Scientific Research Publishing

**Open Journal of Statistics, 2021, 11, 1017-1025**
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# An Introduction to Basic Statistical Models in Genetics

**Tapshir Jahan Setu, Tapati Basak**

Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh
Email: tafsirjahan43@gmail.com, tapati.basak@juniv.edu

**How to cite this paper:** Setu, T.J. and Basak, T. (2021) An Introduction to Basic Statistical Models in Genetics. *Open Journal of Statistics*, **11**, 1017-1025.
https://doi.org/10.4236/ojs.2021.116060

**Received:** November 13, 2021
**Accepted:** December 19, 2021
**Published:** December 22, 2021

## Abstract

The use of the three genetic models viz. additive, dominant and recessive in Genome-wide association study (GWAS) is a common and powerful approach to study the association between genetic variants and a trait (disease). The selection of these models depends on the pattern of inheritance and the scope of the study. GWAS typically focuses on single-nucleotide polymorphism (SNPs) and common human diseases in a case-control setup. In order to study this type of association between the risk genotype and the phenotype for a given inheritance pattern, the use of these genetic models helps to identify the disease risk appropriately. This study provides an overview of the existing genetic models (additive, dominant and recessive) and a practical demonstration of these model tests for the contingency tables of SNP genotypes and the disease phenotypes in a case-control setting.

## Keywords

Genetic Model, Association, GWAS, SNP, Case-Control Study

## 1. Introduction

The main goal of human genetics is to identify genetic risk factors for common and complex diseases [1] [2] [3] [4] [5]. The risks related to allelic variants of candidate genes for which there is evidence of linkage to disease susceptibility are determined [4] [6]. These studies collect valid and precise information on the causes, prevention, and treatment of disease [6].

The genetic association studies such as genome-wide association study (GWAS) is a powerful and complete analysis of the genetic association between certain observable traits and specific genetic variations in the form of Single Nucleotide Polymorphisms (SNPs). GWAS provides a relatively superficial approach to detect potential genetic contributors to phenotypes (common and complex diseases) from a simple case-control setup [1] [3] [7]. These studies attempt to discover novel genes by testing huge number of SNPs for association [3].

The statistical analysis of genetic data can be performed for a study population when a well-defined phenotype is selected, and the genotypes are collected using a sound technique [4]. GWAS perform a series of single-locus statistic tests and examine the susceptibility of each SNP independently for association to the phenotype [1] [4].

The genotypic association tests examine the association between genotypes and the phenotype, where the genotypes for a SNP can also be grouped into different genotype models, such as additive, dominant or recessive models [4] [5].

The main objective of this paper is to provide a practical demonstration of three basic genetic models (additive, dominant, recessive) in the case-control GWAS studies for DNA sequencing data.

## 2. Genetic Models

The existing three genetic models can be rephrased as following.

For a single SNP, the 3 genotypes together with a categorical phenotype with two categories can be presented in a 2 × 3 contingency table (**Table 1**). The counts in the table $\left(n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}\right)$ are the numbers of samples in a case-control with a particular genotype and phenotype combination, where the SNP has two alleles ($D$ = disease-causing allele and $N$ = allele not causing the disease).

Each model makes different assumptions about the genetic effect in the data. For a single SNP with the two alleles, $N$ and $D$, the dominant model (for $D$ allele) assumes that having one or more copies of the $D$ allele increases risk compared to $N$. Hence, the genotypes $DD$ or $ND$ have the higher risk. In case of the recessive model (for $D$ allele), the assumption is two copies of the $D$ allele are required to alter the risk. Hence, the individuals with the genotype $DD$ are compared to individuals having genotypes $ND$ and $NN$. A linear and uniform increase is assumed based on the number of each copy of the disease-causing allele ($D$). Thus, the additive model (for $D$ allele) assumes, if the risk for $ND$ is $k$ then the risk for $DD$ is $2k$ [4] [8] [9].

### Models with the Penetrance Function

Penetrance functions represent one approach to modeling the relationship between SNPs and risk of disease [10] [11] [12]. The penetrance of a genetic disorder is measured by evaluating how often a particular phenotype occurs given a

**Table 1.** A 2 × 3 table of genotype counts for a single SNP in a case control study.

| | ***NN*** | ***ND*** | ***DD*** |
|---|---|---|---|
| Case | $n_{11}$ | $n_{12}$ | $n_{13}$ |
| Control | $n_{21}$ | $n_{22}$ | $n_{23}$ |

particular genotype. This measures the conditional probability $P(x \mid g)$ of being affected with disease $x$ given a specific genotype *g*. Now, the probabilities of being affected depending on a disease-causing genotype with one disease-causing allele $D$ and one allele not causing the disease *N*, can be expressed as [13] [14],

$$f_0 = P(\text{affected} \mid NN), \quad f_1 = P(\text{affected} \mid DN), \quad f_2 = P(\text{affected} \mid DD) \tag{1}$$

Here, $f_0$ is the frequency of individuals who are affected without carrying a disease-causing allele (frequency of phenocopies).

According to Bush (2012), different inheritance patterns (recessive, dominant, additive) can be expressed in terms of mathematical models (**Table 2**). Here, the phenotypes show full penetrance and no phenocopies. That is, no individual without the disease-causing genotype will become affected.

For example, if a disease is transmitted in an additive fashion, the risk for a heterozygous person to be affected is half that of the person who is homozygous $D$ as compared to an individual who is homozygous *N*. Hence, according to the penetrance probabilities shown in **Table 2**, $f_1 = (f_0 + f_2)/2$.

On the other hand, these models could be represented with respect to the genotypic relative risks (GRR) under the assumption of phenocopies that is $f_0 > 0$ (**Table 3**).

For $f_0 > 0$, the GRR can be expressed in terms of the functions $f_0, f_1$ and $f_2$ defined in Equation (1),

$$\gamma_1 = \text{GRR}_1 = \frac{f_1}{f_0} = \frac{P(x \mid DN)}{P(x \mid NN)}, \quad \gamma_2 = \text{GRR}_2 = \frac{f_2}{f_0} = \frac{P(x \mid DD)}{P(x \mid NN)} \tag{2}$$

So, the GRR presents the increased risk of an individual having a disease causing genotype over a person without disease-causing allele. By introducing the GRR, the three parameters ( $f_0, f_1, f_2$ ) defined in Equation (1) are reduced to

**Table 2.** Penetrances for simple Mendelian inheritance patterns.

| Genotype | Genetic model | | | |
|---|---|---|---|---|
| | General | Recessive | Dominant | Additive |
| *NN* | $f_0$ | 0 | 0 | 0 |
| *DN* | $f_1$ | 0 | 1 | 1 |
| *DD* | $f_2$ | 1 | 1 | 2 |

**Table 3.** Genotype relative risks under the assumption of phenocopies.

| Genotype | GRR | Genetic Model | | |
|---|---|---|---|---|
| | | Recessive | Dominant | Additive |
| *DD* | $\gamma_2$ | $\gamma$ | $\gamma$ | $2\gamma - 1$ |
| *DN* | $\gamma_1$ | 1 | $\gamma$ | $\gamma$ |
| Restriction | | $\gamma_1 = 1$ | $\gamma_1 = \gamma_2$ | $\gamma_2 = 2\gamma_1 - 1$ |

the two parameters ($\gamma_1$ and $\gamma_2$). For an additive model, the risk could be expressed as $\gamma_2 = 2\gamma_1 - 1$ (Table 3).

## 3. Genotype Data Preparation

The individual SNP genotype data for single SNPs were generated for 1000 individuals via computer simulation in R-programming language. Then, these 1000 individuals were randomly allocated to the cases and the controls with the equal probability of cases (0.5) and controls (0.5). This random allocation was repeated for 1000 times. The independence test of single SNP was performed in each repetition using the proportion trend test [15] for the three genetic models (additive, dominant and recessive) and the Pearson chi-squared test [16]. The three *p*-values were recorded from the independence tests of the three genetics models along with the *p*-value from the Pearson chi-squared test in each repetition.

## 4. Results and Discussion

Figure 1 is presenting the histograms of the *p*-values obtained from the four types of independence tests using three genetic models (additive, dominant,

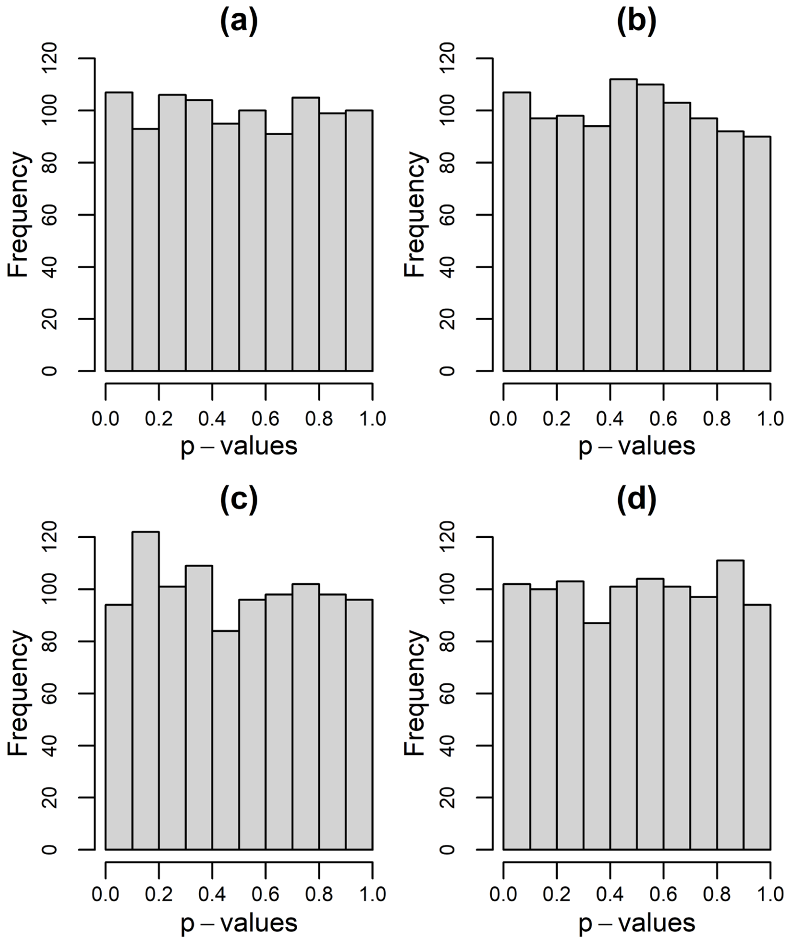

Figure 1. The histogram of the *p*-values from the three genetic model tests and the Pearson chi-square test. (a) Pearson chi-square; (b) Additive; (c) Dominant; (d) Recessive.

recessive) along with the Pearson chi-squared test. Apparently, a flat shaped distribution is observed over the shape of the four histograms shown in **Figure 1**. That is, the *p*-values from each of the independence tests have the uniform distribution under the null hypothesis. The test of uniformity using the Kolmogorov-Smirnov (K-S) test also implies that the *p*-values from each test follow the uniform distribution. The obtained *p*-values from the K-S test are 0.835, 0.689, 0.796 and 0.6255, for the additive, dominant, recessive models and the Pearson chi-squared test, respectively.

But, the critical examination of the **Figure 1** implies that not all the null hypothesis are actually true. A little fluctuation in the heights of the bars in each histogram indicates that there is a small percentage of null hypothesis that are not true (non-null). Different existing correction methods could be applied here in order to control such false discovery rate (FDR).

The *p*-values from each of the three genetic model tests were plotted against the chi-square ( $\chi^2$ )-values along with the Pearson chi-squared test (**Figure 2**). All of the plots are showing that the *p*-values are getting smaller for increasing values of the corresponding $\chi^2$ -statistic.

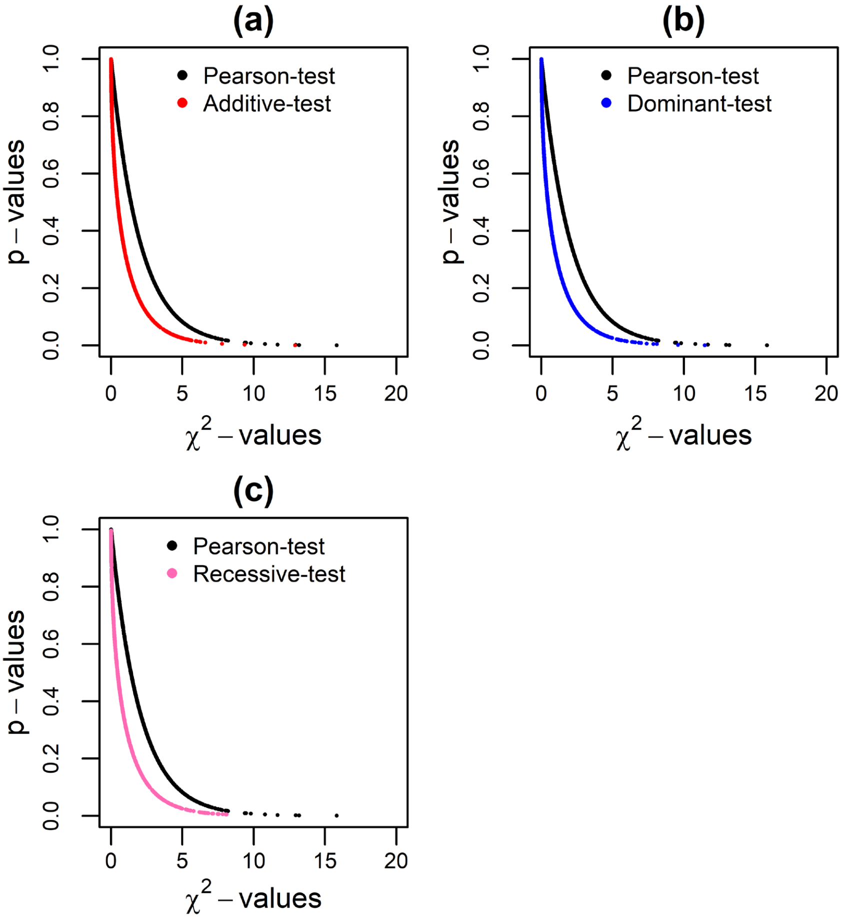

**Figure 2.** The plot of the *p*-values from the three genetic model tests along with the Pearson chi-square test. (a) Additive and Pearson chi-square; (b) Dominant and Pearson chi-square; (c) Recessive and Pearson chi-square.

Apparently, the curve shapes and features of the three tests are seems to be the similar with the Pearson chi-squared test. But, the differences in the results are observed by investigating the Figure 3. Figure 3 is presenting the pairwise difference plots between the *p*-values and $\chi^2$-values of each of the three tests with the Pearson chi-squared test. A positive relation is observed in each of the plot, where many values are grouped together near the origin. This is because, the tables corresponding to these cases have relatively smaller deviations from the Pearson chi-squared test in terms of the *p* and $\chi^2$-values.

On the other hand, the 3-dimensional scatter plot of the *p*-values from the three genetic tests in Figure 4 is indicating that the three genetic tests are producing different *p*-values having a positive relation among them for different tables obtaining from shuffling of the phenotypes.

The result shows, a table with the fixed genotype counts are producing different results while applying the different genetic tests. Also, for a fixed sample size,

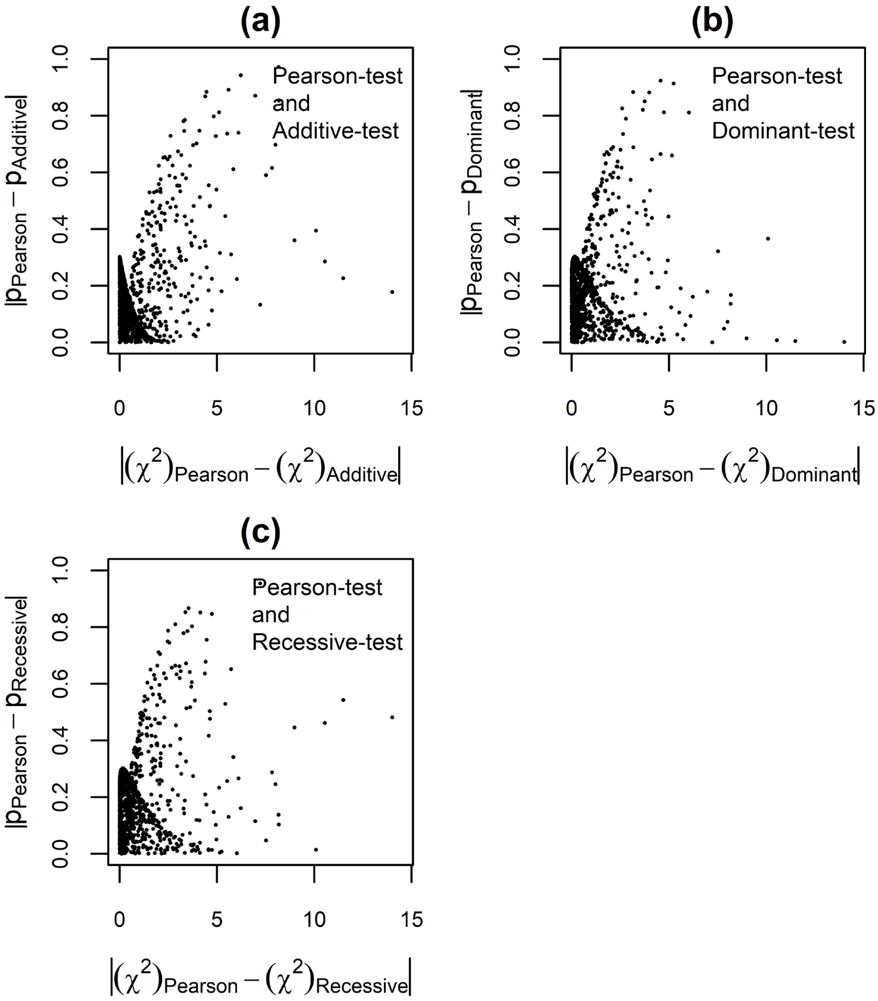

Figure 3. The plot of the absolute pairwise differences between the *p*-values and $\chi^2$-values of each of the three genetic model tests with the Pearson chi-squared test. (a) Additive and Pearson chi-square; (b) Dominant and Pearson chi-square; (c) Recessive and Pearson chi-square.

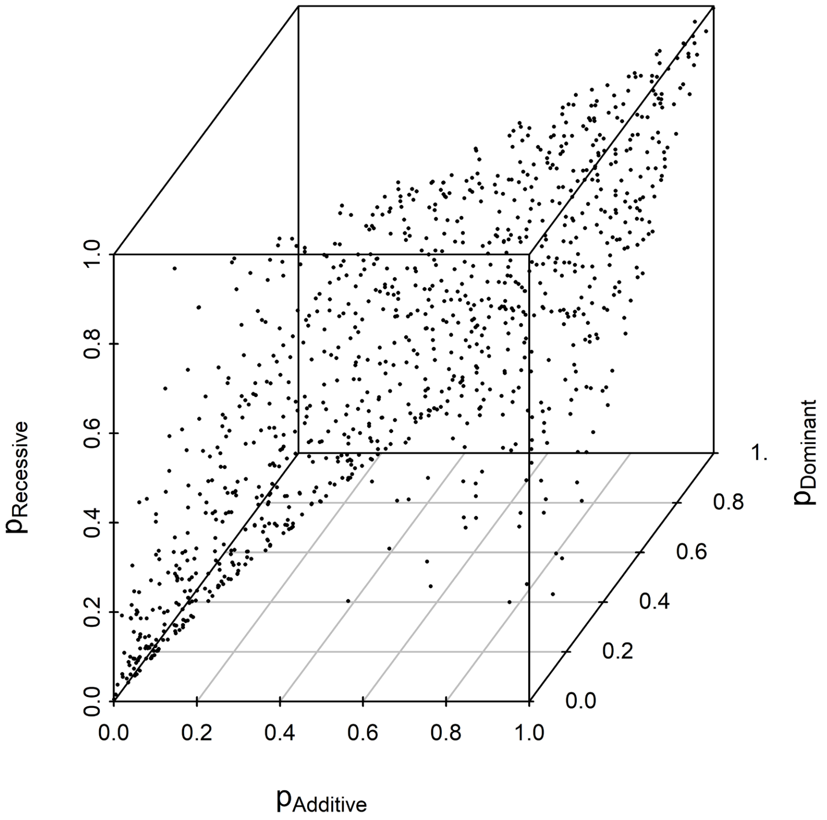

Figure 4. The relation among the *p*-values of the three genetic model tests (Additive, Dominant and Recessive).

the application of a particular genetic test are resulting different *p*-values for the tables that are producing by shuffling the phenotypes.

## 5. Conclusion

This paper is a practical demonstration of the three genetic model tests for the SNP genotype data. Here, the simulated SNP genotype data used in the analysis. But, this application could be extended for the real datasets. The basic structure of both the simulated and real data would be the same. So, the directions of the results would be the same for both the cases. On the other hand, the choice of a proper model is important in such association studies, which generally depends on the inheritance pattern of a disease. So, the investigation of the suitability of these models depending inheritance patterns of disease would be the future directions of this research. The appropriate selection of genetic model in association studies will enhance to detect the risks related to allelic variants of candidate genes. The result of this paper indicates that different genetic model tests are producing different *p*-values for a table of fixed sample size and genotype counts. Also, for the same test, different *p*-values are obtaining for all the tables while the tables were constructed by the shuffling of the phenotypes of the given table. Hence, the models should be correctly chosen according to the mode of inheritance (dominant, additive and recessive).

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Patron, J., Serra-Cayuela, A., Han, B., Li, C. and Wishart, D.S. (2019) Assessing the Performance of Genome-Wide Association Studies for Predicting Disease Risk. *PLoS ONE*, **14**, Article ID: e0220215. https://doi.org/10.1371/journal.pone.0220215

[2] Mills, M.C. and Rahal, C. (2019) A Scientometric Review of Genome-Wide Association Studies. *Communications Biology*, **2**, Article No. 9. https://doi.org/10.1038/s42003-018-0261-x

[3] Shaffer, J.R., Feingold, E. and Marazita, M.L. (2012) Genome-Wide Association Studies: Prospects and Challenges for Oral Health. *Journal of Dental Research*, **91**, 637-641. https://doi.org/10.1177/0022034512446968

[4] Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, **8**, Article ID: e1002822. https://doi.org/10.1371/journal.pcbi.1002822

[5] Lewis, C.M. (2002) Genetic Association Studies: Design, Analysis and Interpretation. *Briefings in Bioinformatics*, **3**, 146-153. https://doi.org/10.1093/bib/3.2.146

[6] Sahebi, L., Dastgiri, S., Ansarin, K., Sahebi, R. and Mohammadi, S.A. (2013) Study Designs in Genetic Epidemiology. *International Scholarly Research Notices*, **2013**, Article ID: 952518. https://doi.org/10.5402/2013/952518

[7] Khandaker, L., Akond, M., Liu, S., Kantartzi, S.K., Meksem, K., Bellaloui, N., Lightfoot, D.A. and Kassem, M.A. (2015) Mapping of QTL Associated with Seed Amino Acids Content in “MD96-5722” by “Spencer” RIL Population of Soybean Using SNP Markers. *Food and Nutrition Sciences*, **6**, 974-984. https://doi.org/10.4236/fns.2015.611101

[8] Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P. and Zondervan, K.T. (2011) Basic Statistical Analysis in Genetic Case-Control Studies. *Nature Protocols*, **6**, 121-133. https://doi.org/10.1038/nprot.2010.182

[9] Horita, N. and Kaneko, T. (2015) Genetic Model Selection for a Case-Control Study and a Meta-Analysis. *Meta Gene*, **5**, 1-8. https://doi.org/10.1016/j.mgene.2015.04.003

[10] Moore, J.H., Hahn, L.W., Ritchie, M.D., Thornton, T.A. and White, B.C. (2004) Routine Discovery of Complex Genetic Models using Genetic Algorithms. *Applied Soft Computing*, **4**, 79-86. https://doi.org/10.1016/j.asoc.2003.08.003

[11] Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. and Kehrer-Sawatzk, H. (2013) Where Genotype Is Not Predictive of Phenotype: Towards an Understanding of the Molecular Basis of Reduced Penetrance in Human Inherited Disease. *Human Genetics*, **132**, 1077-1130. https://doi.org/10.1007/s00439-013-1331-2

[12] Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D.T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M.D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T.R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyfjord, J., Lynch, H., Ponder, B.A.J., Gayther, S.A., Birch, J.M., Lindblom, A., Stoppa-Lyonnet, D., Bignon, Y., Borg, A., Hamann, U., Haites, N., Scott, R.J., Maugard, C.M., Vasen, H., Seitz, S., Cannon-Albright, L.A., Schofield, A., Zelada-Hedman, M. and The Breast Cancer Linkage Consortium (1998) Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Can-

cer Families. *American Journal of Human Genetics*, **62**, 676-689. https://doi.org/10.1086/301749

[13] Ziegler, A. and König, I.R. (2010) A Statistical Approach to Genetic Epidemiology: Concepts and Applications. 2nd Edition, Wiley-VCH, Weinheim.

[14] Gong, G., Hannon, N. and Whittemore, A.S. (2010) Estimating Gene Penetrance from Family Data. *Genetic Epidemiology*, **34**, 373-381. https://doi.org/10.1002/gepi.20493

[15] Armitage, P. (1955) Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, **11**, 375-386. https://www.jstor.org/stable/3001775 https://doi.org/10.2307/3001775

[16] Plackett, R.L. (1983) Karl Pearson and the Chi-Squared Test. *International Statistical Review*, **51**, 59-72. https://www.jstor.org/stable/1402731 https://doi.org/10.2307/1402731