Picking the most likely candidates for further development: Novel intersection-union tests for addressing multi-component hypotheses in comparative genomics

Kyoungmi Kim[1], Stanislav O. Zakharkin[1], Ann E. Loraine[1,2], and David B. Allison[1,2]

[1]Department of Biostatistics, Section on Statistical Genetics, and [2]Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294

Abstract

In the age of modern high dimensional biology, when choosing potential genomic targets and physiological pathways for future study or drug development, investigators need to select a modest number of candidates from among very large numbers of options. In prioritizing these options, they may seek potential targets that meet *all* of several criteria. For example, they might wish to choose genes that are differentially expressed in response to a particular stimulus in each of several model organism species (i.e. evolutionarily conserved responses), or genes that are both differentially expressed in response to a particular stimulus and which produce a predicted phenotypic response when knocked down in an RNAi experiment. Both examples involve testing whether two or more null hypotheses can both be rejected; this entails the conduct of *intersection-union tests* (IUTs). The most common traditional IUT rejects the union of all of k null hypotheses in favor of the intersection of all k alternative hypotheses if a legitimate test for each and every one of the separate k null hypotheses is rejected at level α. This IUT is conservative in all but several unrealistic situations. Moreover, it yields results classifiable as significant or not, but not a single quantitative p-value. Herein, we examine an approach to frequentist testing to overcome these limitations. We then present a Bayesian approach to the problem that makes more complete and intuitive use of the data when many IUTs are being conducted as in high dimensional biology.

Keywords: Comparative genomics, Conserved genes, multi-component hypotheses, Intersection hypothesis**,** Intersection-Union Tests, Bayesian statistics, Mixture models.

INTRODUCTION

Comparing patterns of gene expression across species or tissue types can provide important insights into the molecular or developmental mechanisms underlying semi-universal processes, such as aging, energy metabolism, or certain types of disease states. Homologous genes that exhibit conserved expression patterns in different species or individual genes that are expressed in similar ways across different tissue types within the same species may represent highly conserved pathways or processes; knowledge of these could provide important biological insights into how these pathways operate in healthy as well as diseased tissue.

Effective use of cross-stimulus and cross-species microarray expression studies requires bioinformatic and statistical methodologies that allow for identification of conserved genes across different tissues or organisms. How are the most likely candidates for conserved genes to be identified? To identify those genes, one needs to establish two conditions: (1) that genes are homologous in some biological sense, i.e., that they share some sequence homology or are known through prior studies to participate in the same biological process, and (2) they are differentially expressed in the same way in response to some stimulus of interest. The first condition can be met through relatively well-

established bioinformatic techniques, such as through sequence comparison algorithms or mining databases of biological information. The second condition, however, is not as easily met; we feel that it represents an important, unsolved problem requiring new or reformulated statistical approaches.

For the following discussion, we assume that biologically-informed relations (homologies) between genes are provided and instead focus on the expression component of conservation analysis, which involves finding out whether or not biologically-related genes exhibit equivalent responses to stimuli A, stimuli B, and so on. This latter type of conservation could thus be considered to be a form of *functional* conservation; that is, the genes behave in a similar fashion under similar conditions.

A naïve, or first pass, attempt to determining whether or not two genes behave in a similar fashion, e.g., are differentially expressed in response to the same condition, would be to use a t-test to test each contrast at an overall type I rate of α and require both to be significant. However, in comparative genomics studies, one wishes to test different stimuli simultaneously while retaining the overall significance level. *Intersection-union tests* (IUTs) offer a way to accomplish this goal of simultaneous testing. In addition, they are particularly well-suited for comparative genomic settings[*] involving cross-species microarray studies, which are the focus of this paper.

THE METHODOLOGICAL ISSUE

In order to test two or more hypotheses simultaneously, it is necessary to construct a multi-component[†] hypothesis. Situations in which we would want to test multi-component intersection hypotheses include those where we wish, for example, to (1) find genes that are differentially expressed in response to caloric restriction in all of several selected species (e.g. nematode, fruitfly, and mouse), (2) find quantitative trait loci (QTL) linked to a single trait, for example adiposity, in all of several mouse crosses suggesting that the loci identified are likely to have common variants that are influential and sustainable, (3) find genes that both have their expression levels correlated with body fat (showing physiological relevance) and that do (or do not) change their expression levels when knocked down with RNAi (showing causation); i.e., the intersection of what *does* and what *can* happen, or (4) find genes that independently influence two or more traits (i.e., mosaic pleiotropy) (Harman *et al.* 2000; Kuhel *et al*. 2002; Wang & Paigen 2002).

UNION-INTERSECTION TESTS VS. INTERSECTION-UNION TESTS

A multi-component null hypothesis test includes two or more component null hypotheses, e.g., H_{01}: $\theta_1 = 0$ and H_{02}: $\theta_2 = 0$. Here we only consider two-component hypotheses without loss of generality. However, the rationale can be extended to any number of stimuli. *Union-intersection tests* (UITs) (Roy 1953) apply when the compound null hypothesis is the intersection of all component null hypotheses. The compound (multi-component) hypothesis is rejected if any one of the individual hypotheses is rejected at the multiplicity-adjusted threshold that controls overall experiment type I error rate, as illustrated in Coffman *et al*. 2003. The rejection region for this UIT is the union of rejection regions corresponding to the individual tests (see Figure 1(a)).

Alternatively, *intersection-union tests* (IUTs) (Berger 1982; Berger and Hsu 1996) apply when the compound null hypothesis is the union of two or more hypotheses. In this case, the compound null hypothesis is rejected only when all the component hypotheses are rejected.

Corresopnding author: David B. Allison, Ph. D, Tel: 205-975-9169, Email: dallison@uab.edu.

[*] In this article, we extend the traditional definition of 'comparative genomics' from the comparison of some genomic findings or information across two or more species to the comparison of some genomic findings or information across two or more species, tissues, experiments, or situations. Although methodologies herein are illustrated in microarray studies, the rationale can be extended to allow for tests in identification of conserved quantitative trait loci (QTL) across different stimuli in quantitative genetics and may other situations.

[†] In this paper, we use the terms "compound hypothesis" and "multi-component hypothesis" interchangeably.

An example of the use of *intersection-union tests* in linkage analysis was illustrated by Zhang *et al*. (1998). Tests of one QTL versus two QTL were performed by the Least-squares analysis. The test statistics were computed from the F distribution obtained by data permutation. Under the two QTL model, the regression coefficient for the first QTL was computed while setting the regression coefficient for the second QTL to zero and vice versa. Both resulting F statistics were used in an IUT. If both test statistics were significant, the null hypothesis of the one QTL model was rejected and it was concluded that there was significant evidence of the presence of more than one QTL.

For our purposes, consider an example in which we wish to find genes that are differentially expressed in response to caloric restriction in two organisms, *A* and *B*. To find these, we must consider the union of the null hypotheses that individual genes in A and their counterparts in B are not differentially expressed. Then, if either null hypothesis is "left standing" after the test (i.e., no significant difference observed in A, *or* no significant difference observed in B, or both), then we cannot reject the compound hypothesis.

Formally, we can express this as follows:

Define $\theta_{ik} = |\mu_{1ik} - \mu_{2ik}|$, the absolute mean difference between the expression levels of gene i of the caloric restriction (CR) group and of the placebo group in organism k (e.g., k =A or B). Consider testing the compound hypothesis:

$$H_0: \{ H_{01} \cup H_{02} \}$$

as the **union** of

$$H_{01}: \{ \theta_{iA} \leq \Delta_A \} \ \&$$

$$H_{02}: \{ \theta_{iB} \leq \Delta_B \}$$

versus

$$H_1: \{ H_{11} \cap H_{12} \}$$

as the **intersection** of

$$H_{11}: \{ \theta_{iA} > \Delta_A \} \ \&$$

$$H_{12}: \{ \theta_{iB} > \Delta_B \},$$

where Δ_A and Δ_B are the cut-off values of significance. The i-th gene is acceptable as a conserved gene that is differentially expressed in response to caloric restriction in both organisms *A* and *B* if the global null H_0 is rejected. In other words, H_1 is true if and only if both H_{11} and H_{12} are true. Hence, individual tests for each parameter can be combined by means of the IUTs to yield an overall test of the conserved gene across different organisms.

The IUT rejects the compound hypothesis only if both individual hypotheses, H_{01} and H_{02}, are rejected at level α. In this case, a pre-specified type I error rate is maintained without multiplicity adjustment for multiple components. The rejection region for this test is the intersection of the rejection regions corresponding to the individual tests, that is, $\bigcap_{i=1}^{2} R_i = \{\min(T_A(x), T_B(x)) \geq c_\alpha \}$, where for the *i*-th null hypothesis, R_i is the rejection region, $T_i(x)$ is an appropriate test statistic, and c_α is the threshold value associated with type I error rate of α (see Figure 1(b)). The p-value for the minimum of test statistics (so-called 'min' test), which is the largest p-value, is only used to determine whether the compound null hypothesis will be rejected, regardless of the values of any other p-values. Therefore, the maximum p-value itself does not offer quantitative interpretation as a "weight of evidence".

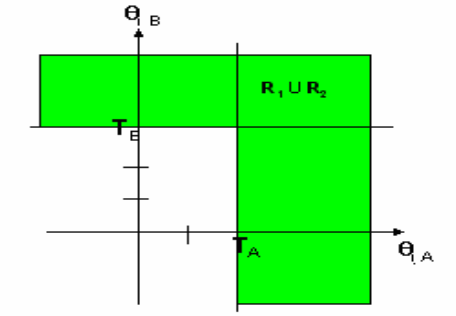
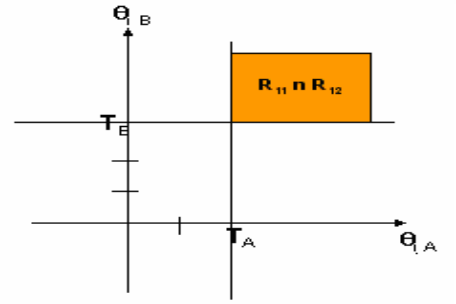


Figure1. (a) Region (Green) of parameter space corresponding to H1 in UITs



(b) Region (orange) of parameter space corresponding to H_1 in IUTs

AN ALTERNATIVE FREQUENTIST APPROACH TO IUTS

Given the fact that the min test is not of exact size, tends to have low power in realistic situations, and does not fully utilize all available information, we have considered and tried to develop an alternative frequentist approach to IUT by reframing the compound null hypotheses. Define $\mu_{iA} = \theta_{iA} - \Delta_A$, $\mu_{iB} = \theta_{iB} - \Delta_B$, and $\mu_i = \mu_{iA} * \mu_{iB}$. Consider testing a reframed compound hypothesis H_0: $\mu_i = 0$ versus H_1: $\mu_i \neq 0$ for any gene i. For H_0 to be true, at least one mean must be zero. Suppose that X and Y are random variables with mean μ_X and μ_Y, and variance σ^2_X and σ^2_Y, respectively. We compute the estimate $\hat{\mu}_{xy} = \overline{xy}$. We then constructed a t-like statistic by dividing $\hat{\mu}_{xy}$ by its estimated standard error. Unfortunately, the best sample estimate of the standard error is not clear. Drawing on the work of Goodman (1960), we tried several variations and then, because there was no way to insure that the test statistics was distributed as t, conducted significance testing via the bootstrap per the guidelines of Hall & Wilson (1991). Figure 2 displays the scatter plot of p-values obtained by the min test (P_{CIUT}) and those obtained by the best performing alternative frequentist IUT (P_{FIUT}) we could construct. As shown in Figure 2, these two tests seem to be very consistent, and our newer alternative is rarely if ever better than the min test suggesting that this approach may offer no advantage over the min test. We have therefore abandoned the attempt to construct a more powerful frequentist IUT for the time being.

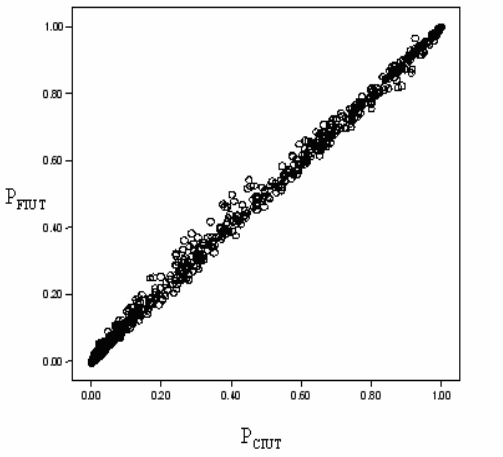


Figure 2. A scatter plot of the p-values (P_{CIUT}, P_{FIUT}).

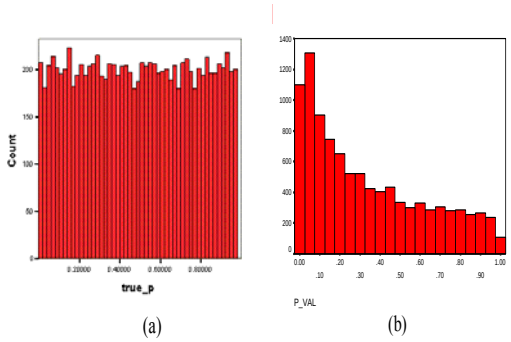
By contrast, the Bayesian approach to hypothesis testing has some advantages over the frequentist approach (Carlin and Louis 2000). Bayes theorem can be applied to compute the posterior probability of uncertainty of each hypothesis based on the data. Under any respective (either informative or non-informative) prior information, Bayes rule applies to obtain the posterior probabilities $P(H_{0i}$ is true given the data) and $P(H_{1i}$ is true given the data) for the null and alternative hypotheses. These probabilities are used to form the Bayes factor, which is the ratio of the posterior odds of H_{0i} to H_{1i}. The Bayes factor provides the odds in favor of H_{0i} over H_{1i} given the data, suggesting a measure of the evidence in favor of one hypothesis over another within the multi-component hypothesis. Thereby, we adapt a Bayesian approach for *intersection-union tests* in an attempt to strengthen statistical analysis and to provide quantitative evidence that supports our decisions.

The power of Bayesian IUT we propose is expected to be greater than that of the traditional IUT, which uses only the largest p-value, due to addition of prior information and full utilization of the information in all p-values. Indeed, a few studies (Sarkar *et al.* 1995; Westfall *et al.* 2001) point out that the min test of IUTs is excessively with lower power near the null component values of zero (e.g., $\Delta_A = \Delta_B = 0$). To enhance power, Snapinn and Sarkar (1996) derived an alternative test that used prior information and calculated accurate estimates of the suggestive null values. They showed that this procedure was substantially more powerful than the min test when the null component values were equal or nearly equal to each other (e.g., $\Delta_A \approx \Delta_B$). Unfortunately, it requires specification of an exact value for the alternative hypothesis which seems impractical in genomic contexts.

BAYESIAN APPROACH TO IUTs

Allison *et al.* (2002) presented a mixture model approach for the analysis of gene expression data. Under the global null hypothesis that there is no difference in gene expression levels between two groups for any gene, the distribution of p-values is uniform on the interval [0,1] regardless of the statistical test used as long as that test is valid. Otherwise, the probability density function (PDF) of p-values would be some monotonically decreasing function on interval [0,1] if the null hypothesis is false. As shown in Figure 3 from Allison *et al.,* under the alternative hypothesis,

the PDF of p-values tends to go higher near zero than around one.



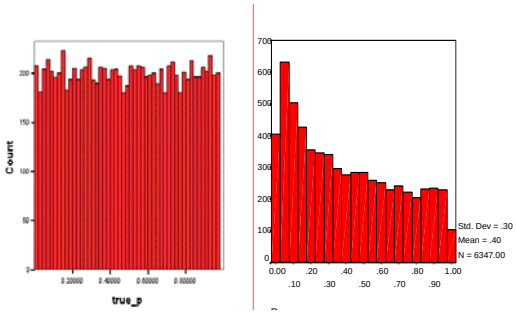
(a) (b)

Figure 3 . Mixture Model Approach from Allison et al. (2002). (a) Under the null hypothesis, the distribution of p-values is uniform on the interval [0,1] regardless of the sample size and statistical test used (as long as that test is valid), and (b) Under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one.

Allison *et al*. (2002) used a Bayesian approach to estimate the number of genes with a real difference in expression levels (e.g., the proportion of p-values that do not fall in a uniform distribution) by fitting the log likelihood function of a mixture of uniform and beta distributions. The log likelihood function of the mixture model with v+1 components is defined as

$$L_{v+1} = \sum_{i=1}^{k} \ln\left[\lambda_0 \beta(1,1)(x_i) + \sum_{j=1}^{v} \lambda_j \beta(r_j, s_j)(x_j)\right], \quad (1)$$

where $\beta(r, s)(x)$ is the density function for the beta distribution with two shape parameters, r and s, and x_j is the p-value for the i-th test, λ_0 is the probability of a randomly chosen test of a true null hypothesis, and λ_j is the probability of a randomly chosen test of a false null hypothesis from the j-th component of beta distribution. If any of v components of the mixture model is not zero, then the null hypothesis is rejected and one concludes that there is statistically significant evidence that one or more of the genes tested is differentially expressed across the groups. Their approach is illustrated by example of a dataset of two groups of mice: old mice versus old mice that had their caloric intake restricted since weaning as described by Weindruch *et al*. (2001). Each group consisted of three mice. Fitting the model (1) results for λ_1, r_1, and s_1 of 0.29, 0.78, and 3.87, respectively. Given these parameters, we estimated that roughly 29% of the genes were differentially expressed and the best estimate for the number of genes of a real difference is 6347 * 0.29 = 1840, where the total number of genes under study is 6347. The distribution of p-values using this dataset is illustrated in Figure 4.



(a) Null-Simulated *(b) Cortex-Real*

Figure 4. The distribution of p-values under the null (a) and alternative (b) hypotheses with the Mouse Cortex Data (Old Ad Lib vs. Old Calorically Restricted), described by Weindruch et al.(2001). Estimated parameters: λ_1 =.29; r_1=.78; s_1 =3.87.

In summary, the estimates of both uniform and beta components in the mixture model provide estimates of the number of genes that have a real difference or no difference in gene expression levels across the two groups, respectively. In addition, the posterior number of genes with a real difference suggests a "best weighted estimate" of the number of genes. The genes with high posterior probabilities are the most promising candidates for further study. The choice of "high" posterior probability is subject to an experimenter's opinion how much error rate she/he is willing to tolerate. For more details, see Allison *et al.* (2002).

The Bayesian approach using the mixture model can be implemented in the IUT for addressing multi-component hypothesis testing in comparative genomics. Consider two datasets to compare: a lean group and an obese group in two different species, human and mice. The first dataset is from a study of adipocyte (fat cell) RNA from 20 lean and 19 obese Pima indians. The biopsies were taken after overnight fast and none of the individuals had any manifested diseases. These data were generated at the NIDDK Phoenix by Dr. Paska Permana. The second is from a study of mouse adipocytes from 5 *ad lib* fed mice and 5 mice with long term caloric restriction. The biopsies were taken after 16 hour overnight fast. These data were generated by Dr. Kazu Hiigami in Dr. R.

Weindruch's Lab (University of Wisconsin-Madison). Now we wish to find homologous genes in humans and mice that are differentially expressed between obese (or heavier) and non-obese (or lighter) groups in both of the two species. The null hypothesis for each homologous gene-pair is that the mouse homolog is not differentially expressed in mice as a function of caloric restriction, its primate counterpart is not differentially expression in humans as a function of obesity, or both.

The *intersection-union tests* of two-component hypothesis were performed by a simple extension of the mixture model approach of Allison *et al.* (2002). The mixture models for the data from each species were fitted separately and both resulting posterior probabilities for individual genes were multiplied to compute the joint [‡] posterior probability for the use of *intersection-union tests*. One would consider genes to have conserved response across two organisms only if the joint posterior probability is sufficiently high; consequently, one can also estimate the number of genes for which the null hypothesis is false in both mice and human by calculating the sum of all posterior probabilities that the compound null hypothesis is false. Such conserved genes are probably "the best investment" in further studies of global patterns of gene expression relevant to obesity. The density function of the joint posterior probability that the compound null hypothesis is false is depicted in Figure 5.

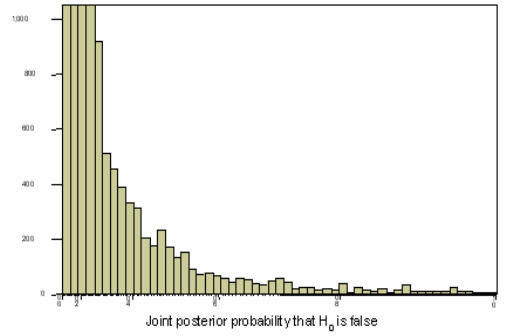


Figure 5. Implementation of IUTs via posterior probabilities by the mixture model. The height represents the frequency of the joint posterior probability that the compound null hypothesis is false.

[‡] The joint posterior probability is computed as the product of two individual posterior probabilities obtained by the mixture model method.

In the example chose, we had two independent datasets. Because of this independence, it is clear that $P(A \cap B) = P(A)P(B)$ and our work was made easy. However, in some other situations, things might not be so simple. For example, if we had taken a single set of mice and measured gene expression via microarrays in their skeletal muscle and adipose tissue, then those measurements would not necessarily be independent and the product rule would not necessarily hold. Adequately estimating $P(A \cap B)$ in such cases remains work for future research.

In conclusion, the potential of comparative genomics in discovery of knowledge about gene function is tremendous. A few approaches are being investigated to identify functionally conserved genes, i.e., genes that yield similar results across species, tissues, or situations. In this paper, we presented a novel *intersection-union test* for multi-component hypothesis to identify functionally conserved genes in different stimuli. The intersection hypothesis helps to compose a single compound hypothesis consisting of multiple components. However, the multiple components of the null hypothesis being tested pose challenges, such as multiple testing and lower statistical power. The *intersection-union tests* enable one to test the intersection hypothesis without multiplicity correction. Moving on to implementation of IUT, we herein proposed the hybrid approach of frequentist and Bayesian that may be a useful method in comparative genomic settings. Further development of this approach and application needs to be undertaken in future studies.

ACKNOWLEDGEMENTS

We gratefully acknowledge Dr. Paska Permana at the NIDDK Phoenix and Drs. Kazu Hiigami and Richard Weindruch (University of Wisconsin-Madison) for sharing their datasets for this paper. This research was supported in part by NSF grants 0090286 and 0217651 and NIH grants U54CA100949 and R01AG018922.

REFERENCES

Allison D. B., Gadbury G. L., Heo M, Fernández J. R., Lee C., Prolla T. A., Weindruch R. (2002) A mixture model approach for the analysis of microarray gene expression data. Computational Statistics & Data Analysis 39:1-20

Bechhofer, R. E. and Dunnett, C. W. (1988) Tables of percentage points of multivariate student t distributions. Selected Tables in Mathematical Statistics 11:1-371, Providence, Rhode Island: American Mathematical society

Berger R. L. (1982) Multiparameter hypothesis testing and acceptance sampling. Technometrics 24:295-300

Berger R. L. and Hsu J. C. (1996) Bioequivalence trials, intersection-union tests, and equivalence confidence sets. Statistical Science 11:283-319

Carlin B. P. and Louis T. A. (2000) *Bayes and Empirical Bayes methods for data analysis*, 2nd ed. Chapman & Hall/CRC

Coffman C. J., R. W. Doerge, M. L. Wayne, and L. M. McIntyre (2003) Intersection tests for single marker QTL analysis can be more powerful than two marker QTL analysis. BMC Genetics, 4(1):10

Dunnett, C. W. and Tamhane, A. C. (1992) A step-up multiple test procedure. Journal of the American Statistical Association 87:162-170

Goodman, L. A. 1960. On the exact variance of products. Journal of the American Statistical Association 55:708-713.

Hall P. and Wilson SR. "Two guidelines for bootstrap hypothesis testing." Biometrics. 1991; 47: 757–62.

Harman K. J., Couper L. L., Linder V. (2000) Strain-dependent vascular remodeling phenotypes in inbred mice. Am. J. Pathol. 156:1741-1748

Kuhel Z. B., Witte D. P., Hui D. Y. Distinction in genetic determinants for injury-induced neointimal hyperplasia and diet-induced atherosclerosis in inbred mice. Arterioscler Thromb Vasc Biol. 22:955-960

McCarroll S. A., C. T. Murphy, S. Zou, S. D. Pletcher, C. S. Chin, Y. N. Jan, C. Kenyon, C. I. Bargmann, and H. Li (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nature Genetics, 36 (2): 197-204

Paigen B., Carey M.C. (2002). Gallstones. In: The *Genetic Basis of Common Diseases*, King R.A., Rotter J. I., Motulsky A. G (eds), 2nd ed. Oxford University Press, pp:298-335

Roy S. N. (1953) On a heuristic method of test construction and its use in multivariate analysis. Ann. Math. Statist. 24:220-38

Sarkar S. M., Snapinn S. M., Wang W. (1995) On improving the min test for the analysis of combination drug trials. J Statist Comput Simul 51:197-213

Snapinn S. M. and Sarkar S. K. (1996) A note on assessing the superiority of a combination drug with a specific alternative. J Biopharm Stat 6:241-251

Stoll M., Kwitek-Black A.E., Cowley A. W. Jr, Harris E. L., Harrap S. B., Krieger J. E.. Printz M. P., Provoost A. P., Sassard J., Jacob H. J. (2000) New target regions for human hypertension via comparative genomics. Genome Res 10:473-482

Sugiyama F., Churchill G. A., Higgins D. C., Johns C., Makaritsis K. P., Gavras H., Paigen B. (2001) Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics 71:70-77

Troendle, J. F. (1996) A permutational step-up method for testing multiple outcomes. Biometrics 52:846-859

Wang X. and Paigen B. (2002) Comparative genetics of atherosclerosis and restenosis: Exploration with mouse models. Arteroscler Thromb Vasc Biol 22:884-886

Weindruch R., Kayo T., Lee C., Prolla T. A. (2001) Microarray profiling of gene expression in aging and its alteration by caloric restriction in mice. The Journal of Nutrition 131:918S-923S

Westfall, P. H. and Young, S. S. (1989) P-value adjustments for multiple hypothesis testing in multivariate binomial models. Journal of the American Statistical Association 84:780-786

Westfall, P.H., Ho, SY, Prillaman, B. A. (2001) Journal of Biopharmaceutical statistics 11:125-138

Zhang Q., D. Biochard, I. Hoeschele, C. Ernst, A. Eggen, B. Murkve, M. Pfister-Genskow, L. A. Witte, F. E. Grignola, P. Uimari, G. Thaller, and M. D. Bishop (1998) Mapping quantitative trait loci for milk production and health of diary cattle in a large outbred pedigree. Genetics, 149: 1959-1973