Meta Gene 9 (2016) 225–229

Contents lists available at ScienceDirect

Meta Gene

journal homepage: www.elsevier.com/locate/mgene

An enhanced version of Cochran-Armitage trend test for genome-wide association studies

Mansi Ghodsi [a], Saeid Amiri [b], Hossein Hassani [c,*], Zara Ghodsi [a]

[a] Translational Genetics Group, Bournemouth University, UK
[b] University of Wisconsin-Green Bay, Department of Natural and Applied Sciences, Green Bay, WI, USA
[c] Institute for International Energy Studies, Tehran, 1967743 711, Iran

ARTICLE INFO

Article history:
Received 2 April 2016
Revised 30 June 2016
Accepted 1 July 2016
Available online 22 July 2016

Keywords:
Bootstrap method
Monte Carlo simulation
Chi-squared test
Contingency table
Genetic association
p-values

ABSTRACT

Genome-wide association studies the evaluation of association between candidate gene and disease status is widely carried out using Cochran-Armitage trend test. However, only a small number of research papers have evaluated the distribution of p-values for the Cochran-Armitage trend test. In this paper, an enhanced version of Cochran-Armitage trend test based on bootstrap approach is introduced. The achieved results confirm that the distribution of p-values of the proposed approach fits better to the uniform distribution, and it is thus concluded that the proposed method, which needs less assumptions in comparison with the conventional method, can be successfully used to test the genetic association.

1. Introduction

A central goal of genome wide association studies (GWAS) is to identify genetic risk factors for complex disorders. In order to find the disease genetic risk factors in a population, GWAS measures DNA sequence variations across human genome (Bush and Moore, 2012). Practitioners in medical sciences and bioinformatics use GWAS to investigate the relations in different disorders; GWAS of different cancers (Easton and Eeles, 2008), GWAS of pancreatic cancer (Amundadottir et al., 2009). The idea of genetic variations with alleles that are common in the population may explain much of the heritability of common diseases, see (Reich and Lander, 2001) and (Schork et al., 2009). Review of GWAS can be found in several texts and papers, see (Moore et al., 2010) among others.

In the simplest form of association mapping, a set of markers are genotyped in both sample of cases and sample of unrelated controls and then using different association tests, allele frequency differences or genotype frequency differences at each marker will be studied (Pritchard and Donnelly, 2001). The main idea behind GWAS studies relies on the fact that if a mutation has positive correlation with susceptibility of a disease, then that mutation is expected to be more frequent among affected individuals than those unaffected individuals (Pritchard and Donnelly, 2001). Hence, considering the existence of linkage disequilibrium (LD) between the marker locus and the susceptibility mutation, the marker close to the disease mutation may also present a frequency difference between case and control group of study (Pritchard and Donnelly, 2001).

Case-control traits can be analysed using either logistic regression or contingency table techniques (Bush and Moore, 2012). Contingency table methods examine the deviation from independence that is expected under the null hypothesis of observing no association between the disease under study and the measured allelic/genotyping frequency differences (Bush and Moore, 2012). Pearson chi-squared test and the related Fisher's exact test are the most widely used tests for independence of the rows and columns of the contingency table (Bush and Moore, 2012).

It should be noted that the association tests are performed separately for each individual marker and depending on the aim of study, the data for each marker with minor allele *a* and major allele *A* can be represented either as genotype count (e.g., *a*/*a*, *A*/*a* and *A*/*A*) or allele count (e.g., *a* and *A*) (Clarke et al., 2011). It is widely believed that the allelic association test with 1 degrees of freedom (df) is more reliable than the genotypic test with 2 df. However, it is imperative to note that this superior performance can only be considered for the case of having the penetrance of the heterozygote genotype between the penetrance of the two homozygote genotypes (Clarke et al., 2011). When the distribution of genotypes in the population deviates from Hardy-Weinberg proportions (HWE), of which additive, dominant and recessive models are all examples (Clarke et al., 2011), the frequency of genotypes rather than alleles should be compared by the Cochran-Armitage test for trend

* Correspoding author.
E-mail address: hassani.stat@gmail.com (H. Hassani).

http://dx.doi.org/10.1016/j.mgene.2016.07.001
2214-5400/

(Sasieni, 1997). For more information on different models see (Clarke et al., 2011).

Thus, the advantage of the Cochran-Armitage trend test in comparison to Pearson's Chi-Square test is that it possesses the superior conservation and is not dependent on the HWE assumption (Sasieni, 1997). Therefore a number of authors have recommended to use the Cochran-Armitage trend test as the genotype-based test for association (Sasieni, 1997; Corcoran et al., 2000; Li, 2008; Risch and Merikangas, 1996; Risch, 2000). It should also be noted that the allelic and trend statistic are equivalent when the combined sample is in HWE (Sasieni, 1997).

However, a major drawback of model based methods is that the statistical properties depend on the choice of weights. Thus, the model miss-specifications minimize the power of the test (Sasieni, 1997; Corcoran et al., 2000; Li, 2008; Risch and Merikangas, 1996; Risch, 2000). Furthermore, Escott-Price et al. (2013) showed that, although in most scenarios the Cochran-Armitage trend test is more powerful than the chi-squared test of genotype counts, the advantage is not substantial. Even, when the disease locus is extremely biased from the additive model, the chi-squared test of genotype counts can be more powerful than the Cochran-Armitage trend test due to the choice of scores for each genotype in the trend test (Escott-Price et al., 2013).

Although, there are considerable studies about the advantages and disadvantages of Cochran-Armitage trend test, to the best of our knowledge, there is a small number of researches which evaluated the distribution of p-values for this association test. In this paper the distribution of the p-values derived by the Cochran-Armitage trend test has been studied and it has been shown that unlike the considered presumption those p-values obtained by this test are not uniformly distributed. To overcome this issue, we introduce a new method, based on the bootstrap technique, for computing the p-value of the Cochran-Armitage trend test.

The bootstrap method has become a standard tool in statistical analysis and is an indispensable tool for testing statistical hypotheses. Using resampling, bootstrap approximates the sampling distribution of a statistic under the null (or the alternative) hypothesis. Bootstrap provides a practical complement to asymptotic parametric inference, hence have attracted many attentions in the applied. The efficiency of the nonparametric bootstrap method has also been shown by Amiri and von Rosen (2011) in which for example in the case of the Pearson chi-squared statistic with a Yates' correction and Fisher's exact test, remarkable improvement has been achieved. The Pearson chi-squared statistic with a Yates' correction and Fisher's exact test, are quite conservative and fail to reject the null hypothesis and can not be recommended to test independence with small sample sizes.

The remainder of this paper is organized as follows. The concept of Cochran-Armitage trend test is explained in Section 2. Section 3 studies the alternative approach to draw the inference including the bootstrap version of Cochran-Armitage trend test. Section 4 investigates the proposed method using the Monte Carlo simulation, which show they are the accurate tests in terms of the significant level and statistical power. Section 4 also demonstrates the improvements in goodness-of-fitness achieved by the introduced bootstrap approach. The paper concludes with a concise summary in Section 5.

2. Cochran-Armitage trend test

The Cochran-Armitage's trend test is a widely used test for trend among binomial proportions which uses the genotype contingency table (Table 1) in a different manner than Pearson's test. Power is very often improved as long as the probability of having disease increases with the number of disease-associated alleles. In genetic association studies in which the underlying genetic model is unknown, the additive version of this test is most commonly used. In order to measure the effect of genotype i and to detect particular types of association, we introduce a weight w_i. The special choice $(w_0,w_1,w_2)=(0,1,2)$, represents the additive effect of allele A. (See Table 2.)

Table 1
Genotype counts distribution for the case-control studies.

	$w_0=0$	$w_1=1$	$w_2=2$	Total
Case	n_0	n_1	n_2	n
Control	m_0	m_1	m_2	m
Total	N_0	N_1	N_2	N

Let us consider a single-marker locus with two possible alleles which are commonly denoted by A and a. Thus, each individual has three possible genotypes AA, Aa, and aa. In the following we denote the two alleles by 0 and 1 instead of A and a and the genotypes by 0, 1, 2, the sum of the two allele indices involved. We assume a random sample of n cases and m unrelated controls. The case-control data can then be summarized according to genotypes as shown in Table 1.

Here, (n_0,n_1,n_2) are counts of the genotypes in cases and (m_0,m_1,m_2) are counts of the genotypes in controls, and (N_0,N_1,N_2) are counts of the genotypes in case-control samples. Let n and m be the total number of cases and controls, respectively, and the total sample size, $N=n+m$. As cases and controls are independently sampled the genotype counts for cases and controls follow independent multinomial distributions with parameters (p_0,p_1,p_2), and (p_0',p_1',p_2'), respectively, where p_i and p_i', $i=0,1,2$, are the genotype probabilities in cases and controls.

$$(n_0,n_1,n_2):\mathrm{Multi}(n,p_0,p_1,p_2),$$

$$(m_0,m_1,m_2):\mathrm{Multi}(m,p_0',p_1',p_2').$$

Under the null hypothesis of no association, $H_0: p_i=p_i'$ for $i=0,1,2$. The Cochran-Armitage's trend test statistic for the data in Table 1 is given by

$$T=\frac{N(N(n_1+2n_2)-n(N_1+2N_2))^2}{n(N-n)\left(N(N_1+4N_2)-(N_1+2N_2)^2\right)}. \quad (1)$$

The statistic in Eq. (1) follows the chi-square distribution with one degree of freedom (df), see (Armitage, 1955). Let us denote the Cochran-Armitage trend test as CA in the rest of work.

Agresti (2007) states CA in terms of the Pearson chi-squared statistic. Consider a contingency table $2\times J$ with ordered column, see Table 1. Let $n_j\sim bin(N_j,p_j)$, $j=0,\ldots,J-1$, it is of interest to test the following null hypothesis

$$\begin{aligned} &H_0: p_0=p_1=\ldots=p_{J-1},\\ &H_1: p_i\neq p_j,\ \exists i\neq j. \end{aligned} \quad (2)$$

It can be carried out by using a linear probability model

$$p_j=\alpha+\beta w_j. \quad (3)$$

Table 2
Frequency table.

score				
w_0	w_1	...	w_{J-1}	total
n_0	n_1	...	n_{J-1}	n
m_0	m_1	...	m_{J-1}	m
N_0	N_1	...	N_{J-1}	N

One can use the ordinary least square approach for testing β. Let $\overline{w} = \sum N_j w_j/N$, $\tilde{p}_j = n_j/N_j$ and $\hat{p} = n/N$. The prediction equation is

$$\hat{p}_j = \hat{p} + \hat{\beta}(w_i - \overline{w}),$$

where

$$\hat{\beta} = \frac{\sum N_j \left(\tilde{p}_j - \hat{p}\right)(w_j - \overline{w})}{\sum N_j (w_j - \overline{w})^2}.$$

Using the Pearson chi-squared statistics

$$X^2 = \sum_j \frac{N_j\left(\tilde{p}_j - \hat{p}\right)^2}{\hat{p}(1-\hat{p})} = Z^2 + X^2(L) \sim \chi^2_{J-1}, \quad (4)$$

Where

$$Z^2 = \frac{\widehat{\beta^2}}{\hat{p}(1-\hat{p})} \sum_j N_j (w_j - \overline{w})^2,$$

$$X^2(L) = \frac{1}{\hat{p}(1-\hat{p})} \sum_j N_j \left(p_{\sim j} - \hat{p}_j\right)^2,$$

under linear probability model $X^2(L) \sim \chi^2_{J-2}$ that using the application of Cochran's theorem, $Z^2 \sim \chi^2_1$. It can be used to test $H_0: \beta = 0$ for the linear trend, the test of independence using Z^2 is called the Cochran-Armitage (CA) trend test.

3. Bootstrap Cochran-Armitag trend test

The bootstrap method has brought a vast new body of statistics in the form of nonparametric approaches to model uncertainty, in which not only the individual parameters of the probability distribution, but also the entire distribution are sought (Amiri, 2013). This has led to a versatile tool for data analysis, in particular in the field of statistical hypothesis tests. Two monographs on the bootstrap method written by Efron and Tibshirani (1994) and Davison and Hinkley (1997) are very useful in this regard in that they focus more on applications than on the theoretical approach. The idea of the bootstrap method is to approximate the sampling distribution of the proposed statistic, and this technique is based on resampling, which provides a practical complement to asymptotic parametric methods. The flexibility and robustness of this technique, especially in situations where the violation of assumptions is being dealt with, can be counted as two advantages of the technique (Good, 2013).

Amiri and von Rosen (2011) use the bootstrap to carry out the test of the contingency table. In order to test the association using the bootstrap method, the resampling should be performed on E_{ij}, where it is held for the expected value in the (i,j)th cell. The principle of the bootstrap test is the performance of bootstrap resampling under the null hypothesis, which is explained in (Efron and Tibshirani, 1994) and (Davison and Hinkley, 1997). The null hypothesis of the lack of the association in the contingency table is $H_0: p_{ij} = p_{i.}p_{.j}$, it leads to $H_0: E_{ij} = O_{i.}O_{.j}/O_{..}$, and therefore resampling under the null hypothesis is resampling on E_{ij} rather than O_{ij}, where O_{ij} is held for the observed value. The dot in the subscription denotes summation. Let X^{2^*} be the resampled that is done under null hypothesis and has χ^2_{J-1}, since $X^{2^*} = Z^{2^*} + X^{2^*}(L)$, Z^{2^*} has χ^2_1 and can be used to test $H_0: p_0 = p_1 = \ldots = p_{J-1}$.

3.1. First approach: NBCA

The test can be done using the following steps,

1. Calculate T or Z^2.
2. Resample data under E_{ij}, and obtain the contingency table $\mathrm{N}^* = \{n^*_0, \ldots, n^*_{J-1}, m^*_0, \ldots, m^*_{J-1}\}$, where $\mathrm{N}^* \sim \mathrm{Multi}(N, \frac{E_{00}}{N}, \ldots, \frac{E_{2(J-1)}}{N})$.
3. Repeat the second step B times, and calculate T^*_b or $Z^{2^*}_b$, $b = 1, \ldots, B$.
4. Estimate p-value using

$$p\text{-value} = \frac{\#\{T^*_b > T\}}{B}.$$

Let us denote the above approach as NBCA.

3.2. Second approach: PBCA

Another approach is to consider a parametric bootstrap. To this end, consider each allele or column are produced from the independent pdf. Under the null hypothesis

$$f_p(n) = \prod_{j=0}^{J-1} \binom{N_j}{n_j} p^{n_j}(1-p)^{N_j - n_j}, \quad (5)$$

that is actually a product of four binomial pdf. In order to estimate p, the maximum likelihood estimation (MLE) of it can be used i.e., $\hat{p} = \frac{n}{N}$. The procedure of the test is the same as below, just in the step 3, the resampled contingency tables are generated using $f_{\hat{p}}(n) = \prod_{j=0}^{J-1} \binom{N_j}{n_j} \hat{p}^{n_j}(1-\hat{p})^{N_j - n_j}$. The test referred to as PBCA in the rest of work.

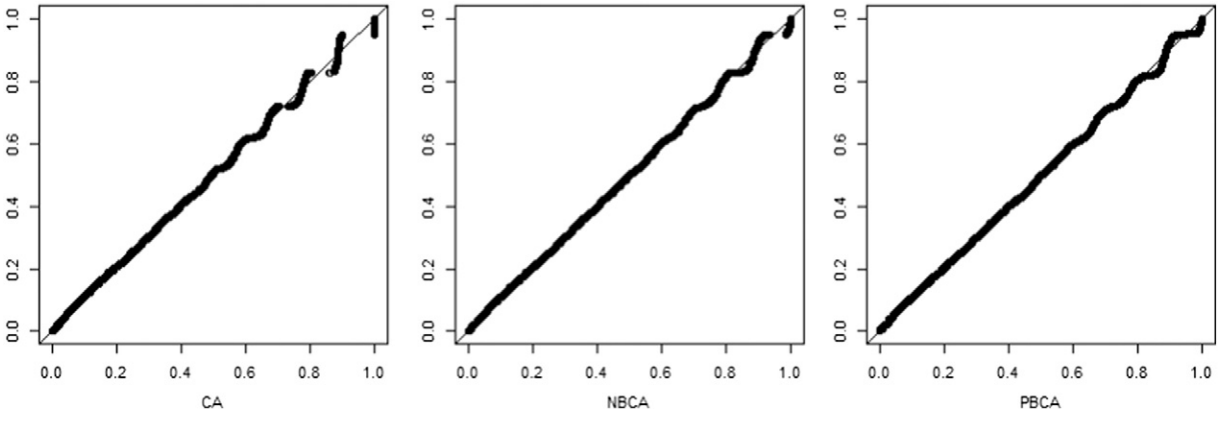

Fig. 1. The Q-Q plot of the p-value for the proposed tests, $n = m = 50$.

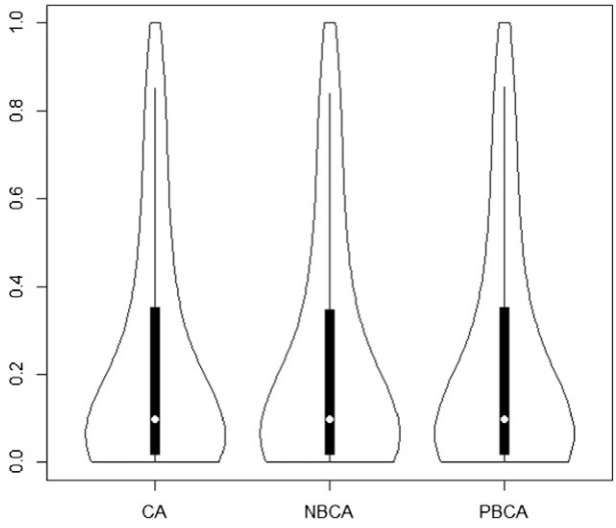

Fig. 2. Violin plot of the simulated p-value for the proposed tests when null hypothesis is not correct.

4. Numerical studies

This section demonstrates the validity of the proposed methods for the inference of Cochran-Armitage trend test. In order to study the finite sample properties of the proposed approaches, Monte Carlo experiments are used. The proposed methods are simultaneously based on the same simulated data in order to provide a meaningful comparison of various algorithms. In total 5000 simulations were performed. In order to make a comparative evaluation of the procedures, we seek the certain desirable features such as the actual significance level.

In order to produce the simulation, the data is generated using $(n_0,n_1,n_2):\text{Multi}(n,(0.2,0.4,0.2))$ and $(m_0,m_1,m_2):\text{Multi}(m,(0.2,0.4,0.2))$. The Q-Q plot of the p-value of the proposed tests are given in Fig. 1, which shows that the p-value using the bootstrap tests fit better to the uniform distribution, that admits the bootstrap can be nominated to draw the inference.

Racine and Mackinnon (2007) suggest

$$p-\text{value} = \frac{\#\{T_b^* > T\}}{B+1} + \frac{U}{B+1}, \tag{6}$$

where $U:Unif(0,1)$. Under null hypothesis, $P(p-\text{value}<\alpha)=\alpha$ for any finite B, specially if the number of bootstrap is not large.

The simulated data are generated using $(n_0,n_1,n_2):\text{Multi}(n,(0.2,0.4,0.2))$ and $(m_0,m_1,m_2):\text{Multi}(m,(0.4,0.2,0.2))$. The Violin plot of the simulated power is given in Fig. 2. The Violin plot is a combination of a box plot and a kernel density plot; it starts with a box plot, and then adds a rotated kernel density plot to each side of box plot that provides a better indication of the shape of distribution and summary of data.

Fig. 3 illustrates the Q-Q plot of generated random number of size 100,000 from χ_1^2. The results confirm that the statistic with distribution χ_1^2 suffer from the lack of goodness-of-fitness in the right tail. This fact is also quite evident for Cochran-Armiatage trend test.

In order to study the efficiency the proposed approaches, we generate 2000 tables with $(n_0,n_1,n_2):\text{Multi}(n,(0.2,0.4,0.2))$ and $(m_0,m_1,m_2):\text{Multi}(m,(0.2,0.4,0.2))$, where $n=m=20$. The Q-Q plot from χ_1^2 and the bootstrap approach is given in Fig. 4, where clearly confirms the superior of the proposed approaches. Note also that both proposed bootstrap approaches perform similarly here.

5. Conclusion

This article explores the genetic association study for the case-control design that draw the inference of the equality of the genotype frequencies. GWAS represent important challenges and opportunities in bioinformatics as they enable modeling of complex genotype-phenotype relationships using the mathematical and statistical approaches. Such models aids us to understand and interpret genetic

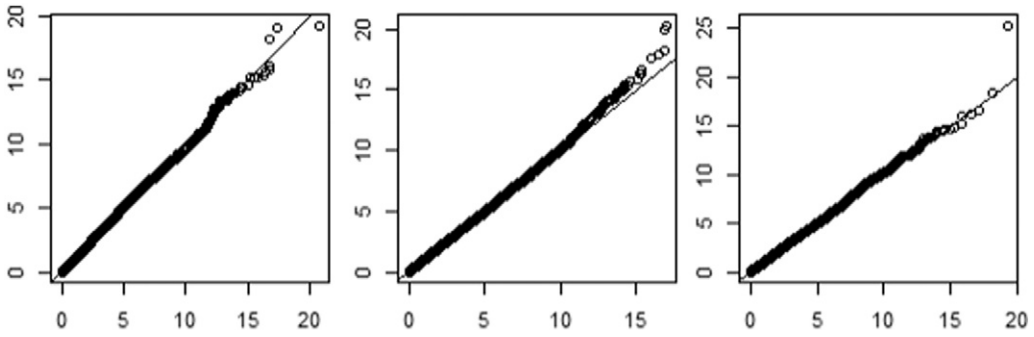

Fig. 3. Q-Q-plots of random number generated from χ_1^2.

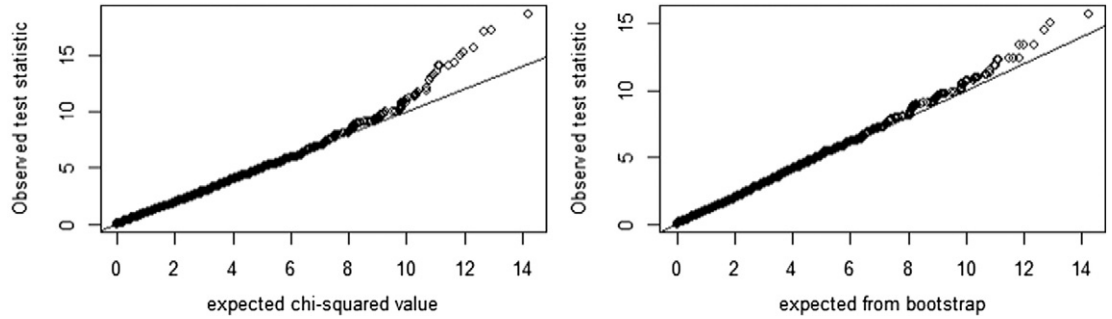

Fig. 4. The Q-Q plot of in terms of the expected values from χ_1^2 and the bootstrap.

association studies and promotes the development of powerful algorithms to examine genotype-phenotype relationships.

In this paper, we explored the Cochran-Armitage trend test and its bootstrap versions. It was shown that the proposed bootstrap can be used to test the genetic association. The results confirm that the p-value of the proposed approaches fits better to the uniform distribution, specially on the right side. Another advantage of the proposed tests require less assumption in comparison with the conventional method. The results also support that the proposed approaches can be successfully employed to test the genetic association. Extending the proposed idea in this paper to obtain a better test that is more robust under choosing weights for the Cochran-Armitage trend test is our future research plan.

References

Agresti, A., 2007. Categorical Data Analysis. second ed. Wiley.

Amiri, S., 2013. Bootstrap test of multinomial population and its application for resampling energy data. Int. J. Energy Stat. 1 (3), 215–224.

Amiri, S., von Rosen, D., 2011. On the efficiency of bootstrap method into the analysis contingency table. Comput. Methods Prog. Biomed 104 (2), 182–187.

Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R.Z., Fuchs, C.S., Petersen, G.M., Arslan, A.A., Bueno-de-Mesquita, H.B., Gross, M., Helzlsouer, K., Jacobs, E.J., LaCroix, A., 2009. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat. Genet 41 (9), 986–990.

Armitage, P., 1955. Tests for linear trends in proportions and frequencies. Biometrics 11 (3), 375–386.

Bush, W.S., Moore, J.H., 2012. Genome-wide association studies. PLoS Comput. Biol 8 (12), p.e1002822.

Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P., Zondervan, K.T., 2011. Basic statistical analysis in genetic case-control studies. Nat. Protoc 6 (2), 121–133.

Corcoran, C., Mehta, C., Senchaudhuri, P., 2000. Power comparisons for tests of trend in dose? Response studies. Stat. Med 19 (22), 3037–3050.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge university press.

Easton, D.F., Eeles, R.A., 2008. Genome-wide association studies in cancer. Hum. Mol. Genet 17 (R2), R109–R115.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.

Escott-Price, V., Ghodsi, M., Schmidt, K.M., 2013. How allele frequency and study design affect association test statistics with misrepresentation errors. Biostatistics p.kxt048.

Good, P., 2013. Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses. Springer Science & Business Media.

Li, W., 2008. Three lectures on case-control genetic association analysis. Brief. Bioinform 9 (1), 1–13.

Moore, J.H., Asselbergs, F.W., Williams, S.M., 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26 (4), 445–455.

Pritchard, J.K., Donnelly, P., 2001. Case-control studies of association in structured or admixed populations. Theor. Popul. Biol 60 (3), 227–237.

Racine, J.S., Mackinnon, J.G., 2007. Simulation-based tests that can use any number of simulations. Comput. Stand.?Simul. Comput.® 36 (2), 357–365.

Reich, D.E., Lander, E.S., 2001. On the allelic spectrum of human disease. Trends Genet 17 (9), 502–510.

Risch, N.J., 2000. Searching for genetic determinants in the new millennium. Nature 405 (6788), 847–856.

Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. Science 273 (5281), 1516–1517.

Sasieni, P.D., 1997. From genotypes to genes: doubling the sample size. Biometrics 1253–1261.

Schork, N.J., Murray, S.S., Frazer, K.A., Topol, E.J., 2009. Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev 19 (3), 212–219.