ABSTRACT

ZAYKIN, DMITRI V. Statistical Analysis of Genetic Associations (Advisor: Bruce S. Weir)

There is an increasing need for a statistical treatment of genetic data prompted by recent advances in molecular genetics and molecular technology. Study of associations between genes is one of the most important aspects in applications of population genetics theory and statistical methodology to genetic data. Developments of these methods are important for conservation biology, experimental population genetics, forensic science, and for mapping human disease genes. Over the next several years, genotypic data will be collected to attempt locating positions of multiple genes affecting disease phenotype. Adequate statistical methodology is required to analyze these data. Special attention should be paid to multiple testing issues resulting from searching through many genetic markers and high risk of false associations. In this research we develop theory and methods needed to treat some of these problems. We introduce exact conditional tests for analyzing associations within and between genes in samples of multilocus genotypes and efficient algorithms to perform them. These tests are formulated for the general case of multiple alleles at arbitrary numbers of loci and lead to multiple testing adjustments based on the closing testing principle, thus providing strong protection of the family-wise error rate. We discuss an application of the closing method to the testing for Hardy-Weinberg equilibrium and computationally efficient shortcuts arising from methods for combining p-values that allow to deal with large numbers of loci. We also discuss efficient Bayesian tests for heterozygote excess and deficiency, as a special case of testing for Hardy-Weinberg equilibrium, and the

frequentist properties of a p-value type of quantity resulting from them. We further develop new methods for validation of experiments and for combining and adjusting independent and correlated p-values and apply them to simulated as well as to actual gene expression data sets. These methods prove to be especially useful in situations with large numbers of statistical tests, such as in whole-genome screens for associations of genetic markers with disease phenotypes and in analyzing gene expression data obtained from DNA microarrays.

STATISTICAL ANALYSIS OF GENETIC ASSOCIATIONS

by

Dmitri V. Zaykin

**A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy**

DEPARTMENT OF STATISTICS

**Raleigh
1999**

APPROVED BY:

Chair of Advisory Committee	**Co-Chair of Advisory Committee**

BIOGRAPHY

Personal History

- Born in Khabarovsk, Russia.

Education

- M.S. in Biology, Far Eastern State University, Vladivostok, Russia.
- Ph.D. in Biomathematics, North Carolina State University, December 1999
  Dissertation topic: Statistical Analysis of Genetic Associations.

Professional Experience

- Junior Research Fellow, Institute of Marine Biology, Vladivostok, Russia
  August 1988 – January 1993
- Visiting Scholar, North Carolina State University, Raleigh, NC
  January 1993 – August 1994
- Graduate Research Assistant, North Carolina State University, Raleigh, NC
  August 1994 – August 1997
- Graduate Research Internship at Glaxo Wellcome, Research Information Systems, Research Triangle Park, NC
  August 1997 – August 1999

- Statistical Geneticist at Glaxo Wellcome, Bioinformatics, Research Triangle Park, NC

  August 1999 – Current

ACKNOWLEDGMENTS

I am most grateful to my advisor, Bruce Weir, for help and support during my research and studies. I am very fortunate to have had the opportunity to learn from him.

I indebted to the rest of my committee for critical assessment of my progress during the writing of my dissertation and during my studies. More specifically, thanks to Sujit Ghosh for valuable advice in the process of converting me into a Bayesian in disguise; to Tom Kepler for ideas, methods and data; to Jeff Thorne for sharing his expertise in phylogenetic analysis and stochastic processes; and, to Shaobang Zeng for teaching me QTL analysis and introducing me to model selection issues.

Alexander Pudovkin has been a colleague and a friend I could always rely upon.

Thanks to Lev Zhivotovsky for the friendship, for sharing his expert mathematical knowledge, and for inspiring discussions.

I benefited from stimulating conversations with Stan Young and from discussions with Peter Westfall during our regular meetings at Glaxo Wellcome, Inc.

I am grateful for the friendship and help of the members of the Programs in Statistical Genetics and Biomathematics. Discussions with Dahlia Nielsen helped to formulate and put things together. Jennifer Shoemaker was helpful and patient with my questions. Chris Basten convinced me that UNIX and LaTeX are better than the alternative.

I thank Sveta for being with me.

Contents

LIST OF TABLES **viii**

LIST OF FIGURES **x**

1 INTRODUCTION **1**

1.1 References 6

2 EXACT TESTS FOR ASSOCIATION BETWEEN ALLELES AT ARBITRARY NUMBERS OF LOCI **9**

2.1 Abstract 10

2.2 Introduction 11

2.3 General method 12

2.3.1 Algorithm 15

2.4 Other tests 19

2.4.1 Products of one-locus frequencies 19

2.4.2 Products of genotypic frequencies at one locus and two allele frequencies at other locus 20

2.5 Discarding some genotypes 21

2.6 Using goodness of fit measures 22
2.7 Numerical results 23
2.7.1 Two loci 24
2.7.2 Many loci 26
2.8 Discussion 28
2.9 Acknowledgements 30
2.10 References 31

3 CLOSED MULTIPLE TESTING ADJUSTMENTS FOR HARDY-WEINBERG EQUILIBRIUM 38

3.1 Abstract 39
3.2 Introduction 39
3.3 Method 41
3.4 Results and discussion 43
3.5 Acknowledgments 47
3.6 References 47

4 COMBINING INDEPENDENT *P*-VALUES 56

4.1 Introduction 57
4.2 Previous methods 58
4.3 Truncated product method 60
4.4 Comparison of methods 63
4.5 Results 64
4.6 Conclusions 66
4.7 References 68

4.8 Appendix 1 71

5 DETERMINING TRUE EFFECTS IN GENE EXPRESSION DATA 78

5.1 Abstract 79

5.2 Introduction 79

5.3 Methods based on distributions of first order statistics. 83

5.3.1 Testing the overall hypothesis 83

5.4 Determining true effects 89

5.4.1 Results from simulations 90

5.4.2 Application to microarray data 91

5.5 Conclusions 93

5.6 Appendix 1 94

5.7 Appendix 2 95

5.8 References 96

6 *P*-VALUE ADJUSTMENTS IN CONFIRMATORY STUDIES 105

6.1 Background 106

6.2 Derivation of the method 106

6.3 Results and discussion 110

6.4 References 112

7 BAYESIAN TESTS FOR HETEROZYGOTE EXCESS AND DEFICIENCY 117

7.1 Abstract 118

7.2 Preliminaries 119

7.3 Method 120
7.4 Results and discussion 123
7.5 Appendix A 125
7.6 Appendix B 125
7.7 Appendix C 127
7.8 Appendix D 128
7.9 References 129

8 SUMMARY **132**

List of Tables

3.1 An example of the closure test for HWE simulated under drift and admixture 50

3.2 Genotype distribution for the example simulated under drift and admixture 51

3.3 Average numbers of rejections for the HWE test per set of k loci at the nominal level of 10%. 52

3.4 Values of y for different p_i's and levels of α (Stouffer et al.'s test) . 53

3.5 Values of ξ critical point for different numbers of hypotheses (Fisher's combination test) 54

3.6 Values of y for different p_i's and levels of α (Fisher's test) 55

5.1 $E(T)$; $(L = 50,000,\ \beta = 0.80)$ 99

5.2 $E\{I(T > 0)\}$; $(L = 50,000,\ \beta = 0.80)$ 99

5.3 $E(\text{FDR} \mid T + F > 0)$; $(L = 50,000,\ \beta = 0.80)$ 99

5.4 $E(T)$ 100

5.5 $E\{I(T > 0)\}$ 100

5.6 $E(\text{FDR} \mid T + F > 0)$ 101

5.7 Confirmatory analysis 102

5.8 $E(T)$, 3567 tests 103

5.9 $E\{I(T > 0)\}$, 3567 tests 103

5.10 $E(\text{FDR} \mid T + F > 0)$, 3567 tests 104

6.1 Power and expected number of rejections under H_0 114

6.2 Power and expected number of rejections under H_0 when all p-values from the first stage are included 114

6.3 Power and expected number of correct rejections under partial H_A with 1/10 tests from H_A 115

6.4 Power and expected number of correct rejections under partial H_A with 10 tests from H_A 116

7.1 Proportions of rejections under H_0 with declared 5% α-level 131

7.2 Proportions of rejections under H_A on the level of 5% 131

List of Figures

2.1 1 37

Chapter 1

INTRODUCTION

The purpose of this research is to develop and study statistical methods for analysis of genetic associations. Most of the motivation arises from the genetics context, however the applicability of approaches described here is broader.

In chapter 2 we developed exact tests for association between alleles at genetic loci and studied their properties. Exact tests are based on Fisher's (1932) idea that, under the multinomial sampling model, the unknown population frequencies of marginal categories (nuisance parameters) can be eliminated by conditioning on their observed counts. This leads to a way of calculating cumulative probabilities of joint multinomial counts under the null hypothesis of independence. Fisher developed his test for 2×2 contingency tables, but it is likely that only the lack of computing power prevented him from extending the test to the general $R \times C$ case.

If observed counts n_{ij} and n_{ji} cannot be distinguished, as is the case with diploid genotypic classification, a "folded contingency table", in I.J. Good's (1965) terminology, could be formed. When there are two alleles at a genetic locus, Haldane's (1954) exact test results. G. Ishi (1962) derived exact conditional probabilities for the general case with many categories and called such tables "intraclass contingency tables". Maiste (1993) studied multiallelic versions of such exact tests for one and two loci. We extended this approach to many loci and suggested an efficient algorithm for calculating exact conditional probabilities of observing counts of multilocus genotypes. These probabilities are calculated under different hypotheses of independence of alleles within and between loci and genotypes across loci. The resulting test addresses an assumption used in forensic identification cases, called "the product rule". The product rule is the procedure of multiplying

frequencies of matching alleles across loci. This procedure forms a combined estimated proportion of multilocus profile and relies on the absence of pronounced population structure. We showed that suggested exact tests perform satisfactorily for testing a hypothesis that this assumption of population homogeneity holds.

In chapter 3 we extended this research and developed multiple testing adjustments for tests for Hardy-Weinberg equilibrium. We proposed a powerful global exact test for Hardy-Weinberg equilibrium. The null hypothesis of this global test is that there is no deviation from equilibrium at any of the genetic loci. The alternative hypothesis is that there are one or more loci in disequilibrium. If the null hypothesis involving all k loci is rejected, it is then possible to proceed to smaller subsets of loci and localize minimal subsets containing one or more loci that retain the evidence of association. The procedure of starting with a global test and continuing with subsets of hypotheses has been termed the "closure" method by Marcus et al. (1976) and has been applied by them in the analysis of variance framework. If a valid α-level test is available for each subset of hypotheses, then the individual significance level for a hypothesis H_i is given by the least significant test of all subsets that involve H_i. Such a procedure protects the family-wise error rate (FWER) in the strong sense, that is, the probability of falsely rejecting one or more true null hypotheses is $\leq \alpha$ for all subsets of k, even in the case when there are some non-true null hypotheses in the set of k.

The disadvantage of the closure method is that there are $\sum_{i=1}^{k} \binom{k}{i} = 2^k - 1$ subsets to consider, a number that rapidly grows with k. This limits the applicability of the method in general, but sometimes shortcuts are available. For example, Holm's (1979) method of adjusting p-values for multiple testing is to start with the

smallest p in the set of k ordered p-values and proceed up as long as $p_i(k-i+1) \leq \alpha$. Then all null hypotheses are rejected for which this inequality holds. Holm's procedure immediately follows from the closure principle if the test statistic at each subset is the smallest p among p-values that correspond to individual hypotheses composing this subset, multiplied by the length of the subset. In other words, the test statistic is min-p, Bonferroni-adjusted by the number of individual hypotheses in the subset. If a subset of length j is declared significant, all shorter subsets with that min-p are automatically declared significant and this results in Holm's procedure, although Holm's proof that the method strongly protects FWER is based on different considerations.

Another example where the closure principle results in a simple way of localizing significant subsets of hypotheses is when the test statistic is based on the "truncated product" method of combining probabilities by Zaykin et al. (1999). The method and the closing procedure were described in chapter 4. The test statistic of the method is the distribution function of the product of p-values that are smaller than some cut-off point, τ, derived under the condition that all null hypotheses are true. Because of the algebraic properties of this test statistic, setting $\tau \leq \alpha$ results in a simple closing procedure where it is not necessary to test all $2^k - 1$ subsets of hypotheses. If the global hypothesis is rejected, the individual adjustments for any $p_i \leq \tau$ are approximately given by $1 - (1 - p_i)^{k+j-1}$, where j is the number of p-values that are $\leq \tau$. The method is found to work well when only some of the null hypotheses are false, and generally compares well with other methods for combining p-values. It is especially useful in situations when the number of tests is very large and a follow-up study is possible for the validation

of some of the tests. The truncated product method also has potential outside of the genetic context. For example, in meta-analysis of published data there is a well known problem of the "publication bias", when only successful findings are reported. Models have been developed for estimating the total number of studies (e.g. Gleser and Olkin, 1996), and therefore an adequate inference can be made using the truncated product method.

In chapter 5 we further developed and studied properties of the combination tests with a special attention to the cases of large k. A new method based on the distribution of first order statistics was suggested. The method also provides individual adjustments that are smaller than ones based on step-wise procedures of Hochberg or Holm. This improvement is obtained at the cost that k must be specified in advance. We described how combination tests are extended to handle correlated data. A series of simulations was conducted to study performance of the combination tests and compare them with step-wise methods of Hochberg (1988) and Benjamini and Hochberg (1995). Finally, the tests were applied to the microarray data and the conclusion has been made that among 3567 genes studied, it is likely that there are 50 to 75 genes that respond to the treatment, and that the individual statistical power of tests for comparisons of levels of DNA expression is 40 to 50%.

A replication experiment of a smaller size was available that confirmed some of the effects that were declared significant in the first experiment. To obtain the overall significance, properly adjusted by the number of tests in both studies, a method of Zaykin and Young (1999) has been used. The theory behind this method, as well as computer simulations, was developed in chapter 6. The method

provides a gain in power by taking into account the ordering of p-values.

Most of this work was concentrating around classical hypothesis testing and methods for adjusting and combining p-values. Chapter 7 was an attempt to look at the behavior of Bayesian methods for characterization of heterozygote deficiency and excess as if they were classical procedures. Similar comparisons exist in the literature. For example, Carlin and Louis (1996) discussed frequentist performance of point and interval estimators obtained from the Bayesian beta-binomial model.

Based on the coalescent process simulations of samples from admixed populations, it has been found that certain non-informative priors provide posterior probabilities with good frequentist properties. It has been found that they compare well with the previously suggested classical methods. These tests provide additional benefits of Bayesian interpretations and entire plots of posterior distributions of coefficients measuring excess or the deficit of heterozygotes. These methods are computationally much faster than frequentist exact tests.

1.1 References

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society series B-methodological, 57: (1) 289–300.

Carlin B. and Louis T.A. 1996. Bayes and empirical Bayes methods for data analysis. Chapman & Hall.

Fisher, R.A. 1932. Statistical Methods for Research Workers. Oliver and Boyd, London.

Gleser L.J. and Olkin I. Models for estimating the number of unpublished studies. Statistics in Medicine 1996. 15: 2493–2507.

Good, I.J. 1965. The estimation of probabilities; an essay on modern Bayesian methods. M.I.T. Press.

Haldane J.B.S. An exact test for randomness of mating. J. of Genetics. 52: 631–635.

Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75: (4) 800–802.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65–70.

Ishi, G. 1962. Intraclass contingency tables. Annals of the Institute of Statistical Mathematics XII: (2) 161–207.

Maiste, P.J. 1993. Comparison of Statistical Tests for Independence at Genetic Loci with Many Alleles. Ph.D. Thesis, North Carolina State University, Raleigh, NC.

Marcus, R., Peritz, E., and Gabriel, K.R. 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 77: 63:655–660.

Stouffer, S.A., E.A. Suchman, L.C. DeVinney, S.A. Star and R.M. Williams, Jr. 1949. The American Soldier, Vol. 1. Adjustment During Army Life. Princeton Univ. Press, Princeton.

Zaykin, D, Zhivotovsky Lev A., and Weir, B.S. 1999. Combining independent p-values (in preparation).

Zaykin D. and Young, S.S. 1999. P-value adjustments in confirmatory studies (in preparation).

Chapter 2

EXACT TESTS FOR ASSOCIATION BETWEEN ALLELES AT ARBITRARY NUMBERS OF LOCI

Zaykin D, Zhivotovsky L and Weir BS (1995)

GENETICA 96:169–178

2.1 Abstract

Associations between allelic frequencies, within and between loci, can be tested for with an exact test. The probability of the set of multi-locus genotypes in a sample, conditional on the allelic counts, is calculated from multinomial theory under the hypothesis of no association. Alleles are then permuted and the conditional probability calculated for the permuted genotypic array. The proportion of arrays no more probable than the original sample provides the significance level for the test. An algorithm is provided for counting genotypes efficiently in the arrays, and powers of the test presented for various kinds of association. The powers for the case when associations are generated by admixture of several populations suggest that exact tests are capable of detecting levels of association that would affect forensic calculations to a significant extent.

Keywords: Exact tests, allelic association, Hardy-Weinberg, Linkage disequilibrium.

2.2 Introduction

In the absence of evolutionary forces such as drift, selection, migration or mutation, genotypic frequencies are expected to be given by the products of corresponding allelic frequencies. Even if these forces are known to be present, however, it may be that genotypic frequencies are very close to the allelic frequency products. There are situations when it is convenient to be able to invoke this "product rule" for multilocus genotypes, as in the use of genetic profiles for human identification. A specific multilocus genotype is unlikely to have been seen in samples collected for the purpose of estimating frequencies, even though all the constituent alleles are present, and then the product rule offers a means of providing an estimate. Of course, it is necessary first to test for consistency of genotype frequencies to products of allele frequencies and such tests are covered in this paper.

With many loci and many alleles per locus, there are very many possible associations among the frequencies of subsets of the alleles. Even for two alleles at two loci, for example, there are six pairs of genes, four triples and one set of four genes to be considered (Weir and Cockerham, 1989). If the only issue is whether allele frequencies can be used to construct genotype frequencies, all these associations are tested for simultaneously in a single test. This has the advantage of avoiding problems with multiple tests, and the power of the test sometimes increases with the number of loci and alleles per locus. If there is interest in some of the individual associations, then specific tests can be constructed, and will be discussed here.

The primary purpose of this paper is to examine the use of "exact" tests, meaning tests based on the probabilities of sets of alleles conditional on observed

counts of subsets of the alleles. In general these tests are expected to perform well and to avoid the problems faced by chi-square goodness-of-fit tests when expected numbers are small. Because of the large number of possible multi-allelic arrays, it is not possible to examine them all to compute significance levels for the tests. Samples of arrays are generated by permutation, following the suggestion of Guo and Thompson (1992).

2.3 General method

The general hypothesis is that there is no association among the frequencies of constituent genes of a genotype. For locus l, each of the two alleles received by an individual at that locus has probability p_{l_i} of being allelic type A_{l_i}. If $P_{1_i1_j,2_i2_j,\cdots,L_iL_j}$ is the population frequency of the genotype $A_{1_i}A_{1_j}A_{2_i}A_{2_j}\cdots A_{L_i}A_{L_j}$, then the hypothesis can be expressed as

$$P_{1_i1_j,2_i2_j,\cdots,L_iL_j} = 2^H \prod_l p_{l_i} p_{l_j} \tag{2.1}$$

where H is the number of loci that are heterozygous. When only one locus is being considered and $L = 1$, Equation 2.1 is just the Hardy-Weinberg law. In other cases, rejection of the hypothesis does not indicate whether it is allelic frequencies within or between loci that are associated.

The development of a testing strategy is based on the multinomial distribution, meaning that each member of a population is assumed to be equally likely to be sampled, and that there is the same probability for each sample member having a particular genotype. Sample sizes are therefore considered to be very much smaller

than the population sizes. If $\mathbf{A}$ indicates a multilocus genotype, and n_{A_g} is the number of individuals of type $\mathbf{A}_g$ in a sample of size n from a population in which those genotypes have frequency P_{A_g}, then

$$\Pr(n_{A_1}, n_{A_2}, \cdots, n_{A_G}) \;=\; \frac{n!}{\prod_{g=1}^{G} n_{A_g}!} \prod_{g=1}^{G} (P_{A_g})^{n_{A_g}}$$

The quantity G is the number of different genotypes possible.

Under the null hypothesis of Equation 2.1, allelic counts n_{l_i} at locus l are multinomially distributed with sample size $2n$ and probabilities p_{l_i}. Furthermore, under the hypothesis these distributions are independent over loci. The joint probability of the sets of allelic counts $\{n_{l_i}\}$ is therefore

$$\Pr(\{n_{l_1}\}, \{n_{2_i}\}, \cdots, \{n_{L_i}\}) \;=\; \prod_{l=1}^{L} \left(\frac{(2n)!}{\prod_i n_{l_i}!} \prod_i (p_{l_i})^{n_{l_i}} \right)$$

and the probability of the genotypic counts conditional on the allelic counts is

$$\Pr(\{n_{A_g}\}|\{n_{l_i}\}) \;=\; \frac{n!}{\prod_{g=1}^{G} n_{A_g}!} \prod_{g=1}^{G} (P_{A_g})^{n_{A_g}} \prod_{l=1}^{L} \frac{\prod_i n_{l_i}!}{(2n)! \prod_i (p_{l_i})^{n_{l_i}}!}$$

Under the null hypothesis of complete independence of allele frequencies

$$\Pr(\{n_{A_g}\}|\{n_{l_i}\}) \;=\; \frac{n! \prod_{g=1}^{G} 2^{n_{A_g} H_g}}{\prod_{g=1}^{G} n_{A_g}!} \prod_{l=1}^{L} \frac{\prod_i n_{l_i}!}{(2n)!} \tag{2.2}$$

where count H_g is the number of heterozygous loci in genotype $\mathbf{A}_g$, of which there are n_{A_g} copies. Note that the unknown allelic frequencies p_{l_i} have cancelled out of this expression.

Genotypic arrays $\{n_{A_g}\}$, generated by permutation, with conditional probabilities equal to or less than that of the observed sample array contribute to the probability with which the null hypothesis would be rejected if it was true. This is

the significance level, or p-value. In general, it is not feasible to calculate this quantity exactly because of the prohibitively large number of genotypic count arrays for a given array of allelic counts.

Gail and Mantel (1977) discussed methods for determining the numbers of two- and three-dimensional contingency tables with fixed marginals. Another approximate method for setting a lower bound on the number will suffice to show that the number is indeed prohibitive. If P_1 is the largest value of all the array probabilities, then the number of arrays must be at least $1 + (1 - P_1)/P_1 = 1/P_1$. This would be the actual number if all arrays had probability P_1. Similarly, if P_1 is the smallest of the T largest probabilities, the number of arrays must be at least $T + (1 - \sum_{i=1}^{T} P_i)/P_1$. For cases when all these P's are small, the lower bound is given essentially by $(1 - \sum_{i=2}^{T} P_i)/P_1$. Since the ordering of all possible P's is unknown, this bound is calculated by finding the set of T largest P's for a large number of permuted arrays. Having $T > 1$ protects against under-estimation. The procedure was applied to STR data from a sample of 182 people typed at loci with 6, 6, 9 and 14 alleles. With $T = 1$ and 65,000 permutations, the lower bound was estimated to be of the order of 10^{748} in each of three separate determinations. Obviously, it is not possible to examine all possible arrays.

Instead of identifying all arrays with lower conditional probabilities than the sample, a set of arrays is generated randomly by permuting those alleles hypothesized to be independent. For the hypothesis in Equation 2.1, this means that alleles are permuted among individuals within loci, and independent permutations performed for each locus. The proportion of permuted arrays as probable or less probable than the sample forms an estimate of the significance level. If the true

significance level is α, then with probability 0.95 the estimate will be within δ of that value after m permutations if $m \approx 4\alpha(1-\alpha)/\delta^2 \leq 1/\delta^2$. Hence, with 95% probability, 10,000 permutations give an estimate accurate to two decimal places. Further discussion of this approach was given by Guo and Thompson (1992).

Applying the permutation method to estimate significance levels can be performed very efficiently. As all the arrays have the same allelic counts, for purposes of comparison it is necessary to compute only

$$P_s = \prod_g \frac{2^{n_{A_g} H_g}}{n_{A_g}!} \tag{2.3}$$

although it is the logarithm of P_s that is computed in practice, using an algorithm of Press et al. (1988). Furthermore, it is not necessary to step through all possible multilocus genotypes, since only those with non-zero counts contribute to Equation 2.3. From now on, this will be indicated by $g \epsilon z$, meaning that only those g values for which $n_{A_g} > 0$ are considered. The computer storage requirements therefore depend on the sample size rather than the number of possible genotypes.

2.3.1 Algorithm

The greatest saving in computing time is made in the way of counting the number of times each genotype appears in one of the genotypic arrays generated by permuting alleles. A naïve way would be to assign each genotype a numerical identifier, sort the identifiers and then count how many times each one occurs. For locus **B** with alleles numbered 1 to n_B, a possible identifier s_B for genotype $B_i B_j,\ j \leq i$ is

$$s_B = \frac{i(i-1)}{2} + j$$

Values of this quantity range from 1 for B_1B_1 to $S_B = n_B(n_B+1)/2$ for $B_{n_B}B_{n_B}$. If there is a second locus **C** with n_C alleles, then the identifier s_C ranging from 1 to S_C for genotypes at that locus can be defined similarly, and the identifier for two-locus genotypes is defined by

$$s_{BC} \;=\; S_C(s_B-1)+s_C$$

which ranges from 1 for $B_1B_1C_1C_1$ to S_BS_C for $B_{n_B}B_{n_B}C_{n_C}C_{n_C}$. The extension to multiple loci is straightforward. If necessary, the genotype can be recovered from the identifier. In the two-locus case, for example,

$$s_C = \frac{i_C(i_C-1)}{2}+j_C \;=\; s_{BC}(\mathrm{mod}S_C)$$

$$s_B = \frac{i_B(i_B-1)}{2}+j_B \;=\; \frac{(s_{BC}-s_C)}{S_C}+1$$

and the one-locus genotypes can be recovered from the one-locus identifiers by

$$\begin{aligned} i &= 1+\text{integer part of}[(\sqrt{8s_B+1}-1)/2] \\ j &= s_B-i(i-1)/2 \end{aligned}$$

The problem with this approach is the need to sort as many identifiers as there are individuals in the sample. The binary search tree method now described is very much faster.

The data set **D** has an element for every individual in the sample. A tree is constructed with nodes for every distinct genotype in **D**. Each element of **D** is placed on the tree, either at an existing node or at a new node, by comparing its identifier to the identifiers for the previously placed elements. This placement

procedure begins at the root node of the tree, and at each node, the tree is followed in one of two directions depending on whether the identifier is greater than or less than the identifier for that node.

The algorithm is as follows:

1. Build the binary search tree

a. Set counter $i = 1$. Take the element of **D** and insert it at the root node as the node identifier (ID). Set the root internal counter C (the number of times that genotype occurs in the sample) to 1 and the number of nodes NN to 1.

b. Increment i. Take the ith element r_i from **D** and recursively traverse the tree. If r_i is equal to the ID from some existing node, taking into account that for each locus genotype AA' is equal to $A'A$, increment the internal counter C at that node by 1. Otherwise, if any outer node (the top) is reached without finding an equal ID, insert r_i into the tree as a new node, setting the node for that ID to r_i, setting its count C to 1, and incrementing NN by 1.

c. If $i < n$, repeat the previous step.

2. Calculate $\ln P_s$ in Equation 2.3.

a. Set $\ln P_s = 0$.

b. For each node j in the tree, calculate $C_j H_j \ln 2 - \ln(C_j!)$ and add this to $\ln P_s$.

c. If **D** is the original sample, set $P_O = P_s$ and set $K = 0$. If **D** is a permuted array, increment K by 1 if $\ln P_s \leq \ln P_O$.

3. Permutation stage.

a. For each locus, randomly permute the $2n$ alleles in **D** for that locus. Return to the binary search tree step. Do this step while the number of permutations is less than the required number NR.

4. The estimated significance level is K/NR.

The binary search tree method for storing and retrieving the numbers of each multilocus genotype in a sample performs best when the genotypes are not sorted. This will certainly be the case for the permuted arrays. In the worst case, when the genotypes have been sorted, the method degenerates to a sequential search. Since the genotypes in the tree are stored in sorted order, it is guaranteed that no parts of the tree other than the current sub-tree can contain the particular genotype being sought. This means that only half the remainder of the tree needs be considered after each comparison. The maximum number of nodes in the tree cannot exceed the sample size, so that the average branch length remains the same as the numbers of loci and alleles increase and the time to complete the algorithm is therefore affected very little by these two numbers. The most time-consuming part of the algorithm is the permutation stage. Its speed depends on the sample size and number of loci, but not the total number of alleles per locus.

The binary search tree method has been programmed in `C++`, taking advantage of features of that language. It may be helpful, however, to illustrate the nature

of the algorithm using the more primitive genotype-identifiers described above. Suppose a sample of 11 multilocus genotypes has been reduced to integer identifiers by the method described above, and these identifiers are 24, 8, 37, 95, 24, 6, 28, 15, 23, 94, 27. The root node of the tree is 24, and at that stage $C = 1, NN = 1$. The tree can be drawn under the rule that the direction of travel is up to the left for identifiers smaller than that at the current node and up to the right for larger identifiers. The resulting tree is shown in Figure 1. A 12th genotype with identifier 25 would be located after only three steps: 25 is greater than 24, less than 37, and less than 28.

2.4 Other tests

More specific tests can be performed in order to characterize associations among subsets of the alleles within genotypes, or to increase the power of detecting specific associations. Details will now be given for tests on two-locus data.

2.4.1 Products of one-locus frequencies

There may be interest in whether two-locus genotypic frequencies can be represented as products of one-locus frequencies without assuming Hardy-Weinberg equilibrium. The null hypothesis can be written as

$$P_{A_iA_jB_kB_l} = P_{A_iA_j}P_{B_kB_l} \tag{2.4}$$

In the human identification case, such a situation might be appropriate when there was evidence of Hardy-Weinberg disequilibrium at each locus, and two-locus

genotypic frequencies were to be constructed as products of observed one-locus genotypic frequencies.

The sample counts can be written in a two-way table with row totals being **A**-locus genotype counts n_{ij} and column totals being **B**-locus counts n_{kl}. The table cell entries are the two-locus counts n_{ijkl}. The probability of the two-locus array conditional on the two one-locus arrays, under the null hypothesis in Equation 2.4, is

$$\Pr(\{n_{ijkl}\}|\{n_{A_{ij}}\},\{n_{B_{kl}}\}) \quad = \quad \frac{\prod_{i,j} n_{A_{ij}}! \prod_{k,l} n_{B_{kl}}!}{n! \prod_{i,j} \prod_{k,l} n_{ijkl}!}$$

In this notation, the subscripts i, j range over all **A**-genotypes in the sample and so include each of the rows in the table, and the subscripts k, l range over all **B**-genotypes and so include all the columns in the table. In estimating the significance level for the exact test of Equation 2.4, the only quantity to be calculated is $P_s = 1/\prod_{g \epsilon z} n_g!$, where g indexes each of the two-locus genotypes with a non-zero count. Permutation proceeds by keeping the one-locus genotypes intact and permuting these genotypes among individuals at one of the two loci.

2.4.2 Products of genotypic frequencies at one locus and two allele frequencies at other locus

To see if the two-locus genotypic frequencies $P_{A_iA_jB_kB_l}$ can be represented as the product of the genotype frequency $P_{A_iA_j}$ at one locus and the product of two allele frequencies $p_{B_k}p_{B_l}$ at the other locus, the hypothesis is

$$P_{A_iA_jB_kB_l} \quad = \quad 2^{H_{kl}} P_{A_iA_j} p_{B_k} p_{B_l} \tag{2.5}$$

where H_{kl} is 1 if $k \neq l$ and is 0 if $k = l$. A human identification setting for this test may be when one of two loci show departures from Hardy-Weinberg frequencies.

The conditional probability of a sample under the hypothesis in Equation 2.5 is

$$\Pr(\{n_{ijkl}\}|\{n_{A_{ij}}\},\{n_{B_k}\}) \;=\; \frac{2^{H_B}\prod_{i,j} n_{A_{ij}}!\prod_k n_{B_k}!}{(2n)!\prod_{i,j}\prod_{k,l} n_{ijkl}!}$$

where H_B is the number of individuals in the sample that are heterozygous at locus **B**. Estimating the significance level in this case requires comparisons among values of $2^{H_B}/\prod_{g\epsilon z} n_g!$, where n_g is still the number of occurrences of the gth two-locus genotype. Permutation proceeds by holding the **A**-locus genotypes intact and shuffling all $2n$ **B**-locus alleles among individuals.

2.5 Discarding some genotypes

There are systems of genetic markers, such as VNTRs, where there is difficulty in assigning genotypes unambiguously (Weir, 1992). A single band on an electrophoretic gel may represent a homozygote, a heterozygote for two alleles with similar copy numbers, or a heterozygote with an allele giving a band outside the scoring region of the gel. It is only individuals scored as heterozygotes for which there is no ambiguity and for which tests of association are required. Only the heterozygotes in the original sample are used, and only permuted arrays where all resulting individuals are heterozygous are used. Since every individual in every array is heterozygous, $H_g = L$ for every genotype A_g, and the conditional probability

in Equation 2.3 for L loci becomes

$$P_s = \prod_{g \epsilon z} \frac{2^{L n_{A_g}}}{n_{A_g}!}$$

2.6 Using goodness of fit measures

Hypotheses have been tested here by determining the proportion of a large number of permuted arrays that have a conditional probability no larger than that of the sample. Instead of using conditional probabilities, arrays could be compared to the hypothesized or expected array on the basis of a goodness of fit statistic such as chi-square. The significance level would be the proportion of arrays with a larger statistic than that of the sample.

For multilocus genotype $\mathbf{A}_g$, the observed count is n_g and the expected count under the hypothesis of interest is written as e_j. The chi-square statistic is the usual

$$X^2 = \sum_g \frac{(n_g - e_g)^2}{e_g}$$

It is necessary to sum only over the observed genotypes, and use can be made of the fact that, since $\sum_g n_g = \sum_g e_g = n$, the statistic is $[\sum_{g \epsilon z}(n_g^2/e_g)] - n$. It has been suggested (Anscombe, 1981) that Karl Pearson may have invented this statistic as an approximation for the exact test only because of the lack of computing resources at that time.

The difference between goodness of fit tests and tests based on conditional probabilities was explored in some detail by Maiste (1993). The difference in the present case is greatest when there are so many different genotypes possible

that each one is likely to be unique in a sample. Certainly this is the situation for forensic databases once five or more loci are scored. If each n_g equals 1, the relevant part of the chi-square test statistic is the sum of reciprocals of expected counts, each of which will be less than 1. The test statistic tends to increase with the number of possible genotypes. The conditional probability, by contrast, depends on the reciprocal of the product of factorial counts, each of which is 1, and this product does not change with the number of possible genotypes. As an extreme example, consider the 20×20 array in Table 1 which represents a sample of 38 two-locus genotype counts. The row and column totals are the one-locus counts. The sample size is much less than the 400 possible two-locus genotypes, and each two-locus genotype seen is unique in the sample. Other arrays, generated by permuting the genotypes at one locus holding both sets of one-locus counts constant, must have the same (when all two-locus counts remain at 0 or 1) or smaller (when some counts are greater than 1) conditional probability so that the significance level for the conditional probability test of no association is 1. However, the sampled individuals fall into cells with low expected frequencies and this is reflected by the goodness of fit test. Based on 17,000 shuffled arrays with the same marginals, the significance level for the X^2 test statistic was found to be 0.01. The same type of difference can happen in sparse arrays when each sampled individual is not unique.

2.7 Numerical results

The performance of the various test statistics and strategies has been investigated by applying them to simulated data sets.

2.7.1 Two loci

Cockerham and Weir (1973) expressed two-locus genotypic frequencies in terms of allelic frequencies and a set of disequilibrium coefficients. For example

$$\begin{aligned} P_{AABB} &= p_A^2 p_B^2 + 2p_A D_{ABB} + 2p_B D_{AAB} + 2p_A p_B \Delta_{AB} \\ &\quad + p_A^2 D_B + p_B^2 D_A + D_{AB}^2 + D_{A/B}^2 + D_A D_B \end{aligned}$$

For either two or four equally frequent alleles, genotypic frequencies were constructed with each disequilibrium coefficient in turn set to one-quarter of its maximum value, the other disequilibria being zero. Samples were taken from populations with these genotypic frequencies. For locus **A**, for example, with two equally frequent alleles, the coefficient D_A is bounded by ± 0.25 so that a population would be constructed with $D_A = 0.0625$. One thousand replicate samples of size 100 were drawn from each population. For each sample and each test, significance levels were computed from 10,000 permuted arrays. Although this number of permutations corresponds to the procedure used to analyze real data sets, it is not really necessary for simulation studies such as this. Oden (1991) and D.D. Boos (personal communication) have shown that the number of permutations can be set to a number very much smaller than the number of simulated data sets.

The entries in Table 2 are the proportions of the 1,000 simulated populations in which the significance level was estimated to be less than or equal to 0.05, and so are the powers of the tests for a 5% significance level. These powers are labelled P_1, P_2, P_3, P_4 for the four hypotheses considered:

- P_1: two-locus genotype frequency is product of four allele frequencies

- P_2: two-locus genotype frequency is product of two **A** allele frequencies and **B** genotype frequency
- P_3: two-locus genotype frequency is product of **A** genotype frequency and two **B** allele frequencies
- P_4: two-locus genotype frequency is product of two one-locus genotype frequencies

Table 2 shows that the test for overall association, as indicated by P_1, is working well in all situations, and is most sensitive to departures from Hardy-Weinberg disequilibrium. The power for detecting Hardy-Weinberg disequilibrium is greater for four than for two alleles. As the number of alleles increases beyond four, however, it has been found that power decreases when only one of the two loci has departures from Hardy-Weinberg. For the tests where genotype frequencies at one locus are held constant, the values of P_2 and P_3 show that Hardy-Weinberg disequilibrium at the other locus can be detected. The test characterized by P_4 is the most powerful of the conditional probability tests but, of course, does not detect departures from Hardy-Weinberg disequilibrium. Apart from the Hardy-Weinberg tests, the powers of tests for four-allele data are less than those for two-allele data. Larger sample sizes are needed to detect associations, although the loss of power with number of alleles varies among the tests.

Powers for the two-allele case can be verified by the method described by Fu and Arnold (1992). They discussed tests for 2×2 tables, and showed how powers could be calculated by restricting attention to those tables that make a significant contribution to power. For hypothesis P_1 when only D_{AB} is non-zero, their method

can be used for the table of four gametes $A_1B_1, A_1B_2, A_2B_1, A_2B_2$ with frequencies $p_{A_1}p_{B_1} + D_{AB}, p_{A_1}p_{B_2} - D_{AB}, p_{A_2}p_{B_1} - D_{AB}, p_{A_2}p_{B_2} + D_{AB}$. Values very similar to those in Table 2 are obtained. The method should be able to be extended to other situations covered in the table.

It should be noted that the power values shown in Table 2 are, to some extent, dependent on the method of simulating data sets. In real populations, where evolutionary forces may create complex patterns of association, the powers may well be higher. For example, P_4 could be high because of the non-zero values of several of the disequilibrium coefficients, whereas only one coefficient at a time was allowed to be non-zero in the simulations.

2.7.2 Many loci

For more than two loci, power was studied only for the test of the hypothesis in Equation 2.1 of no association between any of the alleles. Populations under the alternative hypothesis were simulated in a different way from previously, and in such a way to address the concerns of several authors (e.g. Lewontin and Hartl, 1991) that tests of association have low power as tests for population substructure.

Populations were constructed as amalgamations of several subpopulations that were subjected to drift, and were simulated by the coalescent process (Hudson, 1990) using a program written by P.O. Lewis. The degree of drift was specified by the coancestry coefficient $\theta \equiv F_{ST}$ (Weir, 1990) that measures the probability of any two allelic genes within a randomly mating population being identical by descent (relative to identity between populations), and serves as a measure of

divergence of populations from an ancestral population. Apparent associations between alleles were created by constructing samples as composites of samples of size 20 from 10 such populations. These admixed samples serve to address the issue of subpopulation structure that has caused concern over forensic calculations (e.g. Nichols and Balding, 1991) and correspond to the situation where a population consists of a collection of somewhat distinct subpopulations, but a sample is taken from the population as a whole.

Only unlinked loci were simulated, and either 5 or 10 equally frequent alleles at 1 to 100 loci were used. The number of replicate simulated populations used to determine power values ranged from 500 to 1,000 and the number of permuted arrays used to determine the significance level for each replicate was 3,200. Power values in the case where $\theta = 0$ (no association) should be 0.05.

When $\theta \neq 0$, power always increased with the number of loci, and with the number of alleles per locus. When $\theta = 0.05$, for example, the power shown in Table 3 is 0.969 for four loci and 10 alleles and 0.765 for 5 alleles, but it drops to 0.552 for 2 alleles (data not shown). The explanation is as follows. As the number of loci and alleles per locus increases, there is an increasing chance that each multilocus genotype in a sample becomes unique and the value of P_s in Equation 2.3 reduces to $\prod_{g \epsilon z} 2^{H_g}$. Because of the Wahlund effect of reducing heterozygosity by amalgamating 10 populations to provide the original simulated samples, the P_s value is expected to be smaller for this sample than for a permuted array.

The same does not apply for tests on heterozygotes only since the number of heterozygotes is the same in the sample and the permuted arrays. Indeed, the entries in Table 4 show a decrease in power, and poor performance of the test

when $\theta = 0$ for more than two loci. As Equation 2.3 tends to a value of 2^L, zero power when $\theta = 0$ suggests a very low chance of obtaining a sample with more than one copy of any four-locus genotype. When $\theta \neq 0$, power values are greater than zero, since then the samples contain genotypes homozygous at some loci and these individuals are discarded. This increases the chance of having some duplicate genotypes in the remaining data.

2.8 Discussion

Testing for associations between alleles, within and between loci, can be performed satisfactorily with exact tests conditional on allelic counts. Significance levels can be found by permutation procedures. These tests are an alternative to those based on normal statistics constructed as estimated disequilibrium coefficients divided by their estimated standard deviations (Weir and Cockerham, 1989). In general, Maiste (1993) found that conditional tests performed better than unconditional tests, of which the variance-based tests are an example.

In the current forensic setting, any associations between neutral genetic markers are most likely to be due to population substructure. Comparing multi-locus genotypic frequencies with the products of corresponding allelic frequencies is likely to detect these associations with a power that increases with the number of alleles per locus and/or the numbers of loci. Specifically, samples of size as small as 100 individuals have high powers for detecting the associations accompanying θ values of 0.05 when several loci are used. This θ value is the one suggested by Nichols and Balding (1991) as being a upper bound on actual values in human populations.

The fact that associations are generally not found when exact tests are applied (e.g. Evett et al. 1995) suggests that θ is less than this upper bound.

Conversely, Table 3 shows that the power is low when θ is 0.01. What is the effect of not detecting this level of population substructure? The strength of the evidence of a matching genotype between a evidentiary sample and a person suspected of having contributed that sample can be measured as a likelihood ratio. This is the ratio of the probability of the matching genotypes conditional on the suspected person being the contributor to the probability when another person is the contributor. The ratio, termed the forensic index by Weir (1994) by analogy to the usual paternity index, changes very little for small θ values. For a heterozygous matching genotype, when both alleles have frequencies of 0.05, the one-locus index changes from 200 when $\theta = 0$ to 145 when $\theta = 0.01$ (Weir, 1994). The change is greater for smaller allelic frequencies or larger θ values, but these are less likely for the loci currently being used in human identification.

If the product rule is to be employed to estimate multi-locus genotypic frequencies, it is necessary to check for associations between the constituent allelic frequencies. If exact tests do not detect associations, it appears appropriate to base forensic calculations on the product rule. Associations due to population substructure can be accommodated by modifications of the type proposed by Evett et al. (1995) or Balding and Nichols (1993).

2.9 Acknowledgements

This work was supported in part by NIH grant GM43544. Helpful comments were provided by Dr M.A. Asmussen. The possibility of calculating exact powers was suggested by Dr J. Arnold, who also provided the Anscombe reference.

2.10 References

Anscombe, F.J. 1981. Computing in Statistical Science through APL. Springer-Verlag, New York.

Balding, D.J. and R.A. Nichols. 1993. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci. Int. 64:125–140.

Cockerham, C.C. and B.S. Weir. 1973. Descent measures for two loci with some applications. Theor. Pop. Biol. 4:300-330.

Evett, I.W., P.D. Gill, J.K. Scranage and B.S. Weir. 1995. Establishing the robustness of STR statistics for forensic applications. (submitted)

Fu, Y.X. and J. Arnold. 1992. A table of exact sample sizes for use with Fisher's exact test for 2×2 tables. Biometrics 48:1103–1112.

Guo, S-W. and E.A. Thompson. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics (Eds.) Oxford Surveys in Evolutionary Biology, pp 1–44.

Lewontin, R.C. and D.L. Hartl. 1991. Population genetics in forensic DNA typing. Science 254:1745–1750.

Maiste, P.J. 1993. Comparison of Statistical Tests for Independence at Genetic

Loci with Many Alleles. Ph.D. Thesis, North Carolina State University, Raleigh, NC.

Nichols, R.A. and D.J. Balding. 1991. Effects of population structure on DNA fingerprint analysis in forensic science. Heredity 66:297–302.

Oden, N.L. 1991. Allocation of effort in Monte Carlo simulation for power of permutation tests. J. Am. Stat. Assoc. 86:1074–1076.

Press, W.H., S.A. Teukolsky, W.T. Vetteling and B.P. Flannery. 1988. Numerical Recipes in C. The Art of Scientific Computing. 2nd Edition. Cambridge Univ. Press, New York.

Weir, B.S. 1990. Genetic Data Analysis. Sinauer Associates, Sunderland, MA.

Weir, B.S. 1992. Independence of VNTR alleles defined as fixed bins. Genetics 130:873–887.

Weir, B.S. 1994. The effects of inbreeding on forensic calculations. Ann. Rev. Genet. 28:597–621.

Weir, B.S. and C.C. Cockerham. 1989. Complete characterization of disequilibrium at two loci. In M.W. Feldman (Ed.) Mathematical Evolutionary Theory, Princeton Univ. Press, Princeton, pp 86–110.

Table 1 Sample two-locus array. Marginal totals are one-locus counts.

0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4
0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	2
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	1	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	7
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	4
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
5	2	1	4	2	1	1	1	1	1	1	5	1	2	1	2	2	2	2	1	38

Table 2 Empirical powers of four tests for two loci with $m = 2$ or $m = 4$ equally frequent alleles when significance level is 0.05.

Nonzero disequilibria	m	P_1	P_2	P_3	P_4
None	2	0.049	0.052	0.051	0.056
	4	0.060	0.048	0.061	0.060
D_A	2	0.462	0.484	0.058	0.051
	4	0.951	0.976	0.041	0.056
$D_A + D_B$	2	0.778	0.486	0.498	0.065
	4	1.000	1.000	0.995	0.051
D_{AB}	2	0.708	0.744	0.742	0.783
	4	0.150	0.216	0.218	0.239
Δ_{AB}	2	1.000	1.000	1.000	1.000
	4	0.382	0.455	0.491	0.545
D_{AAB}	2	0.980	0.990	0.970	0.980
	4	0.275	0.351	0.294	0.324
$D_{AAB} + D_{ABB}$	2	1.000	1.000	1.000	1.000
	4	0.555	0.631	0.625	0.724
Δ_{AABB}	2	0.676	0.714	0.714	0.749
	4	0.204	0.226	0.212	0.259

Table 3 Empirical powers (with standard deviations) for exact multilocus test with $m = 5$ or $m = 10$ equally frequent alleles per locus, when significance level is 0.05.

		Number of 5-allele loci									
θ	m	1	2	3	4	10	15	25	50	75	100
0	5	0.050	0.048	0.052	0.044	0.052	0.056	0.047	0.051	0.047	0.049
		(.01)	(.01)	(.01)	(.01)	(.02)	(.01)	(.01)	(.01)	(.01)	(.01)
	10	0.050	0.049	0.051	0.052	0.051	0.048	0.050	0.050	0.050	0.049
		(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)
0.005	5	0.053	0.062	0.074	0.089	0.098	0.147	0.145	0.244	0.275	0.354
		(.01)	(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.03)	(.02)	(.03)
	10	0.065	0.067	0.097	0.126	0.129	0.192	0.240	0.327	0.493	0.600
		(.01)	(.01)	(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.02)	(.02)
0.007	5	0.055	0.082	0.098	0.096	0.127	0.197	0.226	0.370	0.459	0.561
		(.01)	(.01)	(.01)	(.01)	(.01)	(.02)	(.01)	(.03)	(.03)	(.02)
	10	0.067	0.097	0.106	0.125	0.225	0.274	0.360	0.584	0.740	0.823
		(.01)	(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.03)	(.04)	(.03)
0.010	5	0.065	0.078	0.104	0.119	0.217	0.270	0.342	0.519	0.682	0.809
		(.01)	(.01)	(.01)	(.01)	(.03)	(.02)	(.03)	(.02)	(.03)	(.02)
	10	0.068	0.104	0.141	0.193	0.298	0.411	0.583	0.845	0.942	0.986
		(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.02)	(.02)	(.01)	(.01)
0.050	5	0.148	0.323	0.661	0.770	0.989	1.000	1.000	1.000	1.000	1.000
		(.02)	(.02)	(.02)	(.02)	(.01)	(.00)	(.00)	(.00)	(.00)	(.00)
	10	0.237	0.774	0.894	0.969	0.999	1.000	1.000	1.000	1.000	1.000
		(.01)	(.02)	(.01)	(.01)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
0.100	5	0.434	0.856	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(.02)	(.02)	(.01)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
	10	0.757	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(.01)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)

Table 4 Empirical power of exact tests with homozygotes excluded when significance level is 0.05.

	Number of 10-allele loci			
θ	1	2	3	4
0	0.052	0.047	0.005	0.000
0.01	0.062	0.050	0.006	0.002
0.05	0.104	0.065	0.019	0.007
0.10	0.334	0.292	0.083	0.014

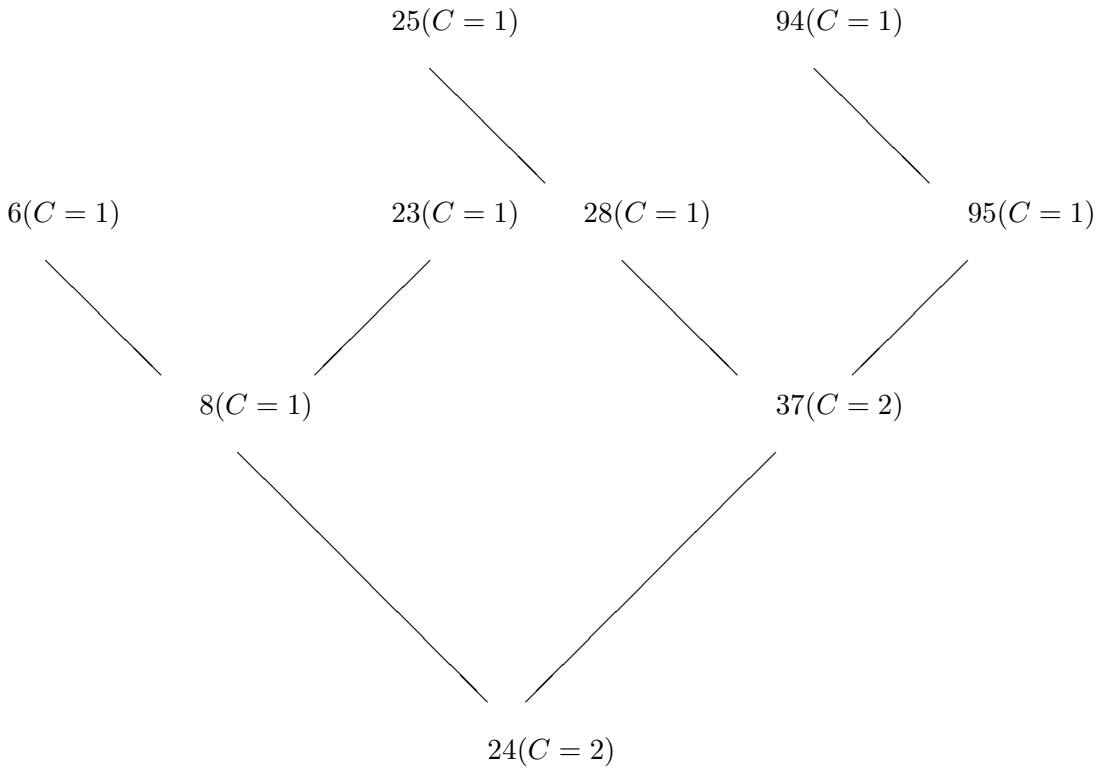

Figure 2.1: Binary Search Tree.

Chapter 3

CLOSED MULTIPLE TESTING ADJUSTMENTS FOR HARDY-WEINBERG EQUILIBRIUM

3.1 Abstract

We propose a closed multiple testing technique for evaluating the significance of tests for association between alleles. The procedure provides a global test for association over all loci as well as individual p-values adjusted for multiple testing. The procedure has high power when all loci are subject to equal evolutionary processes causing deviations from the equilibrium. Individual adjustments are feasible for moderate numbers of loci (k), with $2^k - 1$ amount of computations, but very good approximations that require at most $k(k-1)/2$ tests are available. The global exact test can be performed in all cases.

3.2 Introduction

For a family of hypotheses $\{H_1, ..., H_k\}$, the family-wise error rate (FWER) is defined as

$$\text{FWER} = \Pr(\text{reject one or more } H_i, i \in \mathbf{S} \mid H_i, i \in \mathbf{S} \text{ are true}) \tag{3.1}$$

where $\mathbf{S}$ is any subset of $\mathbf{K}$. The closure method was proposed Marcus, Peritz and Gabriel (1976) as a powerful technique to strongly control FWER. The multiple testing procedure is said to control FWER in the strong sense if FWER $\leq \alpha$ regardless of which subset of null hypotheses happens to be true (Westfall and Young, 1993).

The closure of the family of hypotheses $\{H_1, ..., H_k\}$ consists of the set of all intersection hypotheses, $H_{\mathbf{S}} = \cap_{i \in \mathbf{S}} H_i$ for $\mathbf{S} \subseteq \mathbf{K}$. This assumes that an α-level test $T_{\mathbf{S}}$ is available for each $H_{\mathbf{S}}$. The closed testing procedure rejects $H_{\mathbf{S}}$ at level

α if every $H_{\mathbf{J}}$ such that $\mathbf{S} \supseteq \mathbf{J}$ is rejected by $T_{\mathbf{S}}$.

It is easy to see that this procedure controls FWER. In order to reject one or more true null hypotheses (event A) we need to reject at the α-level both individual hypotheses and the intersection hypotheses containing them (event B). The probability of A is the FWER. But

$$\begin{aligned} \Pr(A) &= \Pr(A \text{ and } B) \\ &= \Pr(B) \Pr(A \mid B). \end{aligned} \tag{3.2}$$

Now, because $\Pr(B) = \alpha$ and $\Pr(A \mid B) \leq 1$, FWER $\leq \alpha$.

It is clear from the statement and the proof of the closure principle that tests at different intersections (or "nodes") need not be the same, as long as all tests to be applied have a size $\leq \alpha$ and are pre-specified.

Some multiple testing procedures result from the application of the closure principle. For example, the step-down procedure of Holm (1979) follows if the Bonferroni-adjusted minimum p-value is a test statistic for all $H_{\mathbf{S}}$. Holm's procedure adjusts p_i as $p_i^* = p_i(k-i+1)$, where p_i's are ordered from the smallest to the largest. The procedure stops when it reaches the first $p_i^* > \alpha$. If the Bonferroni-adjusted minimum p-value is a test statistic for the intersection hypothesis, then having rejected the hypothesis with all tests, all smallest subsets that include the same min p are rejected automatically, because of their smaller Bonferroni multipliers. The $(k-1)$-sized hypothesis that does not involve that min-p has the second smallest p-value of the k-sized hypothesis as its own min p. Therefore, as long as H_1 is rejected, H_2 will be rejected if $(k-1)p_2 \leq \alpha$ and Holm's procedure follows.

In this paper, we concentrate on adjustments with the application to testing for Hardy-Weinberg equilibrium. We suggest that in the presence of equal evolutionary forces acting upon all loci, the intersection hypotheses are more appropriately tested with a test statistic that combines information over all loci.

3.3 Method

Zaykin et al. (1995) suggested an algorithm for performing exact tests at arbitrary numbers of loci. Lewis and Zaykin (1997) implemented that algorithm in the GDA computer package. As a default, GDA performs exact tests for disequilibrium at each locus, then for all pairs, triples, etc., proceeding up to a set of all k loci. The total number of tests (N_T) is the number of possible subsets of loci:

$$N_T = \sum_{i=1}^{k} \binom{k}{i} = 2^k - 1 \tag{3.3}$$

To test the intersection hypothesis at each subset of loci, $\mathbf{S}$, we define a test statistic based on the conditional probability of observing a subset of one-locus genotype counts given allele counts at $\mathbf{S}$.

$$\Pr\left(\{n_{s\in\mathbf{S},ij}\} \mid \{n_{s\in\mathbf{S},i}\}\right) = \prod_{u\in\mathbf{S}} \frac{n!2^{H_u}\prod_i n_{ui}!}{(2n)!\prod_{i,j} n_{uij}!} \tag{3.4}$$

where 2^{H_u} is the number of heterozygotes at the locus u, n_{uij} is the count of genotype ij at the locus u, and n_{ui} is the number of alleles of type i at the locus u. The test statistic, T, is a proportion of genotypic arrays with the same allelic counts for which this probability is at least as small as P_{obs}, the probability calculated for the observed data.

$$T = E\left\{I\left[\Pr\left(\{n_{s\in\mathbf{S},ij}\} \mid \{n_{s\in\mathbf{S},i}\}\right) \le P_{\text{obs}}\right]\right\} \tag{3.5}$$

In practice, the expectation is obtained by averaging over a large number of random permutations. The method is expected to perform well when all loci are subject to the same evolutionary forces creating deviations from the random pairing of alleles. The closure method starts with testing the global hypothesis involving all k loci. If this hypothesis that there are no genetic loci in Hardy-Weinberg disequilibrium is rejected, the algorithm proceeds with testing subsets of $k-1$ loci, etc., down to the single-locus tests. For each intersection hypothesis $\mathbf{S}$, the p-value is given by the largest p-value among the subsets that include $\mathbf{S}$.

A similar procedure would result from the application of Fisher's combined probability test to p-values obtained from testing for HWE at each locus. Such a method will be implemented in the next release of PROC MULTTEST of SAS computer package (Peter Westfall, personal communication). Fisher's test is based on the fact that, under the null hypothesis, $-2\sum_i^k \ln p_i$ has a chi-square distribution.

However, one reason to prefer the global test based on (3.4) is that it accounts for the discrete nature of the test while Fisher's combined probability test assumes a continuous uniform distribution of p-values under the null hypothesis. Another reason is that, as was pointed out by Rice (1990), Fisher's combination test is inappropriate for testing whether several tests support the same null hypothesis. Fisher's combination test is biased in the favor of p-values that are small and Rice (1990) suggests that the test of Stouffer et al. (1949) should be preferred. The test of Stouffer et al. (1949) is based on normal-transformed p-values. If $\Phi(x)$ denotes the probability distribution function for the standard normal distribution,

$$\Phi(x) \;=\; \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{3.6}$$

then each p_i-value can be transformed to a standard normal score by

$$z_i \;=\; \Phi^{-1}(1-p_i) \tag{3.7}$$

The sum $z = \sum_i z_i/\sqrt{k}$ also has a standard normal distribution. The combined p-value is, therefore,

$$p \;=\; 1-\Phi\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\Phi^{-1}(1-p_i)\right) \tag{3.8}$$

Indeed, we found that for sufficient sample sizes and many alleles per locus, an application of the test by Stouffer et al. (1949) often results in p-values that are the same up to several decimal points as those based on (3.4), for all intersection hypotheses. An example of such data set is given in tables 3.1 and 3.2.

3.4 Results and discussion

Tables 3.1 and 3.2 give the genotypic distribution and an example of applying the method to the simulated data with 4 loci. The sample of 100 genotypes scored at four loci was generated by the coalescent process (Hudson, 1990) with consequent admixture. Ten populations drifted until the amount of divergence measured by the coancestry coefficient (Weir, 1996) reached the value of 0.15.

Table 3.1 lists subsets of loci corresponding to each intersection hypothesis, raw p-values obtained by the composite test (3.5), p-values adjusted by the closing procedure, and p-values obtained by applying the Hochberg's (1988) correction. Hochberg's method is a step-wise FWER-controlling multiple testing procedure and is less conservative than the usual Bonferroni correction. Hochberg's method

is to order p_i's, $\{p_1, ..., p_k\}$, start with $i = k$, and once $p_j \leq \alpha/(k - j + 1)$, then reject all H_i for $i \leq j$.

p-values obtained through the closure method appear to be smaller than those adjusted by Hochberg's method. However, it is more interesting to study how a procedure like this would perform on average, both in the presence and in the absence of population disequilibrium.

We repeated the simulations, comparing how often the individual null hypotheses of HWE are rejected with Hochberg and closure methods. Under the null hypothesis ($\theta = 0$), the closure method is found to be more conservative than Hochberg's, but it is more powerful in the presence of Hardy-Weinberg disequilibrium. These results are presented in table 3.3. In the cases studied, the closure method rejected about one more hypothesis, on average.

The usage of the procedure is broader than merely deciding which of the individual hypotheses to reject. It will often happen in practice, that in a set of k loci, some of the individual hypotheses will be rejected, but others will be significant down to, say, pairs of loci. This will indicate that there is evidence that one of the loci or both loci in the pair are in disequilibrium, but there is not enough statistical power to decide which. The global test including all loci is the most powerful one. If it is the only test that has been rejected, the conclusion is that k genetic loci provide overall evidence against the null hypothesis of the HWE, without ability to indicate specific individual loci.

The proposed global test can be carried out for any number of loci, but applicability seems to be limited to tens of loci because $2^k - 1$ intersection hypotheses need to be tested. Since we found that Stouffer et al.'s test closely approximates

the exact method, it is possible to avoid testing most of the intersections.

It follows from the symmetry of the distribution given by (3.8) that combining p-values that are greater than the critical point, ξ=0.5, always results in increasing the combined p-value towards one, and combining p-values that are smaller than ξ always results in decreasing of the combined p-value towards zero. Therefore, if we are concerned only with adjusting individual p-values that are below $\alpha \leq 0.5$, this test provides the following shortcut that avoids testing all $2^k - 1$ hypotheses. First, note that for $p_i \leq \alpha$, the solution of

$$\alpha = 1 - \Phi\left[\frac{1}{\sqrt{2}}\left(\Phi^{-1}(1-p_i) + \Phi^{-1}(1-y)\right)\right] \tag{3.9}$$

provides the value of y such that all pairs p_i, p_j, where $p_j > y$ give combined p-values above α, and following the closure principle the hypothesis H_i is not rejected. Such values could be obtained numerically (table 3.4). If there is no $p_j > y$, then for the ordered set of p_j's ($j = 1, ..., k$) the adjusting procedure for any p_i is as follows. Compute (3.8) for most stringent subsets:

$$\{p_i, p_k\}$$

$$\{p_i, p_k, p_{k-1}\}$$

$$\{p_i, p_k, p_{k-1}, p_{k-2}\}$$

$$...$$

$$\{p_i, p_k, p_{k-1}, p_{k-2}, ..., p_{i+1}\}$$

The adjusted p_i is given by the maximum of these values. Note that many subsets, such as, for example, $\{p_i, p_{k-1}\}$ do not need to be considered, because they will yield p-values smaller than the one for $\{p_i, p_k\}$. Therefore, the computational

reduction is from $2^k - 1$ to at most $k(k-1)/2$ evaluations of (3.8).

These considerations lead to a way of locating smallest subsets of loci that convey evidence of significance, because only a few tests need to be rejected in order to satisfy the closure principle.

The same shortcut method can be used if Fisher's combination test is applied instead, but somewhat more work is required, since the critical point (ξ) depends on the value of k. Let

$$W = -\sum_{i}^{k} \ln p_i \tag{3.10}$$

When all k null hypotheses are true,

$$W \sim \text{Gamma}(k, 1) \tag{3.11}$$

with the mean and the variance equal to k. When k is large,

$$p_w = \Pr\left(W > -\sum \ln \xi\right) \tag{3.12}$$

can be approximated by

$$p_z = \Pr\left(Z > -\sqrt{k}(\ln \xi + 1)\right) \tag{3.13}$$

where Z is the standard normal variable. When $\xi = 1/e$, this probability is 0.5, so that the critical point is given by $1/e$ when the number of tests is infinite. For a finite number of tests, the value ξ can be determined numerically. The values of ξ in the table 3.5 were calculated by using the golden section search (Press et al., 1988) on the function

$$f(x, k) = \text{abs}\left\{(\Psi(k, -k \ln x) - x)/x\right\}, \tag{3.14}$$

where $\Psi(\cdot)$ is the Gamma(k, 1, x) cumulative distribution function. Values in the table show that about two thousand tests are required for the critical point to be sufficiently close to the value of $1/e = 0.36788$, therefore numerical calculations like those described here are needed in practice.

In the case of Fisher's test, the equation analogous to (3.9) is

$$\begin{aligned} \alpha &= 1 - \int_{yp_i}^{1} \int_{v}^{1} (1/w)\ dw\ dv \\ &= p_i y\ (1 - \ln p_i - \ln y) \end{aligned} \tag{3.15}$$

with solution

$$y = -\frac{\alpha}{p_i\ \mathrm{prln}\left(-\frac{\alpha}{e}\right)} \tag{3.16}$$

where $\mathrm{prln}(x)$ denotes the product logarithm function, defined as the function that solves $x = we^w$ for w. Alternatively, the solution can be found numerically. Table 3.6 gives values of y for three levels of α (1%, 5%, and 10%) and a range of p-values.

3.5 Acknowledgments

A C++ source code for the methods described is available at

ftp://statgen.ncsu.edu/pub/zaykin/closed_hwe/

3.6 References

Hochberg Y. 1988 A sharper Bonferroni procedure for multiple tests of significance. Biometrika **75:** 800–802.

Holm, S. 1979 A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6:** 65–70.

Hudson, R.R. 1990 Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics (Eds.) Oxford Surveys in Evolutionary Biology, 1–44.

Lewis, P. O., and Zaykin, D. 1997 Genetic Data Analysis: Computer program for the analysis of allelic data. Free program distributed by the authors over the Internet from the GDA home page at http://alleyn.eeb.uconn.edu/gda/

Marcus, R., Peritz, E., and Gabriel, K.R. 1976 On closed testing procedures with special reference to ordered analysis of variance. Biometrika 77 **63:** 655–660.

Press, W.H., S.A. Teukolsky, W.T. Vetteling and B.P. Flannery. 1988. Numerical Recipes in C. The Art of Scientific Computing. 2nd Edition. Cambridge Univ. Press, New York.

Rice, W.R. 1990 A consensus combined p-value and the family-wide significance of component tests. Biometrics **46:** 303–308.

Stouffer, S.A., E.A. Suchman, L.C. DeVinney, S.A. Star and R.M. Williams, Jr. (1949). *The American Soldier, Vol. 1. Adjustment During Army Life.* Princeton Univ. Press, Princeton.

Weir, B.S. 1996 Genetic Data Analysis II. Sinauer Associates, Sunderland, MA.

Westfall, P.H. and Young, S.S. 1993 Resampling-Based Multiple Testing. Wiley, New York.

Zaykin, D., Zhivotovsky, L. and Weir, B.S. 1995 Exact tests for association between alleles at arbitrary numbers of loci. Genetica **96:** 169–178.

Table 3.1: An example of the closure test for HWE simulated under drift and admixture

loci subset	p-value	closure	Hochberg's
1	0.00621	0.01479	0.02484
2	0.33184	0.33184	0.33184
3	0.03841	0.06250	0.11523
4	0.06887	0.09715	0.13774
1/2	0.01479	0.00534	
1/3	0.00163	0.00345	
1/4	0.00268	0.00534	
2/3	0.06250	0.06250	
2/4	0.09715	0.09715	
3/4	0.01334	0.02128	
1/2/3	0.00345	0.00345	
1/2/4	0.00534	0.00534	
1/3/4	0.00066	0.00139	
2/3/4	0.02128	0.02128	
1/2/3/4	0.00139	0.00139	

Table 3.2: Genotype distribution for the example simulated under drift and admixture

locus 1	locus 2	locus 3	locus 4
BB:5	BB:7	BB:9	BB:8
BC:4	BC:4	BC:4	BC:8
BD:5	BD:7	BD:2	BD:6
BE:7	BE:10	BE:7	BE:4
BF:8	BF:13	BF:8	BF:4
CC:13	CC:2	CC:3	CC:6
CD:8	CD:6	CD:10	CD:11
CE:6	CE:6	CE:8	CE:3
CF:4	CF:6	CF:5	CF:5
DD:10	DD:4	DD:7	DD:14
DE:9	DE:5	DE:8	DE:8
DF:6	DF:13	DF:6	DF:10
EE:7	EE:8	EE:8	EE:4
EF:5	EF:5	EF:11	EF:5
FF:3	FF:4	FF:4	FF:4

Table 3.3: Average numbers of rejections for the HWE test per set of k loci at the nominal level of 10%.

θ	k	N_a	β (closure)	β (Hochberg's)
0.100	4	5	0.490	0.717
0.150	4	5	2.372	1.113
0.150	5	10	4.345	3.188
0.125	6	15	5.356	4.220
0.125	7	15	6.309	5.161

θ is the coancestry coefficient; k is the number of loci; N_a is the number of alleles; β is the average number of hypotheses rejected per simulation.

Table 3.4: Values of y for different p_i's and levels of α (Stouffer et al.'s test)

p_i	y, at α =1%	p_i	y, at α =5%	p_i	y, at α =10%
1e-05	0.83521	1e-05	0.97373	1e-05	0.99291
0.0004	0.52505	0.002	0.70952	0.004	0.79945
0.0008	0.44668	0.004	0.62775	0.008	0.72459
0.0012	0.39964	0.006	0.57377	0.012	0.67175
0.0016	0.36613	0.008	0.53297	0.016	0.63006
0.0020	0.34025	0.010	0.50007	0.020	0.59536
0.0024	0.31925	0.012	0.47248	0.024	0.56552
0.0028	0.30166	0.014	0.44872	0.028	0.53929
0.0032	0.28658	0.016	0.42788	0.032	0.51587
0.0036	0.27342	0.018	0.40934	0.036	0.49471
0.0040	0.26178	0.020	0.39265	0.040	0.47540
0.0044	0.25136	0.022	0.37749	0.044	0.45765
0.0048	0.24195	0.024	0.36362	0.048	0.44124
0.0052	0.23339	0.026	0.35084	0.052	0.42598
0.0056	0.22556	0.028	0.33902	0.056	0.41172
0.0060	0.21834	0.030	0.32802	0.060	0.39835
0.0064	0.21166	0.032	0.31775	0.064	0.38577
0.0068	0.20545	0.034	0.30813	0.068	0.37390
0.0072	0.19966	0.036	0.29908	0.072	0.36267
0.0076	0.19424	0.038	0.29054	0.076	0.35202
0.0080	0.18915	0.040	0.28248	0.080	0.34189
0.0084	0.18435	0.042	0.27484	0.084	0.33224
0.0088	0.17983	0.044	0.26758	0.088	0.32304
0.0092	0.17554	0.046	0.26069	0.092	0.31425
0.0096	0.17148	0.048	0.25411	0.096	0.30583
0.0100	0.16762	0.050	0.24783	0.100	0.29777

Table 3.5: Values of ξ critical point for different numbers of hypotheses (Fisher's combination test)

k	Critical point, ξ
2	0.28467
3	0.29916
4	0.30799
5	0.31409
6	0.31863
7	0.32218
8	0.32505
9	0.32744
10	0.32947
11	0.33122
12	0.33275
13	0.33410
14	0.33530
15	0.33639
16	0.33737
17	0.33827
18	0.33909
19	0.33984
20	0.34054
25	0.34339
1000	0.36396
1900	0.36504

Table 3.6: Values of y for different p_i's and levels of α (Fisher's test)

p_i	y, at α =1%	p_i	y, at α =5%	p_i	y, at α =10%
1e-05	0.99999	1e-05	0.99999	1e-05	0.99999
0.0004	0.99999	0.002	0.99999	0.004	0.99999
0.0008	0.99999	0.004	0.99999	0.008	0.99999
0.0012	0.99999	0.006	0.99999	0.012	0.99999
0.0016	0.81824	0.008	0.99999	0.016	0.99999
0.0020	0.65459	0.010	0.87049	0.020	0.99999
0.0024	0.54549	0.012	0.72541	0.024	0.85213
0.0028	0.46756	0.014	0.62178	0.028	0.73039
0.0032	0.40912	0.016	0.54406	0.032	0.63910
0.0036	0.36366	0.018	0.48361	0.036	0.56808
0.0040	0.32730	0.020	0.43525	0.040	0.51128
0.0044	0.29754	0.022	0.39568	0.044	0.46480
0.0048	0.27275	0.024	0.36271	0.048	0.42606
0.0052	0.25177	0.026	0.33481	0.052	0.39329
0.0056	0.23378	0.028	0.31089	0.056	0.36520
0.0060	0.21820	0.030	0.29016	0.060	0.34085
0.0064	0.20456	0.032	0.27203	0.064	0.31955
0.0068	0.19253	0.034	0.25603	0.068	0.30075
0.0072	0.18183	0.036	0.24180	0.072	0.28404
0.0076	0.17226	0.038	0.22908	0.076	0.26909
0.0080	0.16365	0.040	0.21762	0.080	0.25564
0.0084	0.15585	0.042	0.20726	0.084	0.24346
0.0088	0.14877	0.044	0.19784	0.088	0.23240
0.0092	0.14230	0.046	0.18924	0.092	0.22229
0.0096	0.13637	0.048	0.18135	0.096	0.21303
0.0100	0.13092	0.050	0.17410	0.100	0.20451

Chapter 4

COMBINING INDEPENDENT P-VALUES

Summary

We present a new procedure for combining p-values from a set of L independent hypothesis tests. Our procedure is to take the product of only those p-values less than some specified cut-off value and to evaluate the probability of such a product, or a smaller value, under the overall hypothesis that all L hypotheses are true. We give an explicit formulation for this overall p-value, and find by simulation that it can provide high power for detecting departures from the overall hypothesis. Once the overall hypothesis is rejected, an adjustment procedure with strong family-wise error protection is available for individual p-values.

Key words: p-values, multiple tests, Bonferroni.

4.1 Introduction

When L independent tests of the same hypothesis are all conducted with the same significance level α, then the probability of finding at least one significant result among the L is greater than α. The resulting problem of interpreting the results from multiple tests has been considered many times (e.g. Westfall and Young, 1993). It is also recognized (e.g., Rosenthal, 1990) that a series of non-significant results may together suggest significance: "Two 0.06 results are much stronger evidence against the null than one 0.05." This situation led Fisher (1932) to his method for combining significance values, although it is the problem of possibly spurious single significant tests that is of more concern to us. We have been faced with this situation in two genetic contexts: testing for allelic independence at

several loci from several samples (Zaykin et al., 1995), and testing for marker-disease associations at several marker loci (Kaplan et al., 1995).

We concentrate here on methods that combine p-values from several tests in order to provide a p-value for the overall hypothesis that all single hypotheses are true. If this overall hypothesis is rejected, it is then possible to obtain adjusted p-values for the individual tests. These individual adjustments are derived based on the closure principle of Marcus et al (1976) and yield p-values that are at least as small as ones obtained with the step-wise method of Hochberg (1978).

We review the methods of Edgington (1972), Fisher (1932), Stouffer et al. (1949), and Wilkinson (1951) and describe a new procedure, the "truncated product method," that appears to have good power properties. In the language of Hedges and Olkin (1985), the procedures we discuss are termed omnibus or non-parametric because they depend only on the significance values of individual tests and not on the form of the underlying data.

4.2 Previous methods

We suppose that tests have been conducted for each of L hypotheses $H_i, i = 1, 2, \ldots, L$. For each test the p-value p_i is calculated: if H_i is true, this is the probability of observing a test statistic as extreme as or more extreme than the observed value in the direction of rejection (Popper, 1995). However, $(1-p_i)$ is also the value of the probability distribution function of the test statistic and therefore it is uniformly distributed on the interval [0, 1]. This holds for any continuous test statistic. Moreover, $-2 \ln p_i$ has a chi-square distribution with two df. This led

Fisher (1932) to note that the statistic

$$t \;=\; -2\sum_{i=1}^{L} \ln p_i = -2\ln(\prod_{i=1}^{L} p_i)$$

has a chi-square distribution with $2L$ df when all L hypotheses are true. Therefore, the p-value for the hypothesis that all H_i are true is the probability of a χ^2_{2L} variable being greater or equal to the observed value t. We will write this overall hypothesis as H_T.

Edgington (1972) preferred to work with the sum of the p-values, and he gave the probability of the sum of L uniform [0, 1] variables being less than or equal to S as

$$\sum_{r=0}^{S'} (-1)^r \binom{L}{r} \frac{(S-r)^L}{L!}$$

where S' is the largest integer less than S. This probability also serves as the p-value for the hypothesis H_T. Hedges and Olkin (1985) pointed out that one large p-value can overwhelm many small p-values with this approach.

For L independent tests, Wilkinson (1951) noted that the number of p-values less than some quantity τ has a binomial distribution $B(L, \tau)$ when H_T is true. The probability of finding at least k values less than τ is

$$\sum_{i=k}^{L} \binom{L}{i} \tau^i (1-\tau)^{L-i}$$

Wilkinson set τ to the single-test significance level α. He used this binomial expansion to note, for example, that two tests significant at the 5% level in 14 tests does not imply that two events have occurred where each has a 5% probability, but that one event with a 15% probability has occurred. If $k = 1$, the probability becomes $[1-(1-\tau)^L]$ suggesting that $\tau = 1-(1-\alpha)^{1/L}$ for an overall α-level test.

This is the basis for the usual Bonferroni correction. Any individual test must be significant at the level τ in order for the overall level to be α.

An alternative procedure (Stouffer et al., 1949) uses normal-transformed p-values. If $\Phi(x)$ denotes the probability distribution function for the standard normal distribution

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

then each p_i-value can be transformed to a standard normal score, when the hypothesis is true, by

$$\begin{aligned} 1 - p_i &= \Phi(z_i) \\ z_i &= \Phi^{-1}(1 - p_i) \end{aligned}$$

and $z = \sum_i z_i / \sqrt{L}$ is also standard normal. The p-value for hypothesis H_T is, therefore,

$$p = 1 - \Phi\left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L} \Phi^{-1}(1 - p_i) \right)$$

4.3 Truncated product method

As a procedure that combines the features of Fisher's product method and Wilkinson's truncation method, we suggest the use of the product W of all those p_i values that do not exceed some fixed value τ:

$$W = \prod_{i=1}^{L} p_i \Psi(p_i \leq \tau)$$

where

$$\Psi(p_i \leq \tau) = \begin{cases} 1, & p_i \leq \tau \\ 1/p_i, & p_i > \tau \end{cases}$$

We note that this is a special case of the weighted function Q of p-values discussed by Good (1955)

$$Q = \prod_i p_i^{w_i}$$

in which some weights are equal to 1 and some are equal to 0. Good gave the distribution of Q only in the case where all the weights are different.

Under the null hypothesis H_T, the distribution of W can be evaluated by conditioning on the number, k, of the p_i's less than τ:

$$\begin{aligned}
\Pr(W \leq w) &= \sum_{k=1}^{L} \Pr(W \leq w|k) \Pr(k) \\
&= \int_0^w \sum_{k=1}^{L} \frac{(k \ln \tau - \ln t)^{k-1}}{(k-1)!} I(\tau^k > t) \binom{L}{k} (1-\tau)^{L-k} dt \\
&= \sum_{k=1}^{L} \binom{L}{k} (1-\tau)^{L-k} G(w, \tau, k) \qquad (4.1)
\end{aligned}$$

where

$$G(w, \tau, k) = \begin{cases} w \sum_{s=0}^{k-1} \frac{(k \ln \tau - \ln w)^s}{s!}, & w \leq \tau^k \\ \tau^k, & w > \tau^k \end{cases}$$

Theory related to this statistic rests on Simes' (1986) inequality. If the p_i's are ordered, and if the H_i are all true, then with probability $(1-\alpha)$, $p_i > i\alpha/L$ for all $i = 1, 2, \ldots, L$ and for any α between 0 and 1, provided the L test statistics are independent and continuous. The derivation of equation (4.1) is given in Appendix 1.

The individual adjustments are then available through the application of the closure principle of Marcus et al. (1976). Generally, the procedure considers all possible combination hypotheses obtained via intersection of the set of individual hypotheses of interest. If an individual hypothesis and all intersections that contain

it as a component are rejected by an appropriate α-level test, then the closure principle states that the given hypothesis can be also rejected, at the level α. The closure procedure controls the family-wise error rate (FWER) strongly, meaning that FWER $\leq \alpha$ regardless of which subset of null hypotheses happens to be true (Westfall and Young, 1993). The total number of combination hypotheses (N_h) is

$$N_h = \sum_{i=1}^{L} \binom{L}{i} = 2^L - 1 \tag{4.2}$$

which grows quickly with L and often limits applicability of the method. Fortunately, if τ is chosen to be $\leq \alpha$, only $(L+1)$ tests need to be performed. These are the global test with all j p-values that are $\leq \tau$, and the combination hypotheses involving one p-value $\leq \tau$ with the rest $(L-j)$ p-values that are greater than τ. The L combination hypotheses of the size $(L-j+1)$ give p-values bounded by the Sidak's correction with $(L-j+1)$ tests, i.e. the adjusted p-value, p_i^* for $p_i \leq \tau$ is equal to

$$1 - (1 - p_i)^{L-j+1} \tag{4.3}$$

in the case when p_i is exactly τ.

The proof that for $\tau \leq \alpha$ the suggested adjustment satisfies the closure principle consists of inspecting subsets and noting that (4.1) is an increasing function of L and a decreasing function of W, so that rejecting a combination hypothesis H_i of size $(L-j+1)$ automatically leads to rejection of all subsets that contain i, including the set of all L (the global test).

4.4 Comparison of methods

We compared the various procedures described above on the basis of power to detect departures from null hypotheses concerning the mean of a normal distribution, $H_0 : \mu = \mu_0$, when the variance σ^2 is known. The alternative hypotheses are $H_A : \mu = \mu_A > \mu_0$. We transform the mean $\bar{x}$ of a sample of size n to give the usual z test statistic

$$z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$$

Under H_0, z has the standard normal distribution, and under H_A it is distributed $N(\gamma, 1)$ where $\gamma = \sqrt{n}(\mu_A - \mu_0)/\sigma$.

When H_0 is true, the p-values could be written as p_0:

$$\begin{aligned} p_0 &= \Pr(Z \geq z) \\ z &= \Phi^{-1}(1 - p_0) \end{aligned}$$

where Z is the standard normal variable. The values of p_0 are distributed uniformly on [0, 1]. When H_A is true the p-value is

$$\begin{aligned} p_A &= \Pr(Z \geq z - \gamma) \\ &= 1 - \Phi\left(\Phi^{-1}(1 - p_0) - \gamma\right) \end{aligned}$$

Therefore, if u is drawn randomly from the uniform distribution on [0,1], p-values under the null and alternative hypotheses can be calculated as $(1 - u)$ and $\Phi\left(\Phi^{-1}(1 - u) + \gamma\right)$. This provides us with a means for generating p-values under each hypothesis rather than generating samples, calculating test statistics and then determining p-values.

4.5 Results

In the first place we simulated the case where all L hypotheses were true, and rejected the overall hypothesis H_T that all are true when the overall p-value was less than 0.05.

In Table 1 we show the proportion of 100,000 simulations that resulted in rejection, for two values of τ, 0.05 and 0.5. The truncated product method gave values consistent with the nominal 0.05 value .

We then allowed h_A of $L = 25$ hypotheses to be false, but with the same value of γ so that the tests of each of these h_A hypotheses has the same power. We set $\gamma = 1.64$ so that these powers had the nominal value of 0.50. All other $L - h_A$ hypotheses were set to be true. The p-value for the overall hypothesis H_T was then calculated by the methods of Edgington, Fisher, Stouffer, Wilkinson (with $\tau = 0.05$) and by the truncated product method with $\tau = 0.05$. Empirical powers are shown in Table 2. Fisher's procedure and the truncated product method had substantially the same performance and the other three methods did not perform well for low numbers of false hypotheses.

The alternative hypotheses h_A are all one sided, but in Table 3 we allow some tests to have positive γ and some negative γ. With this combination of different directions of departure from the null, it is clear that the truncated product method performs the best. Wilkinson's procedure now performs better than Fisher's procedure, but the other two methods do not perform well.

As a final example we present an analysis of a set of genetic data (Scholl et al., 1996). Likelihood ratio tests for Hardy-Weinberg equilibrium (Weir, 1996)

were conducted at seven loci in samples from three Native American populations. The p-values are shown in Table 4 and, as is common in such situations, some of them are low. In the first place, we combined p-values over samples for each locus. This addresses the hypothesis that none of the three sampled populations have departures from Hardy-Weinberg, and so is locus-specific. Secondly, we combined p-values over loci for each sample to address the hypothesis that none of the loci within a sampled population departs from Hardy-Weinberg. This procedure is population-specific. For the whole set of p-values we have two possibilities. We could combine p-values over all 21 tests to address the hypothesis that none of the loci departed from Hardy-Weinberg in any of the sampled populations, and these p-values are shown in the table. Alternatively, we could use the locus-specific p-values to make statements about each locus separately but then we may want to use the procedures of this paper to give a p-value for the hypothesis that none of the loci departed from Hardy-Weinberg. With Fisher's procedure, the overall p-value for loci is 0.281 and with the truncated product method it is 0.276. We could also operate on the population-specific values in order to make statements about the populations. With Fisher's procedure, the overall p-value is 0.143 and with the truncated product method it is 0.173. Note that we have ignored the possibility of between-locus dependencies in these tests.

The combination of p-values in Table 4 provides a more satisfactory interpretation of the data than simply ignoring the values in excess of 0.05, and using a Bonferroni correction to claim that the remaining low value (GYPA for Navajo) is not significant. There is support from the truncated product method for regarding the GYPA value to be significant, and from Fisher's method for regarding the

Navajo population as having departures from Hardy-Weinberg.

4.6 Conclusions

Debate over the use of p-values in summarizing the results of hypothesis tests continues (Hagen, 1997) but the values are routinely calculated and reported in many disciplines. Many genetic studies lead to p-values from tests on multiple loci in multiple samples and it is necessary to take this multiplicity into account when drawing inferences. Among the approaches that have been considered in the past it appears that Fisher's method, of taking minus twice the logarithm of the product of a set of L p-values and recognizing that this quantity is distributed as chi-square with $2L$ df when all L hypotheses are true, is a good way of providing an overall p-value.

We have suggested here an alternative procedure, which also takes into account the size of L p-values but uses only the small ones. By "small" we mean a conventional values such as 0.05. Instead of asking whether the set of L tests contains any evidence for departures from all L hypotheses, our procedure can be used to ask whether any of significant test statistics are indeed significant. By focusing on only the set of small p-values we appear to have increased the overall power, especially in situations where a small subset of the hypotheses are false but false in opposite directions.

An important feature of the method is that individual p-value adjustments, and also adjustments for pairs, triples, etc, are readily available from the application of the closure principle. This will be useful in exploratory analysis, where many

tests are involved. In exploratory analysis, some of the effects might be declared to be likely "real" and corresponding experiments worth of a replication. It should be noted, that even methods with the strong FWER protection have larger proportions of falsely rejected hypotheses than α, given that some hypotheses have been in fact rejected. This increase of the FWER, conditional upon rejection, can be large (Zaykin et al., 1999) and it is reasonable to allow that, in those studies where positive results are found, a potentially large percentage of them might be in error.

The truncated product method also has potential outside of the genetic context. For example, in meta-analysis of published data there is a well known problem of the "publication bias", when only successful findings are reported. Models have been developed for estimating the total number of studies (e.g. Gleser and Olkin, 1996), and therefore an adequate inference can be made using the truncated product method.

Acknowledgments

This work was supported in part by NIH Grant GM45344. Helpful advice was given by Drs. R. Berger, S. Ghosh and L. Stefanski.

4.7 References

Edgington, E.S. (1972). An additive method for combining probability values from independent experiments. *J. Psychology* 80:351–363.

Fisher, R.A. (1932). *Statistical Methods for Research Workers.* Oliver and Boyd, London.

Good, I.J. (1955). On the weighted combination of significance tests. *J. Roy. Stat. Soc. B* 17:264–265.

Hagen, R.L. (1997). In praise of the null hypothesis statitical test. *Am. Psychologist* 5:15–24.

Hedges, L.V. and I. Olkin. (1985). *Statistical Methods for Meta- Analysis.* Academic Press, New York.

Hochberg Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70.

Kaplan, N.L. and B.S. Weir. (1995). Are moment bounds on the recombination fraction between a marker and a disease locus too good to be true? Allelic association mapping revisited for simple genetic diseases in the Finnish population. *Am. J. Human Genetics* 57:1486–1498.

Marcus, R., Peritz, E., and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 77 63:655–660.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bull.* 85:185–193.

Scholl, S., B. Budowle, K. Radecki and M. Salvo. (1996). Navajo, Pueblo, and Sious population data on the loci HLA-DQA1, LDLR, GYPA, HBGG, Gc, and D1S80. *J. Forensic Sci.* 41:47–51.

Shaffer, J.P. 1995. Multiple hypothesis testing: A review. *Ann. Rev. Psychology* 46:561–584.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751–754.

Stouffer, S.A., E.A. Suchman, L.C. DeVinney, S.A. Star and R.M. Williams, Jr. (1949). *The American Soldier, Vol. 1. Adjustment During Army Life.* Princeton Univ. Press, Princeton.

Westfall, P.H. and S.S. Young. (1993). *Resampling-Based Multiple Testing.* Wiley, New York.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bull.* 48:156–158.

Zaykin, D., L. Zhivotovsky and B.S. Weir. (1995). Exact tests for association

between alleles at arbitrary numbers of loci. *Genetica* 96:169–178.

Zaykin, D, Young, S.S. and P.H. Westfall. (1999). Using false discovery rate approach in the genetic dissection of complex traits: a response to Weller et al. (submitted to *Genetics*).

4.8 Appendix 1

When H_0 is true and $\tau < 1$, the number of small p-values (k) has a binomial distribution, and p_i's are observations from the uniform $(0, 1)$ distribution, truncated at τ (i.e. the distribution of p_i's is uniform on $(0, \tau)$).

Given k, the conditional distribution of the product (W) can be calculated directly. Let $X_1, ..., X_k$ be independent uniform $(0, \tau)$ random variables. Consider the transformation:

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= X_1 X_2 \\ &... \\ Z_k &= X_1 X_2 ... X_k \end{aligned}$$

So the inverse is

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= Z_2/Z_1 \\ &... \\ X_k &= Z_k/Z_{k-1} \end{aligned}$$

The Jacobian of the transformation ($\mathbf{J}$) has the following structure:

$$\partial x_i/\partial z_j = \begin{cases} 1 & i = j = 1 \\ 1/z_{i-1} & i = j \geq 1 \\ -z_i/z_{i-1}^2 & j = i - 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$\mid \mathbf{J} \mid = \prod_{\mathrm{i}}^{\mathrm{k-1}} 1/\mathrm{z_i}$$

and the joint density is

$$f(\mathbf{Z}) = \frac{1}{\tau^{\mathrm{k}} \prod_{\mathrm{i}=1}^{\mathrm{k}-1} \mathrm{z_i}}$$

Integrating out z_1 through z_{k-1} from the joint density gives the conditional probability, $P(W \leq w \mid k)$:

$$\begin{aligned} P(W \leq w \mid k) &= \int_0^w \left[\int_t^{\tau^k} \int_{z_{k-1}}^{\tau^k} \ldots \int_{z_2}^{\tau^k} \frac{\prod_{i=1}^{k-1} dz_i}{\tau^k \prod_{i=1}^{k-1} z_i} \right] dt \\ &= \int_0^w \frac{(\log \tau^k - \log t)^{k-1}}{(k-1)!\ \tau^k} dt \end{aligned} \tag{4.4}$$

Then the unconditional distribution is found as follows:

$$\begin{aligned} P(W \leq w) &= \int_0^w \sum_{k=1}^{L} \frac{(\log \tau^k - \log t)^{k-1}}{(k-1)!\ \tau^k} \\ &\times \quad I(\log \tau^k > \log t) \binom{L}{k} \tau^k (1-\tau)^{L-k} dt \end{aligned} \tag{4.5}$$

The probability calculated in (4.5) corresponds to the combined p-value. After τ^k in (4.5) is cancelled, this probability is

$$\begin{aligned} P(W \leq w) &= \int_0^w \sum_{k=1}^{L} \frac{(k \log \tau - \log t)^{k-1}}{(k-1)!} \\ &\times \quad I(\tau^k > t) \binom{L}{k} (1-\tau)^{L-k} dt \end{aligned} \tag{4.6}$$

or equivalently

$$\begin{aligned} P(W \leq w) &= \sum_{k=1}^{L} \binom{L}{k} \frac{(1-\tau)^{L-k}}{(k-1)!} \left[\int_0^w (k \log \tau - \log t)^{k-1} \right. \\ &\times \quad \left. I(\tau^k > t) dt \right] \end{aligned} \tag{4.7}$$

Provided $\tau^k > t$, the integral in (4.7), which we denote by I_k is:

$$\begin{aligned} I_k &= \int_0^w (\log \tau^k - \log t)^{k-1} dt \\ &= (\log \tau^k - \log t)^{k-1} t|_0^w \end{aligned} \tag{4.8}$$

$$
\begin{aligned}
&- \int_0^w t d[(\log \tau^k - \log t)^{k-1}] && (4.9)\\
&= w(\log \tau^k - \log w)^{k-1} \\
&- (k-1)\int_0^w t(-\frac{1}{t})(\log \tau^k - \log t)^{k-2} dt && (4.10)\\
&= (k-1)I_{k-1} + wA(\tau,k,w)^{k-1}. && (4.11)
\end{aligned}
$$

where $A(\tau,k,w) = k\log\tau - \log w$. Since $I_1 = w$, then

$$
\begin{aligned}
I_k &= (k-1)!\left[w + w\sum_{s=1}^{k-1}\frac{A(\tau,k,w)^s}{s!}\right] && (4.12)\\
&= w(k-1)!\sum_{s=0}^{k-1}\frac{A(\tau,k,w)^s}{s!} && (4.13)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
P(W \le w) &= w\sum_{k=1}^{L}\left(\begin{array}{c} L \\ k \end{array}\right)\frac{(1-\tau)^{L-k}}{(k-1)!}(k-1)!\sum_{s=0}^{k-1}\frac{A(\tau,k,w)^s}{s!} && (4.14)\\
&= w\sum_{k=1}^{L}\left(\begin{array}{c} L \\ k \end{array}\right)(1-\tau)^{L-k}\sum_{s=0}^{k-1}\frac{A(\tau,k,w)^s}{s!} && (4.15)\\
&= w\sum_{k=1}^{L}\sum_{s=0}^{k-1}\left(\begin{array}{c} L \\ k \end{array}\right)(1-\tau)^{L-k}\ \frac{A(\tau,k,w)^s}{s!} && (4.16)
\end{aligned}
$$

Table 1 Power levels when all L hypotheses are true.

	L					
Method	2	3	5	10	25	50
Trunc. Prod., $\tau = 0.05$	0.046	0.045	0.050	0.051	0.050	0.049
Trunc. Prod., $\tau = 0.5$	0.049	0.049	0.052	0.048	0.053	0.049

Table 2 Power levels when h_A of $L = 25$ hypotheses are false, and $L - h_A$ are true.

	h_A					
Method	1	2	3	4	5	6
Edgington	0.077	0.124	0.186	0.266	0.361	0.470
Fisher	0.119	0.229	0.321	0.502	0.678	0.793
Trunc. Prod., $\tau = 0.05$	0.158	0.265	0.401	0.538	0.653	0.774
Stouffer	0.091	0.159	0.256	0.373	0.502	0.628
Wilkinson	0.075	0.137	0.234	0.355	0.482	0.604

Table 3 Power levels when h_A of $L = 25$ hypotheses are false in each direction and $L - 2h_A$ are true.

	h_A					
Method	1	3	5	7	9	11
Edgington	0.043	0.036	0.027	0.019	0.011	0.005
Fisher	0.093	0.235	0.423	0.617	0.771	0.879
Trunc. Prod., $\tau = 0.05$	0.119	0.353	0.598	0.784	0.891	0.949
Stouffer	0.048	0.051	0.050	0.049	0.050	0.049
Wilkinson	0.066	0.199	0.413	0.638	0.797	0.900

Table 4 p-values for likelihood ratio tests of Hardy-Weinberg equilibrium among three pairs of populations at seven genetic loci.

	Population				
Locus	Navajo	Pueblo	Sioux	Fisher	Trunc. Prod.
LDLR	0.377	0.397	0.599	0.566	0.142
GYPA	0.014	0.470	1.000	0.122	0.045
HBGG	0.136	0.168	0.790	0.235	0.142
D7S8	0.052	1.000	0.804	0.385	0.142
GC	0.259	0.124	0.213	0.125	0.142
HLA-DQA1	0.438	0.368	0.562	0.569	0.142
D1S80	0.750	0.559	0.211	0.563	0.142
Fisher	0.031	0.430	0.812	0.216	
Trunc. Prod.	0.116	0.302	0.302		0.388

Chapter 5

DETERMINING TRUE EFFECTS IN GENE EXPRESSION DATA

5.1 Abstract

DNA microarrays provide the possibility to simultaneously monitor the expression levels of many genes. The most inclusive technique is currently available for the yeast (*Saccharomyces cerevisiae*) genome, where expression of all genes can be studied. The arrays of tens of thousands of genes are becoming common for mammalian genomes and both the number of genes and the sensitivity are continuing to improve. One of the applications is to look for the changes in gene expression in response to drug treatments, where treated and control samples are compared for the abundance of mRNA, measured by the fluorescent intensity. Changes in the intensities between samples could be attributed to the experimental error or to the presence of a real effect of the treatment. To distinguish the true effects from the noise, statistical tests should be carried out for all pairs of intensities. This results in a large number of tests and the multiple testing problem becomes an issue. In this paper we discuss two multiple testing procedures for localizing subsets of true effects. These procedures are intermediate between individual adjustment methods and methods based on combining p-values.

5.2 Introduction

This paper studies performance of two methods for dealing with the multiple testing problems when the number of tests is very large. The usage of our methods is illustrated by applying them to data from comparisons of gene expression changes in Hepatocellular carcinoma cells, untreated and treated with a drug.

Both methods are based on combining subsets of "small" p-values that are outcomes of statistical tests for the difference in measured intensities. A test based on the distribution of the ratio of two intensities has been proposed by Chen et al. (1997), however it relies on the distributional assumptions that might not always hold. For this study, we use p-values obtained with the method of Kepler et al. (1999), who developed a robust semi-parametric normalization technique based on the local regression smoothing.

Multiple testing methods divide into two extremes, the ones that make statements about individual null hypotheses $H_1, ..., H_L$ and ones that are concerned with the global null hypothesis, $\cap_i^L H_i$. The former are traditionally required that they control the family-wise error (FWE). Westfall and Young (1993) argue that it is also desirable that the FWE is controlled in the strong sense. Specifically, for the significance level α, a multiple testing procedure controls the FWE in the strong sense if $FWE \leq \alpha$ for all subsets of hypotheses, regardless of which of them are true. The distinction between the strong and the weak control is thus only the issue in the situation of the "partial null hypothesis", i.e. when some of the hypotheses are false. A common method is the single-step Bonferroni technique. It rejects H_i if $p_i \leq \alpha/L$. Several step-wise procedures have been proposed. A method of Holm (1979) is an example of a step-down procedure. It is based on the Bonferroni inequality and the $\min p$ statistic. A method of Hochberg (1988) is a further improvement and a sequential adaptation of the Bonferroni technique. Hochberg's method is to order p_i's, start with $i = L$ and once $p_j \leq \alpha/(L - j + 1)$, then reject all H_i for $i \leq j$.

Benjamini and Hochberg (1995) proposed a different approach, which is to

maintain the "false discovery rate", or the expected proportion of false rejections. The Benjamini and Hochberg's procedure (BH) is to order p_i's and if $p_j \leq j\alpha/L$, then reject all H_i for $i \leq j$. This error rate is equivalent to the FWER when all hypotheses are true. Clearly, the BH method controls the FWER in the weak sense, however the gain in power to detect true effects can be substantial. The BH method is most appropriate when the number of true effects is known to be large, a moderate amount of the false positive results can be tolerated, or if a follow-up study is anticipated (Weller et al., 1998, Drigalenko and Elston, 1998).

It should be stressed that the BH method does not provide information about the proportion of false positive results for any particular experiment where some of the hypotheses were in fact rejected. Benjamini and Hochberg (1995) stated that "a desirable error rate to control may be the expected proportion of errors among the rejected hypotheses, which we term the false discovery rate (FDR)" and designed a method that controls the false discovery rate unconditionally, weighting FDR by the probability of at least one rejection:

$$\text{FDR} = E\left\{\frac{F}{T+F} \mid T+F \geq 1\right\} \Pr(T+F \geq 1) \tag{5.1}$$

where T, F are the numbers of true and false rejections. The BH method controls this false discovery rate in an *unconditional* manner. Benjamini and Hochberg readily acknowledge that their method cannot control FDR, conditional upon having rejected one or more hypotheses. Not realizing this could lead to some confusion (Zaykin et al., 1999a). It is easy to see that controlling $E\left\{\frac{F}{T+F} \mid T+F \geq 1\right\}$ would require knowledge of the unknown number of true null hypotheses and the power of the test to detect true effects. Consider applying the BH method in the case

when there is exactly one true effect among L hypotheses. The proportion of false discoveries among rejected hypotheses will be at least 50% whenever any of the p-values corresponding to false positives is smaller than the one corresponding to the true effect. The frequency of that happening will depend on whether the distribution of the p-value corresponding to the true effect is stochastically greater than the distribution of the first order statistic from the uniform distribution. Because this distribution is

$$p_{1:L} \sim \text{Beta}(1, L) \tag{5.2}$$

it follows that FDR* will depend on both L and the power to detect the single true effect. We showed by simulations that the expected proportion of false positive results among the rejected hypotheses can increase rather dramatically (Zaykin et al., 1999a). FDR control allows that false positives will occur, in fact they are *expected* in any given study. However, given that a significance has been found, the implied operational interpretation that $(1 - \alpha)$ 100 % of the claimed results will replicate cannot be made, since a smaller percentage is expected to replicate in reality. The problem is more pronounced as the total number of true null hypotheses increases, and thus controlling FDR at a level α allows no clear-cut interpretation.

The combination methods, such as of Fisher (1932), Stouffer et al. (1949) and Edgington (1972) are testing the global null hypothesis, $\cap_i^L H_i$. Once the hypothesis is rejected, it is only possible to conclude that one or more H_i's are false. This is somewhat less so with the Fisher's method, since it is disproportionally influenced by the tests resulting in small p-values (Rice, 1990).

In the next section we review and develop two methods based on distributions of first order statistics that accentuate this property of the Fisher's method by considering a smaller subset of hypothesis $\cap_i^k H_i$, $1 < k < L$. Depending on the value of k, these methods are closer to global methods or methods for individual adjustments. As with the BH method, they control the FWER in the weak sense.

5.3 Methods based on distributions of first order statistics.

5.3.1 Testing the overall hypothesis

When the proportion of true effects is small, both the type-I error rate and the proportion of false positives within the class of rejected hypotheses can be quite high. Then the only sensible conclusion that can be reached is that there are one or more true effects among p-values that satisfy the FWER or FDR conditions. Here we review and develop two methods that address the question of whether there is overall evidence against the null hypothesis among the subset of either "small" or "first k smallest" p-values.

The "truncated product method" (τ-method of Zaykin et al., 1999b) rejects the null hypothesis that there are no true effects among the L tests. The alternative hypothesis is that there is collective evidence that some of the effects are present among the tests with "small" p-values. If the null hypothesis is rejected, the conclusion is more vague than that of the individual adjustment techniques.

The τ-method makes use of the distribution of the product W of all those p_i

values that do not exceed some fixed value $0 < \tau \leq 1$:

$$W_\tau = \prod_{i=1}^{L} p_i^{I(p_i \leq \tau)} \tag{5.3}$$

Under the null hypothesis H_0, the distribution of W_τ is

$$\begin{aligned} \Pr(W_\tau \leq w) &= \sum_{k=1}^{L} \binom{L}{k} (1-\tau)^{L-k} \\ &\times \left(w \sum_{s=0}^{k-1} \frac{(k \ln \tau - \ln w)^s}{s!} I(w \leq \tau^k) + \tau^k I(w > \tau^k) \right) \end{aligned} \tag{5.4}$$

where $I(\cdot)$ is an indicator function. The combined p-value is thus calculated from the equation (5.4).

When τ is set to 1, it becomes equivalent to the Fisher's multiplicative method. Note that instead of looking up the combined p-value from the tail of a chi-square distribution, it can be calculated directly as

$$\Pr(W_\tau \leq w) = w \sum_{s=0}^{L-1} \frac{(-\ln w)^s}{s!} \tag{5.5}$$

When the number of tests is large, it is necessary to resort to the following Monte Carlo algorithm to avoid numerical overflows. The algorithm for estimating (5.4) is as follows:

1 Decide on the value of the truncation point, τ.

2 Calculate $W_0 = \prod_i^L p_i^{I(p_i \leq \tau)}$ for a sample of p_i's (L observed p-values). Set $C = 1$, $R = 0$.

2a Generate L independent uniform $(0, 1)$ random numbers, $u_1, ..., u_L$.

2b Calculate $W_C = \prod_i^L u_i^{I(u_i \leq \tau)}$.

2c If $W_C \leq W_0$, increment R by one. Increment C by one.

6 Repeat steps (2a)-(2c) B times.

7 The empirical combined p-value is R/B.

With this algorithm, it is possible to calculate combined p-values for very large number of tests (logs of products might have to be taken). The empirical p-value obtained with it converges to the one obtained with (5.4) as B increases.

The algorithm can be naturally extended for dealing with non-independent p-values just by generating *correlated* uniforms at step **3**. Denote $\mathbf{U} = (u_1, ..., u_L)'$, a vector of iid uniform(0,1) random variables and let $\mathbf{C}$ be the Cholesky factor of the correlation matrix of p_i's. A vector of correlated uniform random variables ($\mathbf{U}^*$) is then obtained as

$$\mathbf{U}^* = \mathbf{1} - \Phi\left\{\mathbf{C}\ \Phi^{-1}\left(\mathbf{1} - \mathbf{U}\right)\right\} \tag{5.6}$$

where Φ is the standard normal CDF and Φ^{-1} is its inverse, applied to the elements of the vector component-wise. In some situations an alternative way is to sample from a stochastic process. For example, if tests are ordered in some natural way and there is an exponential decay in the correlation, sampling from the Ornstein-Uhlenbeck diffusion process can be used to impose the correlation. The stochastic differential equation for the Ornstein-Uhlenbeck diffusion is

$$\frac{dX}{dt} = -aX(t) + \sigma\tilde{\xi}(t), \tag{5.7}$$

where $\tilde{\xi}(t)$ is the white noise term and a, σ are the drift and the variance parameters. The diffusion increments for some small Δt can be generated as $-aX(t)\Delta t +$

$\sigma Z\sqrt{\Delta t}$, where $Z \sim N(0,1)$. For this process, the autocorrelation function, the mean and the variance are

$$\begin{aligned} r(X(t),X(s)) &= X(0)e^{-a(t+s)} + \frac{\sigma^2}{2a}\left(e^{-a(t-s)} - e^{-a(t+s)}\right) \\ &\xrightarrow{t} \frac{\sigma^2 e^{-a(t-s)}}{2a} \end{aligned} \tag{5.8}$$

$$E(X(t)) = xe^{-at} \xrightarrow{t} 0 \tag{5.9}$$

$$\mathrm{Var}(X(t)) = \frac{\sigma^2}{2a}(1-e^{-2at}) \xrightarrow{t} \frac{\sigma^2}{2a} \tag{5.10}$$

so, for example, setting $X(0) \sim N(0,1)$, $a = 2$, $\sigma^2 = 4$, sampling a vector $\{X(t_i)\}$ from the process, $i = 1, ..., L$, and applying $\Phi^{-1}(\cdot)$ to each observation allows to generate a set of $U(0,1)$ random variables with the auto-correlation function given by $e^{-2(t-s)}$.

Here we introduce another method based on the product of k smallest p-values (κ-method). The main difference between the τ and κ methods is that the number of p-values over which the product is taken is a binomial random variable in the τ-method, whereas this number is fixed (equal to k) in the second method.

Let $V_{1:L}, V_{2:L}, ..., V_{k:L}$ be the first k order statistics in a sample of p-values. When H_0 is partially true, i.e. some of the tests represent true effects, the distribution of the order statistics is obtained as follows. Let g define a subset of indices and $|g|$ be the length of that subset. Let $F_i(x)$ be the distribution function for the p-value of ith test. Then by summing over subsets of k or greater lengths, we obtain the distribution of an individual kth order statistic in a sample of non-identically distributed independent random variables:

$$
\begin{aligned}
P(V_{k:L} \le x) &= \sum_{j=k}^{L} \sum_{g,|g|=j} \left(\prod_{i \in g} F_i(x) \right) \left(\prod_{i \notin g} (1 - F_i(x)) \right) \\
&= \sum_{j=k}^{L} \sum_{g,|g|=j} \prod_{i}^{L} (2F_i(x) - 1) I(i \in g) + 1 - F_i(x) \\
&= \sum_{j=k}^{L} \sum_{g,|g|=j} \prod_{i}^{L} (1 - 2F_i(x)) I(i \notin g) + F_i(x) \qquad (5.11)
\end{aligned}
$$

Under complete H_0, individual p-values have the uniform(0,1) distribution and the distribution of $V_{k:L}$ is Beta$(k, L-k+1)$. Then the distribution of the product of first k order statistics (k smallest p-values) is given by

$$
\begin{aligned}
P(W_k \le w \mid k) &= 1 - \binom{L}{k} \int_w^1 (1 - x^{\frac{1}{k}})^{L-k} \\
&\times \left[1 - \frac{w}{x} \sum_{s=0}^{k-2} \frac{\left(\log \frac{x}{w} \right)^s}{s!} \right] dx \qquad (5.12)
\end{aligned}
$$

(see Appendix 1 for the derivation). The first term corresponds to the distribution of the kth power of a Beta random variable, $V_{k:L} \sim \binom{L}{k}(1-x)^{L-k}\ kx^{k-1}$. Its distribution is

$$
(V_{k:L})^k \sim \binom{L}{k} (1 - x^{\frac{1}{k}})^{L-k} \qquad (5.13)
$$

and the rest of (5.12) is a convolution of smaller order statistics, $V_{j<k:L}$. The probability in (5.12) defines the combined p-value of the κ-method.

An application of the closure principle of Marcus et al. (1976) shows that individual adjustments with the strong FWER control, $V^*_{i:L}$, are available as

$$
V^*_{i:L} = 1 - (1 - V_{i:L})^{L-k+1} \qquad (5.14)
$$

which is a smaller number than the one that would be obtained with step-wise procedures of Hochberg or Holm. This improvement is obtained at the cost that k must be specified in advance.

As with the τ-method, we recommend to compute (5.12) through the following Monte-Carlo algorithm when L is large:

1 Decide on k.

2 Calculate $W_0 = \prod_i^k p_i$ for a sorted sample of p_i's (L observed p-values). Set $C = 1$, $R = 0$.

2a Generate L (possibly correlated) and sorted uniform $(0, 1)$ random numbers, $u_1, ..., u_L$.

2b Calculate $W_C = \prod_i^k u_i$.

2c If $W_C \leq W_0$, increment R by one. Increment C by one.

6 Repeat steps (2a)–(2c) B times.

7 The empirical combined p-value is R/B.

More generally, one can specify individual weights for p-values (powers of p_i's). If $k = L$, this reduces to the method of Good (1955), who gave the distribution of $\prod_i^L p_i^{v_i}$, for the case of all v_i's being different from each other. The Monte-Carlo method removes this restriction. Weights can be derived based on the sample sizes of individual tests. Good (1992) suggested that p-values can be standardized to a case with 100 observations by multiplying each p-value by a $\sqrt{N/100}$.

An improvement in speed of the algorithm can be achieved if u_i's are stored in an ordered manner. Such implementation does not need to keep all L generated

p-values in the computer memory, but only k smallest ones. A C++ example of such function is given in Appendix 2.

5.4 Determining true effects

Once the overall null hypothesis is rejected, it is important to determine which effects are most likely to be the true ones. The BH method reviewed above provides the FDR control, but because of its unconditional nature the utility of the method is limited to cases when the proportion of true effects is high as well as the individual powers to detect them (Zaykin et al., 1999a). In situations when only a very small proportion of null hypotheses is false, the BH method can still be used in an exploratory manner. Here we suggest another powerful algorithm for determining the subset of p-values that are likely to represent true effects.

The W_τ and W_k random variables defined above are the products of p-values. It can happen that all p-values from the ordered subset $p_{(1)}, ...p_{(j)}$ are small enough so that if $W_\tau = p_{(i \leq j)}$ or $W_k = p_{(i \leq j)}$, the overall hypothesis is still rejected. It is therefore possible to successfully divide the original product, W, by $p_{(i)}$'s, starting with $p_{(1)}$, effectively removing p-values from it. As long as the combined probability remains significant, the removed p-value is declared a true effect. We define this as a step-up procedure, in consistency with the notation of Benjamini and Hochberg (1995). The theoretical justification of this scheme follows from the closure principle. The described step-wise procedure arises as a closing method with an early stop. Algebraic forms of the distributions used in τ and κ methods provide shortcuts so that not all $2^L - 1$ intersection hypotheses need to be considered.

Although it is possible to control the FWER in the strong sense with this method (Zaykin et al., 1999b), here we are concerned with increasing the power, allowing for the weak control of the error rate of individual hypotheses. Note that the FWER for the remaining subset (considered as as a single intersection hypothesis) is controlled strongly.

5.4.1 Results from simulations

We performed a series of simulations illustrating performance of the algorithm for determining true effects and compared power, expected proportions of true discoveries and the false discovery rate within the class of rejected hypothesis proposed methods as well as the Hochberg's and the BH methods. All tests were performed at the 5% α-level and different total numbers of tests (L) and individual power (β) were compared. All tests maintained the declared α-level under complete null hypothesis (results not shown).

Tables 5.1-5.3 show results of simulations with 50,000 tests. Only the overall hypothesis was tested by combination tests, with τ set to 0.002 and k set to 100. H_A is the number of true effects in each simulation. All numbers in the tables are averages over 1000 simulations. Table 5.1 shows expected proportion of found true effects. Table 5.2 shows the power values, i.e. the average numbers of times per simulation when one or more true effects were found. Table 5.3 is the conditional false discovery rate. The overall tests make no attempt to adjust p-values individually and the values of the conditional FDR are high. These can be brought down by setting smaller values of τ and k. Note that both H and

BH methods can have quite large proportion of false positives among rejected hypotheses.

Tables 5.4-5.6 are results of simulations with smaller numbers of tests. They show results for overall tests and the step-up algorithm (τ-method) and corresponding results for the H and BH methods. The κ-method has not been included due to its high computational demand, however we expect results to be similar to those obtained for τ-method. Note that values of $E\{I(T > 0)\}$ do not need to be recomputed as the probability of rejecting one or more hypotheses remains the same. The simulation results show that the proportion of false discoveries can be largely reduced by using the step-up algorithm and the expected proportion of found true effects remains high.

5.4.2 Application to microarray data

We applied our algorithms to p-values obtained from tests on mRNA expression changes in Mesothelinoma cells after four hours exposition to a drug. The data were averaged across two arrays. There were 3567 tests in total. We had a repetition of the experiment, but performed with a single array.

We used $\tau = 0.001$ and $k = 50$. Both tests reported an overall p-value less than 0.0001. The step-up algorithm with τ-method declared 22 out of 44 "small" p-values as "true effects". The κ method declared 18 out of 50 as true effects. Both, H and BH methods declared only a single effect corresponding to the Abelson murine leukemia gene as true (raw p-value 0.944310^{-10}). This gene replicated with p-value 0.001. Four other effects replicated with p-values less than 0.05, seven

others gave p-values less than 0.1, and all p-values but one were smaller than 0.5. There was no repetition for one of the genes. Table 5.7 shows results of the analysis. The Abelson murine leukemia gene is not included in the table. The first two columns are unadjusted p-values corresponding to the two experiments. The last column contain p-values combined across both experiments. They are adjusted for both 3567 tests from the first experiment and for the 21 repetitions. The adjusted p-value (p_c) was calculated according to Zaykin and Young (1999c) as

$$p_c = 1 - L\binom{n}{k}\int_w^1 \Big\{(1-w)^{n+L-k}w^{k-1} \tag{5.15}$$

$$\times \int_0^1 \frac{(1-x)^{n-k}x^{L-1}}{[1-x(1-w)]^n}dx\Big\}\,dw \tag{5.16}$$

where n is the number of p-values that have been replicated, $m = n-k+1$, where k is the rank of p-value from the replication, and W is the product of two p-values from both experiments.

The overall significance for last two columns was obtained with Fisher's combined probability method, yielding very small p-values. Results in the table 5.7 indicate that data contained true effects that were missed by H and BH methods. To explore this situation further, we conducted a series of simulations in the way that has been described above. We set $L = 3567$ (the number of tests in the first experiment). We used $\beta = 0.40, 0.50, 0.80$ and numbers of true effects varying from 15 to 100 (tables 5.8 – 5.10). The results show that the power for the individual tests cannot be as high as 0.80. The false discovery rate is low for all methods. As the τ-method reaches the expected number of true discoveries around 20, all three methods attain very high power (5.9) and the expected number of true discoveries

for H and BH methods is also high. In fact, when $\beta = 0.80, H_A = 25$, the probability that the BH method finds 2 or less true effects is only 2% (data not shown). The data is most consistent with the situation of the average power between 0.40 and 0.50 and the number of genes that responded to the treatment between 50 and 75. Indeed, BH and H methods reject about one test and the τ-method rejects about 20 tests in such situation.

5.5 Conclusions

We suggest that "intermediate" combination methods have good potential in exploratory analysis. Using an expression array data set, we demonstrate how to make inferences about the average numbers of true effects in the class of rejected hypotheses. Combination methods perform well when used together with simulations and methods of Hochberg (1988) and Benjamini and Hochberg (1995). We showed that it is possible to infer the plausible number of false null hypotheses from comparisons of numbers of rejected hypotheses made by individual adjustments methods and by combination methods.

5.6 Appendix 1

Let $V_{1:L}, V_{2:L}, ..., V_{k:L}$ be first k order statistics from $U(0,1)$. We are looking for the distribution of $W_k = V_{1:L}V_{2:L}...V_{k:L}$. Write $V_i \equiv V_{i:L}$ and

$$W_k = \left(\frac{V_1}{V_2}\right)\left(\frac{V_2}{V_3}\right)^2 ... \left(\frac{V_{k-1}}{V_k}\right)^{k-1} V_k^k \tag{5.17}$$

$$\equiv U_1\ U_2^2\ ...\ U_{k-1}^{k-1}\ V_k \tag{5.18}$$

$U_1\ U_2^2\ ...\ U_{k-1}^{k-1}\ U_k$ are independent and uniformly distributed on (0-1), since

$$U_i \sim \text{Beta}(i, 1) \tag{5.19}$$

and

$$U_i^i \sim \text{Beta}(1, 1). \tag{5.20}$$

V_k^k is a kth power of a Beta$(k, L-k+1)$ random variable with the following density

$$V_k^k \sim \binom{L}{k}(1 - x^{\frac{1}{k}})^{L-k} \tag{5.21}$$

Then

$$P(W_k \leq w) = 1 - \int_w^1 \binom{L}{k}(1 - x^{\frac{1}{k}})^{L-k}$$

$$\times \int_w^1 \int_{\frac{w}{x_2}}^1 \int_{\frac{w}{x_2 x_3}}^1 ... \int_{\frac{w}{x_2 ... x_{k-1}}}^1 dx_1 ... dx_k \tag{5.22}$$

$$= 1 - \binom{L}{k}\int_w^1 (1 - x^{\frac{1}{k}})^{L-k}$$

$$\times \left[1 - \frac{w}{x}\sum_{s=0}^{k-2} \frac{\left(\log \frac{x}{w}\right)^s}{s!}\right] dx \tag{5.23}$$

5.7 Appendix 2

A C++ function for the Monte-Carlo computation of the weighted κ-method. User needs to supply a uniform (0,1) random numbers generator, random_u01(). The ln_W0 parameter is $\sum_i^k v_i \ln p_i$ calculated for the original set of p-values. Other parameters are as in the text.

```
#include <set>
#include <vector>
#include <math.h>
double random_u01();

double k_method (double ln_W0, size_t L, size_t k,
                 size_t B, const vector<double>& v) {
    struct Smallest {
        multiset<double, greater<double> > l;
        size_t lsz;
        Smallest(size_t lsz_) : lsz(lsz_) {}
        void Update(double p) {
           l.insert(p); if(l.size()>lsz) l.erase(l.begin());
        }
    };
    size_t R=0, i;
    for(size_t j=0; j<B; j++) {
        Smallest s(k);
        for(i=0; i<L; i++) s.Update(random_u01());
        double ln_WC=0;
        multiset<double, greater<double> >::iterator C=s.l.begin();
        for(i=0; C!=s.l.end(); C++,i++) ln_WC+=v[i]*log(*C);
        if(ln_WC<=ln_W0) ++R;
    }
    return double(R)/B;
}
```

5.8 References

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society series B-methodological, 57: (1) 289-300.

Drigalenko, E.I and Elston, R.C. 1997. False discoveries in genome scanning. Genetic Epidemiology, 14: 779-784.

Edgington, E.S. 1972. An additive method for combining probability values from independent experiments. J. Psychology 80:351–363.

Fisher, R.A. 1932. Statistical Methods for Research Workers. Oliver and Boyd, London.

Good, I.J. 1955. On the weighted combination of significance tests. J. Roy. Stat. Soc. B 17:264-265.

Good, I.J. 1992. The Bayesian/non-Bayesian compromise: a brief review. JASA 87: 597-606.

Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of signifi-

cance. Biometrika 75: (4) 800-802.

Holland, B.S., Copenhaver, M.D. 1987. An improved sequentially rejective Bonferroni test procedure. Biometrics 43 (2): 417-423.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian J of Stat 6: 65-70.

Kepler, T.B., Crosby, L. and Morgan, K. 1999. Normalization and analysis of DNA microarrays by self-consistency and local regression (in preparation).

Marcus, R., Peritz, E., and Gabriel, K.R. 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 77: 63:655–660.

Rice, W.R. 1990. A consensus combined p-value and the family-wide significance of component tests. Biometrics 46 (2): 303-308.

Stouffer, S.A., E.A. Suchman, L.C. DeVinney, S.A. Star and R.M. Williams, Jr. 1949. The American Soldier, Vol. 1. Adjustment During Army Life. Princeton Univ. Press, Princeton.

Weller, J.I., Song, J.Z., Heyen, D.W., Lewin, H.A. and Ron M. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. 1998. Genetics 150: 1699-1706.

Westfall, P.H. and Young, S.S. 1993. Resampling-Based Multiple Testing. Wiley, New York.

Zaykin, D, Young, S.S. and Westfall, P.H. 1999a. Using false discovery rate approach in the genetic dissection of complex traits: a response to Weller et al. (submitted to Genetics).

Zaykin, D, Zhivotovsky Lev A., and Weir, B.S. 1999b. Combining independent p-values (in preparation).

Zaykin D. and Young, S.S. 1999c. P-value adjustments in confirmatory studies (in preparation).

Table 5.1: $E(T)$; $(L = 50,000,\ \beta = 0.80)$

H_A	κ (overall)	τ (overall)	H	BH
15	1.419	1.267	0.227	0.244
25	4.214	4.052	0.294	0.362
50	13.466	10.310	0.641	0.892
100	30.682	34.162	1.074	2.250
200	53.391	70.270	2.324	7.419

Table 5.2: $E\{I(T > 0)\}$; $(L = 50,000,\ \beta = 0.80)$

H_A	κ (overall)	τ (overall)	H	BH
15	0.244	0.227	0.190	0.195
25	0.460	0.457	0.245	0.258
50	0.816	0.794	0.466	0.492
100	0.993	0.980	0.675	0.750
200	1.000	1.000	0.891	0.950

Table 5.3: $E(\text{FDR} \mid T + F > 0)$; $(L = 50,000,\ \beta = 0.80)$

H_A	κ (overall)	τ (overall)	H	BH
15	0.942	0.954	0.214	0.231
25	0.910	0.920	0.174	0.175
50	0.835	0.854	0.059	0.073
100	0.691	0.742	0.041	0.064
200	0.466	0.587	0.007	0.063

Table 5.4: $E(T)$

β	L	H_A	τ (step-down/overall)	H	BH
0.50	100	5	0.483/1.048	0.246	0.300
0.50	100	15	3.529/7.163	0.747	1.214
0.50	1000	5	0.076/0.309	0.063	0.068
0.50	1000	15	0.598/2.610	0.188	0.227
0.50	1000	25	1.837/8.232	0.314	0.422
0.50	1000	50	7.529/24.507	0.609	1.041
0.80	1000	5	0.348/0.886	0.406	0.485
0.80	1000	15	3.310/9.541	1.201	1.940
0.80	1000	25	8.150/19.851	1.989	4.011
0.80	1000	50	22.54/39.937	4.020	11.565

Table 5.5: $E\{I(T > 0)\}$

β	L	H_A	τ (step-down)	H	BH
0.50	100	5	0.293	0.223	0.237
0.50	100	15	0.916	0.534	0.576
0.50	1000	5	0.058	0.063	0.067
0.50	1000	15	0.289	0.172	0.227
0.50	1000	25	0.607	0.314	0.422
0.50	1000	50	0.978	0.461	0.511
0.80	1000	5	0.192	0.347	0.362
0.80	1000	15	0.788	0.712	0.750
0.80	1000	25	0.991	0.876	0.912
0.80	1000	50	0.999	0.996	0.984

Table 5.6: $E(\text{FDR} \mid T + F > 0)$

β	L	H_A	τ (step-down/overall)	H	BH
0.50	100	5	0.437/0.674	0.155	0.173
0.50	100	15	0.187/0.357	0.050	0.068
0.50	1000	5	0.797/0.955	0.444	0.454
0.50	1000	15	0.546/0.873	0.202	0.221
0.50	1000	25	0.420/0.799	0.143	0.158
0.50	1000	50	0.322/0.653	0.075	0.089
0.80	1000	5	0.488/0.932	0.100	0.124
0.80	1000	15	0.168/0.807	0.038	0.065
0.80	1000	25	0.164/0.707	0.023	0.053
0.80	1000	50	0.162/0.540	0.044	0.047

Table 5.7: Confirmatory analysis

	p-values		
Gene	1-st experiment	repetition	overall (adjusted)
V-abl Abelson murine leukemia	$0.46\ 10^{-05}$	0.032	0.005
Human CDK inhibitor p19INK4d	$0.34\ 10^{-04}$	0.044	0.031
Murine leukemia viral (bmi-1)	$0.34\ 10^{-04}$	0.051	0.029
Tetranectin	$0.46\ 10^{-04}$	0.081	0.024
Tyrosin E-protein kinase ITK/T	$0.60\ 10^{-04}$	0.128	0.044
Human cytochrome bc-1 complex	$0.65\ 10^{-04}$	0.075	0.037
Integrin, beta 8	$0.87\ 10^{-04}$	N/A	
Human placenta (Diff33) mRNA	$0.11\ 10^{-03}$	0.022	0.089
Homo sapiens Su(var)3-9 homol	$0.17\ 10^{-03}$	0.054	0.119
Spleen focus forming virus	$0.21\ 10^{-03}$	0.098	0.134
Homo sapiens pyruvate dhg	$0.24\ 10^{-03}$	0.615	0.436
Homo sapiens RaP2	$0.27\ 10^{-03}$	0.220	0.244
Human putative endothelin rec	$0.28\ 10^{-03}$	0.185	0.238
Myosin light chain (alkali)	$0.31\ 10^{-03}$	0.212	0.284
Laminin, gamma 1	$0.40\ 10^{-03}$	0.266	0.359
Human phosphatidylinositol	$0.41\ 10^{-03}$	0.055	0.230
Human prepromultimerin mRNA	$0.44\ 10^{-03}$	0.077	0.236
H.sapiens hPTPA mRNA	$0.44\ 10^{-03}$	0.073	0.264
Human protein kinase (zpk)	$0.44\ 10^{-03}$	0.134	0.285
CD83 antigen precursor	$0.50\ 10^{-03}$	0.017	0.400
Interferon (gamma)	$0.56\ 10^{-03}$	0.486	0.667
Lysyl oxidase	$0.57\ 10^{-03}$	0.240	0.445

Fisher's combined p-value	
$2.84\ 10^{-06}$	$9.71\ 10^{-05}$

Table 5.8: $E(T)$, 3567 tests

β	H_A	τ (step-down)	H	BH
0.40	50	12.376	0.753	1.055
0.40	75	21.517	1.147	1.945
0.40	100	31.009	1.494	2.843
0.50	15	2.111	0.423	0.494
0.50	25	6.183	0.713	0.940
0.50	50	17.501	1.404	2.348
0.50	75	29.258	2.158	4.189
0.80	15	6.381	2.187	3.182
0.80	25	13.782	3.648	6.264
0.80	50	32.839	7.153	15.496

Table 5.9: $E\{I(T > 0)\}$, 3567 tests

β	H_A	τ (step-down)	H	BH
0.40	50	0.999	0.538	0.558
0.40	75	1.000	0.701	0.730
0.40	100	1.000	0.787	0.812
0.50	15	0.777	0.355	0.360
0.50	25	0.988	0.508	0.527
0.50	50	1.000	0.762	0.790
0.50	75	1.000	0.921	0.911
0.80	15	0.999	0.895	0.907
0.80	25	1.000	0.982	0.991
0.80	50	1.000	1.000	1.000

Table 5.10: $E(\text{FDR} \mid T + F > 0)$, 3567 tests

β	H_A	τ (step-down)	H	BH
0.40	50	0.058	0.001	0.001
0.40	75	0.055	0.001	0.001
0.40	100	0.047	0.000	0.001
0.50	15	0.048	0.003	0.004
0.50	25	0.052	0.001	0.001
0.50	50	0.050	0.001	0.001
0.50	75	0.050	0.001	0.001
0.80	15	0.019	0.001	0.002
0.80	25	0.022	0.001	0.001
0.80	50	0.023	0.001	0.001

Chapter 6

P-VALUE ADJUSTMENTS IN CONFIRMATORY STUDIES

6.1 Background

Consider the following situation. In a marker/disease association study a whole genome scan has been performed. One or more smallest p-values have been recorded. Suppose there are other studies that could be used for confirming the results. When there is only one pair of p-values, one resulting from a study, and another from the replication, both can be combined with Fisher's method (Fisher, 1935). Multiple p-values from two studies, however, cannot be simply combined with the Fisher's procedure. An obvious possibility is to combine each pair first, and then adjust by the total number of pairs, L Here we describe a better method that allows to include one or more smallest p-values from the first study into the combination test.

6.2 Derivation of the method

First, let X_1 be the smallest p-value from the first study involving L tests and X_2 be the p-value obtained from the confirmatory study, testing the same null hypothesis. Such situation is relevant, in particular, for the validation of nodes of recursive partitioning trees, where tests are ordered by p-values and the smallest p-value is used for determining a "descriptor" for splitting (e.g. Hawkins, D. M. and Kass, 1982).

Under H_0, $X_1 \sim \text{Beta}(1, L)$ and X_2 has the uniform(0,1) distribution.

Consider a random variable $W = X_1 X_2$. Its density is

$$f_W(w) = \int_{x_1 x_2}^{1} \frac{1}{t}\, L(1-t)^{L-1} dt \tag{6.1}$$

which can be expressed in a semi-closed form. Let

$$\Psi(i, L) = \frac{(-1)^{i+1} \prod_{j=0}^{i} (L-j)}{i^2 (i-1)!}. \tag{6.2}$$

$\Psi(i, L)$ is the result of the repeated indefinite integration by parts of the equation (6.1).

Then the density of W is

$$\begin{aligned} & f_W(w) = & (6.3) \\ & \left(-L \log w + \sum_{i=1}^{L} \Psi(i, L) w^i\right) - \left(-L \log 1 + \sum_{i=1}^{L} \Psi(i, L) 1^i\right) \\ = \;& -L \log w + \sum_{i=1}^{L} \Psi(i, L) w^i - \sum_{i=1}^{L} \Psi(i, L) & (6.4) \\ = \;& \sum_{i=1}^{L} \Big\{ (w^i - 1) \Psi(i, L) - \log w \Big\} & (6.5) \end{aligned}$$

The combined p-value is therefore

$$\begin{aligned} p_c &= 1 - \int_w^1 \sum_{i=1}^{L} \Big\{ (t^i - 1) \Psi(i, L) - \log t \Big\}\, dt & (6.6) \\ &= 1 - \int_w^1 \sum_{i=1}^{L} \left\{ (t^i - 1) \frac{(-1)^{i+1} \prod_{j=0}^{i} (L-j)}{i^2 (i-1)!} - \log t \right\} dt & (6.7) \end{aligned}$$

This distribution function is most easily evaluated through the Monte-Carlo simulation. For an observed value of the product of two p-values, w_0, the empirical p-value is obtained as the proportion of times when $w_i \leq w_0$, where w_i is a product of Beta(1, L) and U(0, 1) random numbers.

The above might be generalized to the case when there are several smallest p-values. For this, we will need the distribution function of the product of two Beta random variables.

Define the random variables $X_1 \sim \text{Beta}(b, g)$ and $X_2 \sim \text{Beta}(k, m)$. After making a transformation $W = X_1 X_2$, $V = X_2$, substituting $x = \frac{1-v}{1-w}$ and integrating, the pdf of $W = X_1 X_2$ is found as

$$f_W(w|k, m, b, g) = \frac{\Gamma(g+b)\Gamma(m+k)}{\Gamma(m)\Gamma(k)\Gamma(g)\Gamma(b)} \times (1-w)^{m+g-1} w^{k-1} \int_0^1 \frac{(1-x)^{m-1} x^{g-1}}{[1-x(1-w)]^{m+k-b}} dx \tag{6.8}$$

Note that the closed form of the integral is not available in general. In a particular case when $m + k - b = 0$, the integral is over support of the non-normalized Beta density. In such situations the product is itself a Beta(g+m, k) random variable.

Define τ being the number of p-values that passed to the second stage. Then $m = \tau - k + 1$, where k is the rank of p-value at the the second stage, $g = L$ and $b = 1$. $f_W(w)$ is therefore

$$f_W(w) = \frac{L\tau!}{(\tau-k)!k!}(1-w)^{\tau+L-k} w^{k-1} \int_0^1 \frac{(1-x)^{\tau-k} x^{L-1}}{[1-x(1-w)]^{\tau}} dx$$

The combined p-value is

$$p_c = 1 - \int_w^1 f_W(t) dt \tag{6.9}$$

It is easy to compute the Monte Carlo equivalent for the CDF of the product. If the product of two Beta random variables is W, and rBeta(x,y) is a Beta(x, y) random numbers generator, then the pseudo-code for obtaining the cumulative probability ("combined p-value") is

```
CNT = 0;
Do N times:
    if [rBeta(b, n)*rBeta(k, m) <= W] then CNT = CNT+1;
done;
PVALUE = CNT / N;
```

A valid and simpler way to correct for multiple tests performed during the first study is to combine all p-values across studies where replicates are available with Fisher's procedure, and then adjust each combined p-value by L. Since we are interested in only one pair of tests, the combined p-value (p_n) for the product, w, is

$$p_n = L\left(1 - \int_w^1 \int_v^1 \frac{1}{t}\, dt\, dv\right) \tag{6.10}$$

$$= Lw(1 - \log w) \tag{6.11}$$

It can be shown that p-value calculated with (6.11) is always greater than or equal to the probability obtained with (6.7). In fact, the difference between the two probabilities, $\delta = p_n - p_c$, is bounded by L: $0 \leq \delta \leq L$. The following table shows the difference as a function of w for for the first four values, $L = 1, ..., 4$

L	$\delta = p_n - p_c$
1	0
2	$2w - w^2$
3	$\frac{9w-6w^2+w^3}{2}$
4	$\frac{22w-18w^2+6w^3-w^4}{3}$

The difference between probabilities increases with w and L. suggesting that p_c is a correct although conservative probability for any "significant" p-value, and not necessarily the smallest.

Consider an improved version of p_n, based on Šidák's correction:

$$p_{\check{\mathrm{S}}} = 1 - (1 - w(1 - \log w))^L \tag{6.12}$$

First four differences are again listed below. As before, they are always positive, but with the maximum at intermediate values of w.

L	$\delta = p_n - p_{\check{\mathrm{S}}}$
1	0
2	$w\,(2 - 2w - w\,(\log w - 2)\log w)$
3	$1 + \frac{w(3+(w-6)w+6\log w)}{2} - (1 - w + w\log w)^3$
4	$1 - \frac{w(w(18+(w-6)w)-12\log w-10)}{3} - (1 - w + w\log w)^4$

6.3 Results and discussion

p-values under H_A have been generated for the case of testing the difference between two normal means as described in Zaykin et al. (1999), with the individual power of each test being equal to 80%. If u is a random number from $U(0, 1)$, then

an observation from the distribution under H_A can be obtained as

$$p = 1 - \Phi\left(\Phi^{-1}(1-u) + \gamma\right) \tag{6.13}$$

with

$$\gamma = \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma} \tag{6.14}$$

Therefore, splitting the sample size in half (with other conditions held equal) is equivalent to dividing γ by $\sqrt{2}$. This allows to generate p-values as if all "pooled" data from both experiments were available.

We recorded standard type-1 error rate (realized α-level when all tests are performed under H_0), statistical power, and the expected number of rejections. This has been done both under complete H_0 and under "partial H_A", when some of the tests have been generated under H_A and the remaining tests have been generated under H_0. The α-level was 5% and the number of simulations was 10,000.

Note that under H_0, the expected number of rejections might not necessarily be equal to the α-level. As an example, consider the case of Bonferroni ($\alpha_1 = \alpha/L$) and Šidák ($\alpha_2 = 1 - (1-\alpha)^{1/L}$) corrections. With L tests, the expected number of rejections with Bonferroni correction is equal to the α level: $L\alpha_1 = \alpha$. Since $\alpha_2 > \alpha_1$ when $L > 1$, it follows that the expected number of rejections for the Šidák method is always greater than the α-level. With an exaggerated example of $\alpha = 0.5$ and $L = 100$, the expected number of rejections for the Šidák method becomes 0.69 (almost 20% increase).

As it has been shown above, when only the experiment with the smallest p-value is replicated, combining two studies and multiplying the result by L is not an optimal procedure. At the another extreme, when all experiments are replicated, (6.11) or (6.12) adjustments are equivalent to dividing the whole sample from both replicates into two parts, performing tests on both and then combining resulting p-values. Bauer and Kohne (1994) suggested that the power of this procedure is almost equivalent to that of the test performed on the pooled sample. That is, the loss of power following artificial partitioning of the total sample is quite small. Our simulations confirm this observation (data not shown). Therefore, we expect that p-value adjustments calculated with (6.9) should be optimal when only part of experiments is replicated. We considered situations when the most significant result is replicated, or all experiments with p-values smaller than α are replicated.

We confirmed that both methods maintain correct type-I error rate (tables 6.1, 6.2). In particular, table 6.2 shows that the method has the right rejection level even when *all*, and not only small tests have been replicated.

For non-H_0 situation (tables 6.3, 6.4), the results indicate that when the proportion of tests from H_A is 1/50 or higher, or when only the experiment with the smallest p-value is replicated, p_c method provides higher expected number of rejected non-true null hypotheses.

6.4 References

Fisher, R. 1932. Statistical methods for research workers. London: Oliver and Boyd.

Bauer, P. and K. Kohne. 1994. Evaluation of experiments with adaptive interim analysis. Biometrics 50:1029-1041.

Hawkins, D.M. and Kass, G.V. 1982. Automatic Interaction Detection. Topics in Applied Multivariate Analysis, 269–302 (D. M. Hawkins, editor). Cambridge University Press. Cambridge.

Zaykin D., L. Zhivotovsky and B.S. Weir. 1999. Combining independent p-values (in progress).

Table 6.1: Power and expected number of rejections under H_0

Test	Replication condition	L	E(rej)	Type-I error
p_n	Smallest p	100	0.0284	0.0284
p_c	Smallest p	100	0.0522	0.0522
p_n	$p \leq \alpha$	100	0.0367	0.0362
p_c	$p \leq \alpha$	100	0.0445	0.0433
p_n	Smallest p	1000	0.0248	0.0248
p_c	Smallest p	1000	0.0476	0.0476
p_n	$p \leq \alpha$	1000	0.0385	0.0371
p_c	$p \leq \alpha$	1000	0.0472	0.0476

Table 6.2: Power and expected number of rejections under H_0 when all p-values from the first stage are included

Test	L	E(rej)	Type-I error
p_n	10	0.0510	0.0495
p_c	10	0.0525	0.0502
p_n	100	0.0524	0.0514
p_c	100	0.0538	0.0530
p_n	1000	0.0523	0.0508
p_c	1000	0.0528	0.0515

Table 6.3: Power and expected number of correct rejections under partial H_A with 1/10 tests from H_A

Test	Replication condition	L	E(rej)	power
p_n	Smallest p	100	0.9552	0.9552
p_c	Smallest p	100	0.9686	0.9686
p_n	$p \leq \alpha$	100	5.1913	0.9993
p_c	$p \leq \alpha$	100	6.2254	0.9995
p_n	Smallest p	1000	0.9970	0.9970
p_c	Smallest p	1000	0.9978	0.9978
p_n	$p \leq \alpha$	1000	31.1573	1
p_c	$p \leq \alpha$	1000	40.8573	1
p_n	Smallest p	10000	0.9997	0.9997
p_c	Smallest p	10000	0.9997	0.9997
p_n	$p \leq \alpha$	10000	158.9855	1
p_c	$p \leq \alpha$	10000	206.8952	1

Table 6.4: Power and expected number of correct rejections under partial H_A with 10 tests from H_A

Test	Replication condition	L	E(rej)	power
p_n	Smallest p	100	0.9521	0.9521
p_c	Smallest p	100	0.9671	0.9671
p_n	$p \leq \alpha$	100	5.1913	0.9993
p_c	$p \leq \alpha$	100	6.2254	0.9995
p_n	Smallest p	500	0.8391	0.8391
p_c	Smallest p	500	0.8847	0.8847
p_n	$p \leq \alpha$	500	3.6973	0.9892
p_c	$p \leq \alpha$	500	3.7217	0.9787

Chapter 7

BAYESIAN TESTS FOR HETEROZYGOTE EXCESS AND DEFICIENCY

7.1 Abstract

We propose a multiallelic Bayesian test for the excess or the deficit of heterozygotes (HDE) and compare its performance with exact frequentist procedures. Based on the coalescent process simulations of samples from admixed populations we found that our Bayesian test has desirable frequentist properties. It compares well with exact tests proposed previously and provides additional benefits of Bayesian interpretations and entire plots of posterior distribution of coefficients measuring HDE. The method is computationally much faster in comparison with the frequentist exact tests.

7.2 Preliminaries

Shoemaker et al (1998) discussed a Bayesian approach for characterization of the HWD for the case of two alleles. With two alleles, there is a single disequilibrium parameter and there is no distinction between the testing for Hardy-Weinberg disequilibrium (HWD) or the heterozygote deficiency/excess (HDE). Extension to multiple alleles is most straightforward for the testing of the overall HDE. Such extension allowed us to construct a test with good frequentist properties. Although we find the Bayesian approach appealing by itself, in this article we concentrate on how our test behaves as if it were a classical test in comparison with the previously suggested methods.

Rousset and Raymond (1995) provided a thorough comparison of several classical exact procedures for the HDE. The tests are termed "exact" for their conditioning on the observed allele numbers, which allows to eliminate the unknown population allele frequencies (nuisance parameters) from the test statistic. The Bayesian approach deals with the nuisance parameters by integrating them out; conditioning on all the data, not just the marginal counts. It is often the case that the Bayesian approach results in the point and the interval procedures with excellent frequentist properties in both simple and complicated modeling scenarios (Carlin and Louis, 1996). Rousset and Raymond (1995) found that the exact tests based either on the number of heterozygotes in permuted samples (N_{het}) or their score test ($U = \sum n_{ii}/\hat{p}_i$, where $\hat{p}_i$'s are the observed allele frequencies) have the optimal performance in most situations. We used U, N_{het}, and some other tests defined below as the basis for comparison with the Bayesian tests.

We require that the HDE measure is zero when there is no excess/deficit of heterozygotes and that it is negative or positive otherwise. Suppose we calculate the following posterior probability

$$\tau = \Pr(\text{heterozygote excess} > 0) \tag{7.1}$$

Then

$$p = 2\tau \; I(\tau < \frac{1}{2}) + 2(1-\tau) \; I(\tau \geq \frac{1}{2}) \tag{7.2}$$

(where $I()$ is the indicator function) can serve as a Bayesian surrogate for "two-tailed" p-value, provided that τ has the Uniform(0,1) distribution under H_0 (Appendix A). In practice, τ in (7.1) is estimated by sampling N times from the posterior distribution and counting the number of times (y) when the heterozygote excess was greater than zero in the posterior samples. Note that y is the binomial random variable, Bin(N, τ). The prior on τ might be given by the Uniform(0,1) distribution. Then the posterior Bayesian p-value is calculated from (7.2) using

$$\hat{\tau} = \frac{y+1}{N+2} \tag{7.3}$$

which is the posterior expectation for τ. This has the advantage over reporting the sample proportion, $\hat{\tau}_{MLE} = y/N$ in that the value of τ is never exactly zero.

7.3 Method

We consider the Dirichlet-multinomial model for our analysis. The multinomial likelihood of the data is

$$Pr(\{n_{ij}\} \; | \{P_{ij}\}) = \frac{n!}{\prod n_{ij}!} \prod P_{ij}^{n_{ij}} \tag{7.4}$$

where n_{ij} is the observed genotype counts of the type ij, $n = \sum n_{ij}$ and P_{ij} are the population genotype frequencies.

The prior distribution for the genotype frequencies is the Dirichlet$(\gamma_{11}, \gamma_{12}, ..., \gamma_{kk})$, where k is the number of alleles:

$$\pi(\mathbf{P}) = \frac{\gamma_.}{\prod_{i \leq j} \gamma_{ij}} \prod_{i \leq j} P_{ij}^{\gamma_{ij}-1} \tag{7.5}$$

Then the posterior distribution is the Dirichlet$(n_{11} + \gamma_{11}, n_{12} + \gamma_{12}, ..., n_{kk} + \gamma_{kk})$.

We define folowing measures of HDE for the Bayesian tests.

The total disequilibrium coefficient is defined as

$$D = \sum_{i<j} D_{ij} = 2 \sum_{i} D_{ii} \tag{7.6}$$

which could be scaled by its maximum value

$$\Psi = \sum \Psi_{ij}, \tag{7.7}$$

$$\Psi_{ij} = \begin{cases} p_i p_j, & D_{ij} \geq 0 \text{ and } i \neq j \\ -\min\left[p_i(p_j - 1),\ p_j(p_i - 1)\right], & D_{ij} < 0 \text{ and } i \neq j \\ p_i(1 - p_i), & D_{ij} \geq 0 \text{ and } i = j \\ -p_i^2, & D_{ij} < 0 \text{ and } i = j \end{cases} \tag{7.8}$$

$$D_{ij} = p_i p_j - \frac{1}{2} P_{ij} \tag{7.9}$$

$$D_{ii} = P_{ij} - p_i^2 \tag{7.10}$$

where p_i is a frequency of the i-th allele. The scaled weighted disequilibrium coefficient is

$$D_s = \frac{1}{\Theta} \sum_{i<j} \frac{D_{ij}}{\sqrt{p_i p_j}} \tag{7.11}$$

where

$$\Theta = \sum_{i \neq j} \Theta_{ij}, \tag{7.12}$$

$$\Theta_{ij} = \begin{cases} \sqrt{p_i p_j}, & D_{ij} \geq 0 \\ -\min\left[\sqrt{\frac{p_i}{p_j}}(p_j - 1),\ \sqrt{\frac{p_j}{p_i}}(p_i - 1)\right], & D_{ij} < 0 \end{cases} \tag{7.13}$$

The coefficients Ψ_{ij} and Θ_{ij} are the bounds for the corresponding disequilibrium coefficients and are derived from the constraints imposed by allele frequencies on the genotype frequencies (Weir, 1996).

The third measure ($\delta = -1 + \sqrt{P_{AA}} + \sqrt{P_{aa}}$) was discussed in Ghosh and Weir (in preparation) as a measure of the departure from HWE in the case of two alleles. In the multiallelic case δ is a measure of homozygote excess or deficiency (Appendix D) and is defined as

$$\delta = -1 + \sum \sqrt{P_{ii}} \tag{7.14}$$

δ varies between -1 and 1 is somewhat analogous to the U statistic.

In a single-parameter case (the beta-binomial model) the parameters of the prior distribution (γ_{ij}-s here) are not likely to affect the posterior conclusions, because the change in the prior parameters from 0 to 1 amounts to adding at most a single observation to the data (Gelman et al 1995). This is not the case with our multivariate model. We have found that the uniform prior on the genotype frequencies grossly overestimates the probability that the null hypothesis is false (Fig 1). This happens because the quantities of interest (measures of HDE) are the transformations of the original parameters. The square root of the determinant of the Fisher information matrix was suggested by Jeffreys (1961) as the prior that

is invariant to the form of the parameterization. The Jeffreys prior distribution (symmetric Dirichlet with all parameters equal to 1/2) is derived in Appendix B. We also considered another symmetric prior, Dir(1/3). Appendix C gives the contingency tables testing argument in favor of this prior.

Our simulations show that when the H_0 is true, the Jeffreys prior generates the uniform (0,1) distribution of p-values for δ. We obtained better results for D using the Dir(1/3) prior.

7.4 Results and discussion

We studied and compared frequentist performance of coefficients D_s and δ with measures U, F, and H, considered in Rousset and Raymond (1995). The measures are defined as follows. Since significance were determined by shuffling, only variable parts of statistics are considered.

U-statistic is defined as

$$U \quad \propto \quad \sum_{i=1} \frac{n_{ii}}{\hat{p}_i} \tag{7.15}$$

"Exact test"

$$F \quad \propto \quad N_h \ln 2 - \sum_{i \leq j} \ln (n_{ij}!) \tag{7.16}$$

Observed number of heterozygotes

$$H \quad \propto \quad N_h \tag{7.17}$$

We also studied a "frequentist" version of the δ coefficient, defined as

"frequentist" δ

$$\delta_f \quad \propto \quad \sum_{i=1} \sqrt{n_{ii}} \tag{7.18}$$

Samples of 100 genotypes scored at four loci were generated by the coalescent process (Hudson, 1990) with the consequent admixture. Ten populations were allowed to drift until the prescribed amount of divergence, as measured by the coancestry coefficient (Weir, 1996). Values of θ varied from 0.0 to 0.3, tables 7.1, 7.2.

Simulations under H_0 shows that proportions of rejections, determined in 10,000 simulations are satisfactory for all tests, indicating that the posterior probabilities can indeed be treated as frequentist p-values, when the Dir(1/2) and Dir(1/3) priors were used.

Under H_A, Bayesian tests show good power, which is increasing with the number of alleles. It should be noted that while we restricted attention to the prior parameters, fixed for each coefficient, we were able to find that the method based on the coefficient D_s can always outperform frequentist tests if the prior parameters are set to values higher than 1/2 for smaller number of alleles, while still keeping the type-I error on the declared level. In addition, samples from entire posterior distribution of D_s and δ are most easily produced, providing means for the Bayesian interval inference.

A future extension of this work will include investigation of non-informative classes of conjugate priors with less restrictive covariance structure than that of the Dirichlet distribution (e.g. Rayens and Srinivasan, 1994, also Wong, 1998).

7.5 Appendix A

Let τ be a uniform random variable, $\tau \sim \mathrm{U}(0,1)$

$$p = 2\tau\ I(\tau \leq \frac{1}{2}) + 2(1-\tau)\ I(\tau > \frac{1}{2}) \tag{7.19}$$

The moment generating function of p is

$$\begin{aligned} m_p(t) &= E\{\exp(tp)\} \\ &= \int_0^1 \exp(2\tau I(\tau < \frac{1}{2})t) \exp(2(1-\tau)I(\tau \geq \frac{1}{2})t) d\tau \\ &= \int_0^{1/2} \exp(2\tau t) d\tau + \int_{1/2}^1 \exp(2(1-\tau)t) d\tau \\ &= \frac{e^{1t} - e^{0t}}{(1-0)t} \end{aligned} \tag{7.20}$$

so that p in (7.19) has a Uniform$(0,1)$ distribution.

7.6 Appendix B

Finding the Jeffreys prior

The Jeffreys prior is defined as the square root of the determinant of the Fisher information matrix: $\pi(\mathbf{P}) = |I(\mathbf{P})|^{1/2}$. Let P_i denote $m = k(k+1)/2$ multinomial genotypes. The log-likelihood is

$$\ln f(\mathbf{n}|\mathbf{P}) \propto n_m \ln(1 - \sum^m P_i) + \sum^m n_i \ln P_i \tag{7.21}$$

The elements of the Fisher information matrix are

$$-E\left(\frac{\partial^2}{\partial P_i \partial P_j}\ln f(\mathbf{n}|\mathbf{P})\right) \tag{7.22}$$

$$= \begin{cases} \frac{n}{P_i} + \frac{n}{1-\sum_{v\neq m} P_v}, & i = j \\ \\ \frac{n}{1-\sum_{v\neq m} P_v}, & i \neq j \end{cases} \tag{7.23}$$

where $n = \sum n_i$.

This matrix is of the form

$$\begin{pmatrix} a_1 + x & x & x & ... \\ x & a_2 + x & x & ... \\ x & x & a_3 + x & ... \\ ... & ... & ... & ... \end{pmatrix} \tag{7.24}$$

with the determinant

$$\det = \prod a_i + \sum_i^m \prod_{j\neq i} a_i x \tag{7.25}$$

so that

$$\frac{1}{n}|I(\mathbf{P})| = \prod \frac{1}{P_i} + \sum_i^m \prod_{j\neq i} \frac{1}{P_i\left(1-\sum_{v\neq m} P_v\right)} \tag{7.26}$$

$$= \frac{1}{\prod P_i\left(1-\sum P_j\right)} \tag{7.27}$$

and

$$|I(\mathbf{P})|^{1/2} \propto \prod P_i^{-\frac{1}{2}}\left(1-\sum P_j\right)^{-\frac{1}{2}} \tag{7.28}$$

Therefore, the Jeffreys prior is the conjugate Dir(1/2) distribution.

7.7 Appendix C

We have found that under HWE the median of the posterior distribution of D_s is at zero with the symmetric Dir(1/3) prior. A similar result holds for a simpler case of testing the heterogeneity between rows in a 2×2 contingency table. Let the table be

$$\begin{bmatrix} n_1 & n_2 \\ n_3 & n_4 \end{bmatrix} \tag{7.29}$$

Let π_1 be the population frequency of the first binomial sample (first row) and π_2 be the population frequency of the second binomial sample. The exact Bayesian test calculates the posterior probability that $\pi_1 < \pi_2$ as

$$\begin{aligned} P(\pi_1 < \pi_2) = \int_0^1 \int_t^1 & B(n_1+\gamma, n_1+\gamma) s^{n_1+\gamma-1}(1-s)^{n_2+\gamma-1} \\ & \times B(n_3+\gamma, n_4+\gamma) t^{n_3+\gamma-1}(1-t)^{n_4+\gamma-1} dtds \end{aligned} \tag{7.30}$$

where γ is the common prior parameter of the Beta distribution. Now, suppose that entries in the second row are precise multiples of the first,

$$\begin{bmatrix} a & b \\ xa & xb \end{bmatrix} \tag{7.31}$$

For the test to be valid in the classical sense, we require that the prior parameter for the beta distribution satisfies

$$\frac{1}{2} = P(\pi_1 < \pi_2) = \int_0^1 \int_t^1 B(a+\gamma, b+\gamma) s^{a+\gamma-1}(1-s)^{b+\gamma-1} \times B(xa+\gamma, xb+\gamma) t^{xa+\gamma-1}(1-t)^{xb+\gamma-1} dt ds \tag{7.32}$$

This can be solved numerically for γ showing that γ needs to be 1/3.

7.8 Appendix D

The δ coefficient with multiple alleles

Consider a genetic locus with k alleles. Let δ be defined as

$$\delta = -1 + \sum_{i=1}^{k} \sqrt{P_{ii}} \tag{7.33}$$

When $\delta = 0$ and HWE holds,

$$1 = \sum_{i=1}^{k} \sqrt{P_{ii}} \tag{7.34}$$

$$1 = \left(\sum_{i=1}^{k} \sqrt{P_{ii}} \right)^2 \tag{7.35}$$

$$1 = \sum_{i=1}^{k} p_i^2 + 2 \sum_{i<j}^{k} p_i p_j \tag{7.36}$$

When HWE does not hold,

$$\left(\sum_{i=1}^{k} \sqrt{P_{ii}} \right)^2 \tag{7.37}$$

$$= \left(\sum_{i=1}^{k} \sqrt{p_i^2 + D_{ii}}\right)^2 \tag{7.38}$$

$$= \sum_{i=1}^{k} p_i^2 + \sum_{i=1}^{k} D_{ii} + 2\sum_{i<j}^{k} \sqrt{(p_i^2 + D_{ii})(p_j^2 + D_{jj})} \tag{7.39}$$

Which can also be written in a form involving disequilibrium coefficients for heterozygotes:

$$\sum_{i=1}^{k} p_i^2 + 2\sum_{i<j}^{k} D_{ij} + 2\sum_{i<j}^{k} \sqrt{(p_i^2 + \sum_{u\neq i}^{k} D_{iu})(p_j^2 + \sum_{v\neq j}^{k} D_{jv})} \tag{7.40}$$

The last equation illustrates that δ measures the excess or the deficit of heterozygosity, rather than the departure from HWE. In other words, δ can be zero when HWE does not hold (this applies when $k > 2$).

The δ coefficient is an appealing measure for the Bayesian posterior inference, because it is directly calculated from posterior samples of genotypes and its range [-1 to 1] is independent of k.

7.9 References

Carlin B. and Louis T. A. 1996. Bayes and empirical Bayes methods for data analysis. Chapman & Hall.

Gelman A., Carlin J. B., Stern H. S., Rubin D. B. Bayesian data analysis. 1995. Chapman & Hall.

Ghosh S. and Weir B. S. Detecting departures from HWE (in prep)

Jeffreys H. 1961. Theory of probability. London, Oxford University Press.

Rayens W and Srinivasan C. 1994. Dependence properties of generalized Liouville distributions on the simplex. JASA 89: 1465–1470,

Rousset F. and Raymond M. 1995. Testing heterozygote excess and deficiency. Genetics **140:** 1413-1419.

Shoemaker J., Painter I. and Weir B. S. 1998. A Bayesian characterization of Hardy-Weinberg disequilibrium. Genetics **149:** 2079-2088.

Weir, B. 1996. Genetic Data Analysis II. Sunderland, Mass.: Sinauer.

Wong T. 1998. Generalized Dirichlet distribution in Bayesian analysis. Applied mathematics and computation. 97: 165–181

Table 7.1: Proportions of rejections under H_0 with declared 5% α-level

	Type-I error, H_0					
# alleles	D_s	δ	F	H	U	δ_f
2	0.053	0.052	0.042	0.042	0.051	0.056
3	0.056	0.052	0.048	0.044	0.050	0.051
4	0.056	0.057	0.047	0.048	0.051	0.050
5	0.058	0.067	0.048	0.050	0.052	0.053
6	0.051	0.071	0.053	0.045	0.048	0.048
10	0.039	0.063	0.051	0.048	0.049	0.048
15	0.035	0.010	0.050	0.045	0.046	0.037
20	0.047	0.010	0.043	0.034	0.043	0.034

Table 7.2: Proportions of rejections under H_A on the level of 5%

	Power, $H_A(\theta = 0.3 - 0.1)$					
# alleles	D_s	δ	F	H	U	δ_f
2	0.719	0.722	0.714	0.692	0.741	0.785
3	0.801	0.786	0.763	0.804	0.812	0.811
4	0.787	0.742	0.701	0.790	0.795	0.789
5	0.672	0.586	0.539	0.681	0.681	0.667
6	0.456	0.329	0.311	0.460	0.465	0.443
7	0.519	0.362	0.357	0.521	0.517	0.483
10	0.663	0.413	0.484	0.642	0.622	0.581
15	0.811	0.508	0.673	0.770	0.731	0.724
20	0.892	0.565	0.792	0.832	0.780	0.808

Chapter 8

SUMMARY

We investigated frequentist and Bayesian statistical methods for studying genetic associations. Exact conditional methods for testing associations between genes, Bayesian tests for the heterozygote excess and deficiency based on tail probabilities of the posterior distribution, and computationally efficient algorithms have been described in detail.

Great improvement in computational speed has been achieved for the exact conditional tests through storing permuted multilocus genotypes in a tree-like structure. We showed that traditional test statistics, such as chi-square, can also be used with this algorithm, since they can be expressed as a sum extending across non-zero entries of contingency tables or sets of multilocus genotypes. We studied these exact methods in the closing procedure framework, with regard to testing for Hardy-Weinberg equilibrium. The procedure provides a global test for association over all loci as well as individual p-values adjusted for multiple testing, accounting for discrete nature of the data.

The procedure has high power when all loci are subject to similar evolutionary forces causing deviations from the Hardy-Weinberg equilibrium. Individual adjustments with the exact test are feasible for moderate numbers of loci (k), with $2^k - 1$ computations, but very good approximations have been found that reduce computational load from exponential to quadratic dependence on the number of loci.

Special attention has been paid to situations with large numbers of tests and to specific issues of combining and adjusting p-values resulting from them. The questions of combining and adjusting p-values address different statistical hypotheses, but they are related through the closure principle.

With many tests, false positives are expected, and benefits and disadvantages of global methods with the weak FWER protection, methods that strongly control FWER, and methods for controlling the false discovery rate have been discussed and studied.

We proposed new methods for combining sets of independent p-values within and across studies and described extensions that allow for correlations between p-values. These procedures are based on the distributions of truncated products and products of first order statistics from the uniform and beta distributions. Depending on the value of the truncation point or on the number of p-values included into the product, the null hypotheses of τ-method and κ-method reduce to those of methods for individual adjustments or of Fisher's combination test. Closed form expressions are not always computationally efficient or available, but Monte Carlo algorithms are easily implemented. These algorithms readily extend for the cases of weighted and correlated observations. We found that these methods have good potential in exploratory analysis. Using expression array data, we demonstrated that it is possible to make inferences about average numbers of true effects in the class of rejected hypotheses. We showed that it is possible to infer the plausible number of false null hypotheses from comparisons of numbers of rejected hypotheses made by individual adjustments methods and by combination methods.