability of making at least one error of the first kind (Family Wise Error Rate, FWER). The simplest is the Bonferroni correction, in which each test is performed at the level of α/m, where m is the number of markers. Then we have the guarantee that FWER will not exceed α. This adjustment, however, becomes problematic, when the genotypes of the markers are strongly correlated, which in experimental populations is typical. Then the level of α/m is too low and it may happen that an essential gene escapes our attention. One solution is to use permutation tests [9], which adjust the critical value for the test to the correlation structure between the markers (in fact, between the values of the statistics). The procedure goes in such a way that we permute the vector Y several times, for each permutation we count values of test statistics and we find their maximum. As the critical value we take the $1-\alpha$ quantile of the distribution of the resulting maxima.

In case of the backcross we can use still another approach. Authors of [12] proposed to approximate the distribution of the likelihood ratio statistics throughout the chromosome using the square of Ornstein-Uhlenbeck process, thanks to which we will count (numerically) the critical value c for a single test from the following estimation:

$$\alpha \approx 1 - \exp\left[-2\left\{1-\Phi(\sqrt{c})\right\} - 0,04L\sqrt{c}\phi(\sqrt{c})\nu\left(\sqrt{0,04c\delta}\right)\right],$$

where δ is the distance between adjacent markers (in cM), L is the length of the chromosome (in cM), ϕ is the density of and Φ denotes the standard normal distribution, while $\nu(t)$ can be calculated form the formula

$$\nu(t) = 2t^{-2}\exp\left\{-2\sum_{n=1}^{\infty} n^{-1}\Phi(-|t|n^{1/2}/2)\right\}.$$

A new unit of length has appeared here, centimorgan (cM). At the distance of one centimorgan, the expected number of recombination, i.e. the exchange of genetic material, is 1%.

Yet another solution to the problem of low level of significance for a single test is a little different approach: we accept the fact that the false discoveries will occur, as long as they are not much in relation to all discoveries. If it suits us, we can apply the Benjamini-Hochber correction [4]. We count the p-values for tests in each marker, we sort them non-decreasingly: $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[m]}$ and we calculate index k_F according to the formula

$$k_F := \max\left\{i:\ p_{[i]} \leq \frac{i\alpha}{m}\right\}.$$

Then, we reject k_F hypotheses with the p-values less than or equal to $p_{[k_F]}$. The procedure may seem strange, but it was shown that it controls at a level not exceeding α the so-called fraction of false discoveries (False Discovery Rate, FDR), i.e.

$$FDR = E\left(\frac{V}{R}|R > 0\right),$$

where R is the number of all rejected hypotheses, and V is the number of false rejections.

4. Multiple regression. The main problem of testing in single markers is the fact that we completely ignore the impact of other markers. If more genes have connection with a trait (it is usually true), it is a better idea to attempt to fit a model that contains all these essential genes. In addition, genes may interact with each other. All of this can be modeled using the multiple regression. If we consider only interactions of second order, then a model for the case of two versions of genotypes is of the form of

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + \sum_{1 \leq j < l \leq m} \gamma_{jl} X_{ij} X_{il} + \varepsilon_i. \tag{6}$$

In practice, because m is large, we limit ourselves just to interactions of second order, and sometimes give them up at all.

4.1. Model selection criteria. If we already decide to apply the multiple regression, we need to establish criterion which particular model we want to consider as the best. It is known that adding more variables will certainly not make worse the fitting to the date, so we need to decide on a compromise between the fitting and the number of variables. For this purpose, we can use the model selection criteria, which take into account a penalty for the size. It turns out, however, that the classic criteria as AIC [1] or BIC [20] are not suitable for this purpose because they overestimate the number of significant variables [6].

Why is this happening? Shortly speaking, while deriving the BIC, the element responsible for the a priori distribution for the considered model is omitted, thereby we assume that everyone is equally probable. Unfortunately, as a result, we in fact prefer models with sizes closer to $m/2$, because they are most numerous (if we assume the uniform distribution on models then the distribution for the size of a model is binomial $B(m, \frac{1}{2})$). Because traits most frequently depend on a small number of genes (e.g. a dozen or so), or at least we are looking for such small models, the described property of BIC criterion is not acceptable. And since BIC is rather known as a conservative criterion, which usually choose models of small size, we need to look around for something else.

4.2. Modifications of BIC. One idea is to replace the uniform distribution on models with different distribution, so that we will get a criterion with desired properties. This led to establishing the criteria mBIC [5] and EBIC [7]. In the first one, by using a suitable a priori distribution, we still obtain binomial distribution on a model size, but the probability of success equal to E_m/m where E_m is the expected value of the significant variables. The criterion minimizes the expression

$$mBIC = n\ln(RSS) + p\ln n + 2p\ln\left(\frac{m}{E_m} - 1\right), \qquad (7)$$

where RSS is the residual sum of squares, p is the size of the relevant model. The difference with the standard BIC is the addition of the last element. In the event that we do not have any expectations to the number of relevant variables, it was shown that for typical sample sizes the choice of $E_m = 4$ results in the fact that the total type I error is at a level close to 0.05 [5]. It was also shown that the criterion is consistent in a situation where both n and m tend to infinity [22].

For EBIC we assume that the a priori distribution a on model is proportional to $\binom{m}{p}^{\kappa-1}$ for some κ greater than zero, with the result that the minimized expression is

$$EBIC = n\ln RSS + p\ln n + 2(1-\kappa)\ln\binom{m}{p}. \qquad (8)$$

The κ parameter should be chosen by ourselves, taking into account that for $\kappa = 1$ we will receive the usual BIC, while taking $\kappa = 0$ we assume that the a priori distribution on model size is uniform. It was also shown that the criterion is consistent when n and m tend to infinity [18].

The above criteria apply to a situation in which we neglect the interaction effects. Nevertheless, their inclusion is not a problem and the appropriate expression to minimize takes the form of

$$mBIC = n\ln RSS + (p+q)\ln n + 2p\ln\left(\frac{m}{E_m} - 1\right) + 2q\ln\left(\frac{N_e}{E_e} - 1\right), \qquad (9)$$

where q is the number of interactions in the considered model, N_e is the number of all interactions (amounting to $\binom{m}{2}$, if we consider only interactions of second order), and E_e is the expected number of interactions. It was shown that if we do not have any expectations, adoption of $E_m = N_e = 2.2$ should ensure control over FWER at the level of 0.05 [5]. Similarly, we can specify the version with interactions for EBIC.

When the genotype of the marker can occur in three versions, we can simply use the formula above, substituting $2m$ in place of m, i.e Z_{ij} are treated as additional markers [2].

If controlling FWER is not so important for us, then similarly to single marker tests we can focus on the control of FDR. An appropriate criterion in this case is mBIC2 [13], defined as (10).

$$mBIC2 = n \ln RSS + p \ln n + 2p \ln \left(\frac{m}{E_m} - 1 \right) - 2 \ln p!. \tag{10}$$

Of course, after appropriate modifications, this criterion can be also used in a situation when we take into account interactions and genotypes in three possible versions.

These criteria can also be used if the trait distribution is not continuous, e.g. in the logistic or Poisson regression. One should then replace the term $n \ln RSS$ with minus doubled logarithm of the maximum likelihood function for a given model [23].

Finally, we should say a few words about how to use the above-mentioned criteria in practice. Typically, the number of markers is large enough (when we consider the SNPs of about 500 thousand), that we are not able to consider all the possible models to choose the best. Therefore, at the outset we perform tests in individual markers and organize the variables according to the statistics' values. Then we can use the forward or backward selection procedures, giving priority to those markers that from the initial analysis seem to have the greatest impact on a trait. We can also completely reject the markers with low values of statistics, since it is very unlikely to have them included in the final model, and thanks to this we are able to offer a good model in finite time [14].

Readers interested in expanding their knowledge about statistical approach to the location of genes are referred to the book [19].

5. Summary. Classic and new approaches to the problem of localizing genes have been presented. When using conventional methods we encounter a number of problems which we are able to cope with only partially. Simulations and analysis of real data show that the new methods behave better: they enable to find genes that would escape our attention during classic proceedings [7,14,23] and allow to design models closer to reality, for example by taking into account interactions [5]. These methods are still being developed and adjusted to the genetic data, which character is even more specific [17,21].

References

[1] H. Akaike, *A new look at the statistical model identification.* IEEE Transactions on Automatic Control **19** (1974): 716–723. doi: 10.1109/TAC.1974.1100705 MR 0423716 Zbl 0314.62039

[2] A. Baierl, M. Bogdan, F. Frommlet, A. Futschik. *On locating multiple interacting Quantitative Trait Loci in intercross designs.* Genetics **173** (2006): 1693–1703. doi: 10.1534/genetics.105.048108 PMID: 16624924 [PubMed]

[3] D.J. Balding, *A tutorial on statistical methods for population association studies.* Nature Reviews Genetics **7** (2006): 781–791. doi: 10.1038/nrg1916

[4] Y. Benjamini, Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. Roy. Statist. Soc. Ser. B. **57** (1995): 289–300. MR 1325392 Zbl 0809.62014

[5] M. Bogdan, J.K. Ghosh, R.W. Doerge, *Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci.* Genetics **167** (2004): 989–999. doi: 10.1534/genetics.103.021683

[6] K.W. Broman, T.P. Speed, *A model selection approach for the identification of quantitative trait loci in experimental crosses.* Journal of the Royal Statistical Society: Series B **64** (2002): 641–656. Zbl 1067.62108 doi: 10.1111/1467-9868.00354

[7] J. Chen, Z. Chen, *Extended Bayesian Information criteria for model selection with large model spaces.* Biometrika **95**(3) (2008): 759–771. Zbl 05609546 doi: 10.1093/biomet/asn034

[8] M. Chorąży, *Gen strukturalny – ewolucja pojęcia i dylematy.* Nauka **3** (2009): 57–108.

[9] G.A. Churchill, R.W. Doerge, *Empirical threshold values for quantitative trait mapping.* Genetics **138** (1994): 963–971. PMID: 7851788 [PubMed]

[10] R.W. Doerge, *Mapping and analysis of quantitative trait loci in experimental populations.* Nature Reviews Genetics **3** (2002): 43–52. PMID: 11823790 [PubMed]

[11] R.W. Doerge, Z-B. Zeng, B.S. Weir, *Statistical issues in the search for genes affecting quantitative traits in experimental populations.* Statistical Science **12** (1997): 195–219. doi: 10.1214/ss/1030037909

[12] J. Dupuis, D.O. Siegmund, *Statistical methods for mapping quantitative trait loci from a dense set of markers.* Genetics **151** (1999): 373–386. PMID: 9872974 [PubMed]

[13] F. Frommlet, A. Chakrabarti, M. Murawska, M. Bogdan, *Asymptotic Bayes optimality under sparsity for general distributions under the alternative*, Technical Report (2011), arXiv:1005.4753v2.

[14] F. Frommlet, F. Ruhaltinger, P. Twaróg, M. Bogdan, *A model selection approach to genome wide association studies.* Computational Statistics and Data Analysis **56** (2012): 1038–1051. doi: 10.1016/j.csda.2011.05.005

[15] C-H. Kao, Z-B. Zeng, *Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model.* Genetics **160** (2002): 1243–1261. PMID: 11901137 [PubMed]

[16] S. Keller-Przybyłkowicz, M. Korbin, *Lokalizacja genów wpływających na jakość jabłek na mapie referencyjnej genomu jabłoni.* Zeszyty Naukowe Instytutu Sadownictwa i Kwiaciarstwa **16** (2008): 69–81.

[17] W. Li, Z. Chen, *Multiple interval mapping for quantitative trait loci with a spike in the trait distribution.* Genetics **182**(2) (2009): 337–342. doi: 10.1534/genetics.108.099028