Composite Hypothesis Testing from Multiple Experiments: An Approach Built on Intersection-Union Tests and Bayesian Posterior Probabilities

Kyoungmi Kim and David B. Allison

Department of Biostatistics, Section on Statistical Genetics,
University of Alabama at Birmingham

This chapter provides knowledge of analytical statistical methods in the search for genes that are consistently differentially expressed in a response of interest across multiple experiments. Specifically, we discuss methods for combining multiple individual hypotheses for multiple experiments into a single composite hypothesis-based statistical test built on intersection-union tests (Kim *et al*., 2004). Moreover, we consider some of issues involved in those methods.

Introduction

When there are multiple experiments, all of which address the same hypothesis, combining information from several experiments or studies can draw a general conclusion. If the individual experiments are identical and carried out independently, then it may be possible and desirable to combine the raw data into a single data. An optimal test is then based on significance of a statistical test using a pooled test statistic obtained the overall mean of the total observations and their pooled variance. Alternatively, if they are not identically conducted, then it is inappropriate to combine data. In this case, it is reasonable to combine results of individual experiments, rather than the original observed raw data. The first case can be met through relatively well-established statistical techniques. The later case, however, is not as easily met and poses challenges in high dimensional biology; it is important to present reformulated statistical approaches to overcome such challenges. Herein, we specifically consider situations in which multiple experiments are conducted under different conditions such as cross-stimulus and cross-species studies. For example, we seek to identify genes 1) that are differentially expressed in response to a particular stimulus across several model organism species or 2) that are differentially expressed in the similar way in response to multiple stimuli. There are two broad ways of combining information across multiple studies: one that concerns the combination of hypotheses (Rhodes *et al.*, 2002, 2004; Kim *et al*., 2004), and the other that concerns the combination of the results of individual statistical tests (Choi *et al.*, 2003; Moreau *et al.*, 2003, Parmigiani *et al.*, 2004; Stevens and Doerge, 2005). In this chapter, we focus on how multiple individual hypotheses can be combined in the context of microarray studies.

One of the questions of interest when an investigator conducts n multiple microarray experiments addressing the same question would be “Are there genes that are differentially expressed in the same fashion across the multiple experiments?” A first attempt to finding an answer would be to combine the significances of the gene g from the n experiments into a single combined significance. For example, one uses a traditional statistical method, such as t-test, to obtain a significance score, such as a p-value that assesses the probability that the observed level of differential gene expression could have occurred by change. However, this procedure tests thousands of genes per experiment simultaneously and thus causes many false positives. Here, a false positive is defined as a gene that is truly not differentially expressed but appears to indicate differential expression by chance and because of experimental noise. Therefore, there is a need to adjust p-values to control false positives for massive multiple testing (Westfall and Young, 1993; Benjamini and Hochberg, 1995; Reiner et al., 2003). Once a p-value per gene for each study is obtained, we combine the n gene-specific p-values to determine whether the same gene g shows significant difference in expression in all of the n experiments. Combining the significance results, such as the maximum p-value of the n p-values, can assess the overall significance for all experiments even this approach is very conservative. If the maximum p-value is less than a pre-specified rejection threshold, then all p-values are below the threshold. Thereby, the null hypothesis that the gene g is not differentially expressed in response to a stimulus of interest for each of all n experiments is rejected, and the alternative hypothesis that the gene g shows differential expression for each of the n experiments is accepted. As a result, one can conclude that there are genes that are differentially expressed in the equivalent fashion across the multiple experiments. However, this approach is very sensitive to outliers. Although such combination methods of p-values are easy to implement, they also pose challenges both statistically and computationally. All multiple experiments should address the same hypothesis, all of which infer the unique parameter of interest. The p-value-based combination methods within parametric settings do not provide information regarding the magnitude of significance of differential expression for genes, other than classified conclusion between “there is significance of differential expression” and “there is no significance of differential expression.” Therefore, we desire integrative procedures that address multiple hypotheses simultaneously and assess all results as “evidence” that can be thought of as a “degree” of the confidence on the final inferential decisions.

Through this chapter, we discuss theoretical details of the composite hypothesis testing methods, especially intersection-union tests, and illustrate their applications to gene expression studies. Although the intersection-union test provides a convenient way of performing simultaneous testing per gene while retaining the overall significance level, it also has disadvantages. These disadvantages include being overall results conservative and only yielding them classifiable as significant or not, but not the degree of uncertainty of the significance. Therefore, we explore alternative strategies for applying the intersection-union method to detect genes that exhibit equivalent responses across multiple experiments. Lastly, we discuss some of issues involved in these methods.

Composite Hypothesis Testing

Suppose that n experiments have been conducted independently to determine whether the gene g is differentially expressed in a response of interest. For simplicity, we consider one-side hypotheses without loss of generality: for the particular gene g, suppose that it is of interest to test the individual hypotheses

$$H_{0i}: \theta_i = 0 \qquad \text{vs.} \qquad H_{ai}: \theta_i > 0, \qquad i=1, 2, \ldots, n, \tag{1}$$

using observed values x_i of the random variable X_i having a distribution depending on $\theta_{i,}$ the true mean expression level of the gene g in the ith experiment . A statistical test is on the basis of the test statistic T_i, which large values of T_i resulting small p-values lead to rejection of the null hypothesis $H_{0i}: \theta_i = 0$ in favor of the alternative hypothesis $H_{ai}: \theta_i > 0$. If t_i is the real value of the test statistic T_i, then the p-value, or significance level, of the test for the ith experiment is computed as

$$p_i = \text{Prob}(T_i > t_i \mid H_{0i}).$$

Assuming that the test statistic T_i has a continuous distribution under the null H_{0i}, p_i is uniformly distributed on the interval [0, 1]. The individual hypotheses do not need to have the same substantive meaning across all the experiments.

With respect to testing of multiple hypotheses, there are two types of tests: a union-intersection test and an intersection-union test. A union-intersection test (UIT) (Roy, 1953) is to determine if the gene is differentially expressed in “at least” one of the experiments. In other words, we wish to test the composite null hypothesis, which is the union of all the alternative hypotheses,

$$H_0 = \bigcap_{i=1}^{n} H_{0i} \quad \text{vs.} \quad H_a = \bigcup_{i=1}^{n} H_{ai}. \tag{2}$$

Note that the composite null hypothesis is rejected if any one of the individual null hypotheses is rejected at the multiplicity-adjusted threshold that controls an overall experiment type I error rate.

Alternatively, an intersection-union test (IUT) (Berger, 1982; Berger and Hsu, 1996) is to determine if the gene is differentially expressed in “all” of the experiments. We wish to test the composite null hypothesis, which is the intersection of the n alternative hypotheses,

$$H_0 = \bigcup_{i=1}^{n} H_{0i} \quad \text{vs.} \quad H_a = \bigcap_{i=1}^{n} H_{ai}. \tag{3}$$

For this case, the composite null hypothesis is rejected only if all the individual null hypotheses are rejected. The IUT does not need a multiplicity adjustment. Therefore, the IUT is well-suited for determining whether genes are differentially expressed in the same way in response to a stimulus of interest across the multiple experiments. We present greater details in subsequent sections.

An example of the application of IUTs would be a situation in which we wish to find genes that are differentially expressed in response to caloric restriction in two experiments, A and B. To find these genes, 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 no significant difference is observed in A, or no significant difference is observed in B, or both, then we can not reject the composite null hypothesis. Mathematically, we can express this argument as follows:

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

$$H_0 = H_{01}\ U\ H_{02}$$

as the union of

$$H_{01}\text{: } \theta_{iA} = 0 \text{ and } H_{02}\text{: } \theta_{iB} = 0$$

vs.

$$H_a = H_{a1} \cap H_{a2}$$

as the intersection of

$$H_{a1}\text{: } \theta_{iA} > 0 \text{ and } H_{a2}\text{: } \theta_{iB} > 0 .$$

A layout of this two-component composite hypothesis is illustrated in Table 1. The ith gene is acceptable as a consistent gene that is differentially expressed in response to caloric restriction in both experiments A and B if the composite null hypothesis H_0 is rejected (i.e., if either H_{01} or H_{02}, or both, are rejected at level α). In other words, H_a is true if and only if both H_{a1} and H_{a2} are true. Hence, individual tests for the multiple experiments can be combined by means of the IUTs to yield a single, overall test of the consistent gene across the different experiments. The individual tests do not need to address the same consistent hypothesis across the multiple experiments.

Table1. A layout of two-component hypothesis for intersection-union test

	Null $H_{01}: \theta_{iA} = 0$	Alternative $H_{a1}: \theta_{iA} > 0$
Null $H_{02}: \theta_{iB} = 0$	Null H_0	Null H_0
Alternative $H_{a2}: \theta_{iB} > 0$	Null H_0	Alternative H_a

Assessing significance of the composite hypothesis

There are several strategies for creating a comprehensive p-value for the composite hypothesis in IUTs combining individual p-values corresponding to individual hypotheses. The p-value-based combining methods include Fisher's inverse chi-square method (Fisher, 1932), Tippett's minimum method (Tippett, 1931), the inverse normal method (Stouffer et al., 1949), average method, Pearson's method, and maximum method. Table 2 lists the combined p-value used in each case of a function of p_i. Combination methods based on p-values are nonparametric procedures because under the null H_{0i}, p-values are uniformly distributed on the interval [0, 1] and independent regardless the distribution of the test statistics $\text{T}_1, \ldots, \text{T}_n$. Such methods that are not distribution specific are sometimes referred as "omnibus" procedures. We discuss theoretical aspects of the first three commonly used methods as follows.

Table 2. P-value-based Combining Methods.

Method	Combined p-value

Fisher	$\prod_{i=1}^{n} p_i$
Tippett	$\min_{1\le i\le n} p_i$
Inverse Normal	$\frac{1}{\sqrt{n}}(\sum_{i=1}^{n}\Phi^{-1}(p_i))$
Average	$\frac{1}{n}\sum_{i=1}^{n} p_i$
Pearson	$1-\prod_{i=1}^{n}(1-p_i)$
Maximum	$\max_{1\le i\le n} p_i$

Fisher's inverse chi-square method. Fisher's method is the most widely used procedure. Given *n* independent experiments and their corresponding p-values, $p_1, \ldots, p_n$, obtained by a valid statistical test, this method uses the product $\prod_{i=1}^{n} p_i$ to combine the p-values for the overall significance. Under the H_{0i} for the ith experiment, $-2\log p_i$ has a chi-square distribution with 2 degrees of freedom. Because the sum of independent chi-square variables has also a chi-square distribution with 2n degrees of freedom, the composite null hypothesis H_0 is rejected if

$$-2\log(\prod_{i=1}^{n} p_i) = -2\sum_{i=1}^{n}\log p_i \ge \chi^2_{2n,\alpha},$$

where the critical value $\chi^2_{2n,\alpha}$ is the upper α percentile of the chi-square distribution with 2n degrees of freedom. In situations where cross-study variation is significant, Fisher's method can be modified by assigning weights to reflect differences across individual experiments, that is, $-2\log(\prod_{i=1}^{n} p_i^{w_i}) = -2\sum_{i=1}^{n} w_i \log p_i$, where w_i is the weight for the ith experiment.

Tippett's minimum p-value method. Under the H_{0i}, each p-value has a uniform distribution and the minimum of n p-values, $p_{(1)}$, is compared with $1-(1-\alpha)^{1/n}$ to determine whether the composite null hypothesis H_0 is rejected at level α. The null hypothesis H_0 is rejected if

$$p_{(1)} < 1-(1-\alpha)^{1/n}.$$

Inverse normal method. This procedure involves transforming each p-value to the standard normal score, $\Phi^{-1}(p_i)$ and then averaging the transformed scores. When the H_0 is true, the statistic $\frac{1}{\sqrt{n}}(\sum_{i=1}^{n}\Phi^{-1}(p_i))$ has the standard normal distribution. The H_0 is rejected if

$$\frac{1}{\sqrt{n}}(\sum_{i=1}^{n}\Phi^{-1}(p_i)) \geq \Phi^{-1}(1-\alpha),$$

Where Φ is the standard normal cumulative density function. The weighted statistic $\frac{1}{\sqrt{n}}(\sum_{i=1}^{n} w_i\Phi^{-1}(p_i))$ can be also used.

Any of the listed methods can be easily implemented to assess the significance of composite hypothesis testing. For example, we consider testing a two-component composite hypothesis as described in the caloric restriction example and determining its significance with a combined p-value based on the maximum method. In UITs, each component of the composite null hypothesis is determined to be rejected or not at the multiplicity-adjusted threshold that controls an overall experiment type I error rate. The rejection region for this UIT is the union of rejection regions that correspond to the individual tests (see the left panel of Figure 1). In contrast, IUTs maintain a pre-specified type I error rate without multiplicity adjustment for multiple components. The rejection region for this intersection-union test is the intersection of the rejection regions that correspond to the two individual tests, that is,

$$\bigcap_{i=1}^{2} R_i = \{\min(T_1(x), T_2(x)) \geq c_\alpha\},$$

where for the ith 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 the type I error rate of α (see the right panel of Figure 1). The p-value for the minimum of test statistics, which is the maximum p-value, is only used to determine whether the composite null hypothesis will be rejected, regardless of the magnitudes of any other p-values. The other procedures listed can be applied in place of the maximum method.

< Insert Figure 1 here >

Rhodes *et al.* (2002) have illustrated the use of Fisher's inverse chi-square method in the context of IUTs. They proposed a meta-analytic approach to microarrays that combined results of four individual studies to determine genes that were differentially expressed in response to prostate cancer. Their approach was a variant of Fisher's inverse chi-square method. More details on Fisher's inverse chi-square method will be addressed later in this chapter. They first performed individual study analysis across the four studies by treating each gene in each study as an independent hypothesis. A significance score, q-value, was assigned to each gene based on a multiple testing correction through false discovery rate (FDR), which is defined as the expected ratio of the number of true positives over the total number of true positives plus false positives (Storey and Tibshirani, 2003). Then a combined summary statistic from the different q-values was computed and used as a test statistic for testing the hypothesis that significant results from the different studies did not correspond to the same gene. For meta-analysis, their significance for each gene was evaluated based on the distribution of randomly generated summary statistics from the random permutation t-tests. Each gene then estimated the lowest FDR, expressing the likelihood that the q-values of the gene from the individual studies were assigned to the gene by random selection from the respective studies. If the lowest FDR rate of the gene was significant, then one would consider that the gene showed differential expression in all of the four studies. In other words, if the gene was significant only in some studies, but not all of them, the lowest FDR rate would not be significant for differential expression in response to prostate cancer at a threshold. This method identified and assessed the intersection of multiple gene expression profiles from the four microarray datasets.

Despite the appealing property of omnibus procedures combining p-values, they do not provide quantitative measure of “evidence” against the null hypothesis. Therefore, the statistical significance of omnibus tests is very poor to draw general conclusions about the magnitude, direction, and consistency of differential expression across multiple experiments. To obtain quantitative measures of evidence, we turn to a Bayesian deviation to see if a Bayesian-type measure, expressed in terms of posterior probability that the null hypothesis is true, provides actual quantitative evidence against the null hypothesis. We discuss Bayesian procedures of how to utilize information about the distributions of individual test statistics or p-values given that the alternative hypothesis is true in the following subsection.

Measuring Bayesian evidence in multiple hypothesis testing

By considering multi-component hypotheses as in (2) or (3), Pratt (1965) showed that in one-sided hypothesis testing problems, the p-value can be approximately equal to the posterior probability of the null hypothesis and the posterior probabilities of the individual hypotheses can replace the individual p-values instead. Hence, it is possible to reconcile the Bayesian posterior probability that the null is true and the frequentist p-value (see Casella *et al.*, 1987). We can directly use the product of the posterior probabilities of individual hypotheses, a Bayesian variant of Fisher's method, to assess evidence against the composite hypothesis. However, the prior probability of the intersection of parameters, $\bigcap_{i=1}^{n}\theta_i$, is much smaller than the prior probability of each θ_i. As a result, the posterior probability of $\bigcap_{i=1}^{n}\theta_i$ is also smaller to the posterior probability of each θ_i, even the posterior probability of each θ_i is substantial. This problem arises more seriously as the number of experiments, n, is growing larger. To avoid an undesirable favor of the alternative hypothesis against the null hypothesis, we may need to adjust the prior distributions of p_i, i=1, …, n, so that the chances of being the null and the alternative true are equal.

Application of Bayesian measure to Intersection-Union Tests

Here, we try to utilize a Bayesian-type measure to assess the uncertainty of the combined results based on the estimated Bayes factor, which is the ratio of the posterior probabilities of the null hypothesis to the alternative hypothesis, providing the odd in favor of the null over the alternative. Specifically, we will implement a Bayesian approach to intersection-union tests in an attempt to strengthen statistical analysis and to provide quantitative “evidence” that supports our inferential decisions.

Recall that the null and alternative hypotheses in (3) of the intersection-union test procedure is

$$H_0 = \bigcup_{i=1}^{n} H_{0i} \quad \text{vs.} \quad H_a = \bigcap_{i=1}^{n} H_{ai}, \quad i = 1, \cdots, n.$$

That is, the null H_0 states that the gene g in the question is not differentially expressed in at least one of the experiments, and the alternative H_a states that the gene g is differentially expressed in all of the experiments. If we reject H_0 using any of the omnibus procedures, one would consider reasonable evidence to support that the gene g is differentially expressed in the equivalent (at least similar) way to the response of interest across the experiments and refer this gene to a “conserved” gene. Here we apply a mixture model approach (Allison *et al.*, 2002) to calculate a combined p-value for the null hypothesis in (3) using individual resulting p-values of the individual component hypotheses based on the Fisher’s method.

Briefly, the mixture model approach presented Allison *et al.* (2002) for the analysis of gene expression data was on the base of the uniformity of p-values under the null hypothesis was ture. Under the null hypothesis that there was no difference in gene expression levels between two groups for any gene, the distribution of p-values was uniform on the interval [0, 1], regardless of the statistical test used as long as that test was valid. Otherwise, if the null hypothesis was false, the probability density function (PDF) of p-values would be some monotonically decreasing function on interval [0, 1]. As shown in Figure 2 from Allison *et al.* (2002), under the alternative hypothesis that there is at least one gene differentially expressed between two groups, the PDF of p-values tends to go higher near zero than around one.

< Insert Figure 2 here >

Based on the distribution of p-values under the alternative hypothesis, Allison *et al*. (2002) used the Bayesian approach to estimate the number of genes with a real difference in expression levels by fitting the log likelihood function of a mixture of uniform and beta distributions. In other words, they estimated the proportion of p-values that might not fall in a uniform distribution by calculating weights of a uniform distribution under the true null hypothesis and a beta distribution under the false null hypothesis. The log likelihood function of the mixture model with *m*+1 components was defined as

$$L_{m+1} = \sum_{i=1}^{k} \ln\left[\lambda_0 \beta(1,1)(x_i) + \sum_{j=1}^{m} \lambda_j \beta(r_j, s_j)(x_j) \right], \qquad (4)$$

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 *m* components of the mixture model is not zero, then the null hypothesis is rejected, indicating that there is at least one gene that behaves differently between the two groups. Therefore, one can conclude that there is statistically significant evidence that one or more of the genes tested is differentially expressed across the groups. The fitted model was used to calculate the posterior probability that a gene was differentially expressed between the two groups. This posterior probability per gene provided a quantitative measure that the gene was truly different in expression levels between the two groups. An example of the distribution of p-values is illustrated in Figure 2. Once the posterior probabilities of the individual hypotheses are estimated, we adapt the Fisher's method for assessing Bayesian-type significant evidence against the composite null hypothesis H_0 in (3). We calculate the product of the posterior probabilities as the joint posterior probability per gene. The genes with high joint posterior probabilities are the most promising candidates as conserved genes, which are targets for further study. Nevertheless, it is noteworthy to mention that the choice of "high" joint posterior probability is subject to an investigator's opinion of how much error rate she/he is willing to take at risk. This demonstrated approach is a variant of the weighted Fisher procedure by incorporating prior information about the distribution of data (Kim *et al.*, 2004).

For example, we consider two datasets to compare two groups: a lean group and an obese group in two different species, human and non-human (mice). The first dataset is from a study of adipocyte (fat cell) RNA from 20 lean and 19 obese Pima Indians. 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). 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 test of two-component composite hypothesis was 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 to obtain the posterior probabilities for all the genes. Then two resulting posterior probabilities of the two species per gene were multiplied to compute the joint posterior probability for the use of intersection-union test. One would consider a gene to have conserved response across the two species 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 the joint posterior probabilities across all the genes that the composite null hypothesis is false. Such conserved genes are probably "the best investment" in further studies of global patterns of gene expression relevant to caloric restriction and its influence on obesity. The density function of the joint posterior probability that the composite null hypothesis is false is depicted in Figure 3.

< Insert Figure 3 here >

Issues related to intersection-union tests

There are a few issues related to the use of intersection-union tests for examining multiple microarray experiments. The issues include cross-experiment variation due to different experimental conditions and dependence between experiments. Different experiments can be quite different regarding some factors. Examples of the factors would be the use of different microarry platforms (Moreau *et al.*, 2003; Parmigiani *et al.*, 2004) and the inconsistency of the number of genes under study (Choi *et al.*, 2003), especially when two different, distance species are compared from each other. This inconsistency may cause results to be biased and incomplete by mismatching pairs of genes across experiments. Another important issue is independence between datasets. In the human-nonhuman example described, we had two independent datasets. Because of this independence, it was clear that the joint posterior probability was equal to the product of two marginal posterior probabilities (e.g., $P(A \cap B) = P(A)P(B)$). However, in some other situations, datasets obtained from multiple experiments are not mutually independent, so the multiplication rule for the joint probabilities no longer holds. For instance, multiple case-control experiments were conducted by crossing two types of experimental model strains: one was a wild type and another was a genetically modified type of the wild type. All experiments compared a common wild type as control with different modified types as cases. Then gene expression studies were conducted using microarray technology based on a common reference design to identify genes that were commonly differentially expressed in all of the varied cases. In this case, all the experiments shared the common biological source because the cases were generated by mutating the same wild strain. Thus they were not biologically independent and caused confounding effects on the biological variation. Therefore, the datasets of gene expression contained the common variant. Another example would be a case in which we had taken a single set of mice and measured gene expression via microarrays in their skeletal muscle and adipose tissue. These two tissues from the same mouse are not independent. As in these two examples, the measurements would not necessarily be independent, and the multiplication rule does not hold. Therefore, when the underlying assumption of independence among multiple experiments or datasets is violated, adaptive modification for estimation of the joint posterior probabilities is required.

Summary

Cross comparisons for multiple experiments are sometimes invaluable in helping to global understanding of the genetic pathways of complex diseases. However, the common strategy often involves confirming that the genes that are differentially expressed in one experiment are also differentially expressed in other experiments. This confirmatory analysis requires greater efforts both statistically and computationally in order to compute and summarize all results and to draw a comprehensive, generalized conclusion for the multiple experiments. Therefore, methods for combining all multiple tests into a single hypothesis-based test are very attractive. In this chapter, we have introduced theoretical and practical aspects of composite hypothesis testing built on traditional intersection-union tests in the search for “conserved” genes that are differentially expressed in response to a stimulus of interest across multiple experiments or “consistent” genes that are differentially expressed in the same fashion to multiple stimuli. Moreover, we have also provided a basic understanding of the traditional intersection-union test and shown how this method can be applied to gene expression studies. We have also pointed out the limitations of the intersection-union tests. As an alternative approach to overcome some of the limitations, we have discussed a hybrid approach that mixes frequentist and Bayesian that improves the flexibility and efficiency of the traditional intersection-union tests by incorporating information of data properties. Given the growing demands and rapid advances in genome technology, the continuing development of these statistical methods is needed to be undertaken in future studies.

Software Availability

Software that includes the methods introduced in this paper is available in the web site of the Section on Statistical genetics of Department of Biostatistics at University of Alabama at Birmingham: http://www.soph.uab.edu/ssg_content.asp?id=1163.

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

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Series B 57:289-300

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

Casella, G. and Berger, R. L. (1987) Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. Journal of the American Statistical Association 82: 106-111

Choi, JK., Yu, U., Kim, S., Yoo, O.J. (2003) Combining multiple microarray studies and modeling interstudy variation. Bioinformatics 19 (Suppl. 1): i84-i90

Fisher, R. A. (1932) Statistical Methods for Research Workers (4th ed.), Edinburgh: Oliver and Boyd.

Kim, K., Zakharkin, S. O., Loraine, A. E., Allison, D. B. (2004) Picking the most likely candidates for further development: Novel intersection-union tests for addressing multi-component hypotheses in comparative genomics. Proceedings of the American Statistical Association, ENAR Section [CDRom]. Alexandria (VA): American Statistical Association.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., Dabrowski, M. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. TRENDS in Genetics 19: 570-577

Parmigiani, G., Garrett-Mayer, E.S., Anbazhagan, R., Gabrielson, E. (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clinical Cancer Research 10:2922-2927

Pratt, J. W. (1965) “Bayesian Interpretation of Standard Inference Statements” (with discussion). Journal of the Royal Statistical Society, Series B, 27: 169-203

Reiner, A., Yekutieli, D., Benjamini, Y. (2003) Identifying differentially expressed genes using false discovery controlling procedures. Bioinformatics 19:368-375

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

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., Chinnaiyan, A. M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Research 62: 4427-4433

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T. R., Pandey, A., Chinnaiyan, A. M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci 101:9309-9314

Stevens, J. R. and Doerge, R.W. (2005) Combining Affymetrix microarray results. BMC Bioinformatics 6:57

Storey, J.D. and Tib Tibshirani R. (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci 100:9440-5.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams, R. M., Jr. (1949) The American Solider: Vol 1. Adjustment during Army life. Princeton: Princeton University Press.

Tippett, L. H. C. (1931) The Methods of Statistics. London: Williams and Norgate.

Westfall, P. H. and Young, S.S. (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. John Wiley & Sons, Inc. New York

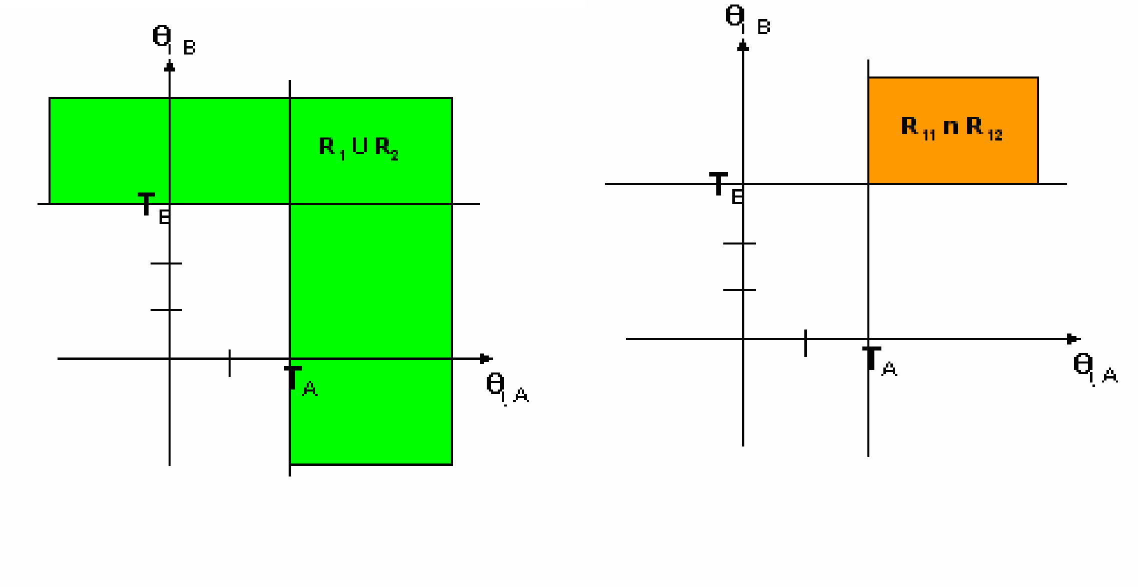
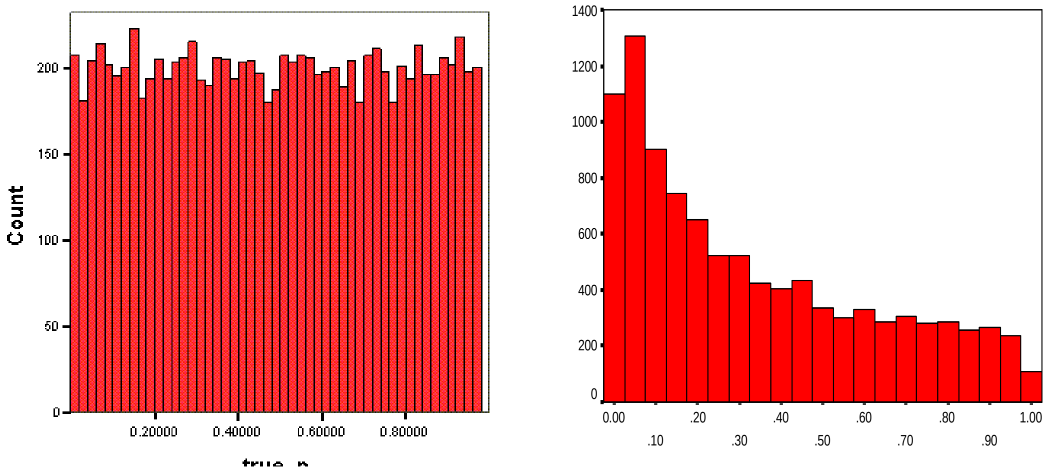


Figure 1. Rejection regions of parameter space corresponding to the alternative hypothesis H_a in UITs (the left panel) and IUTs (the right panel).



(a)

(b)

Figure 2 . 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.

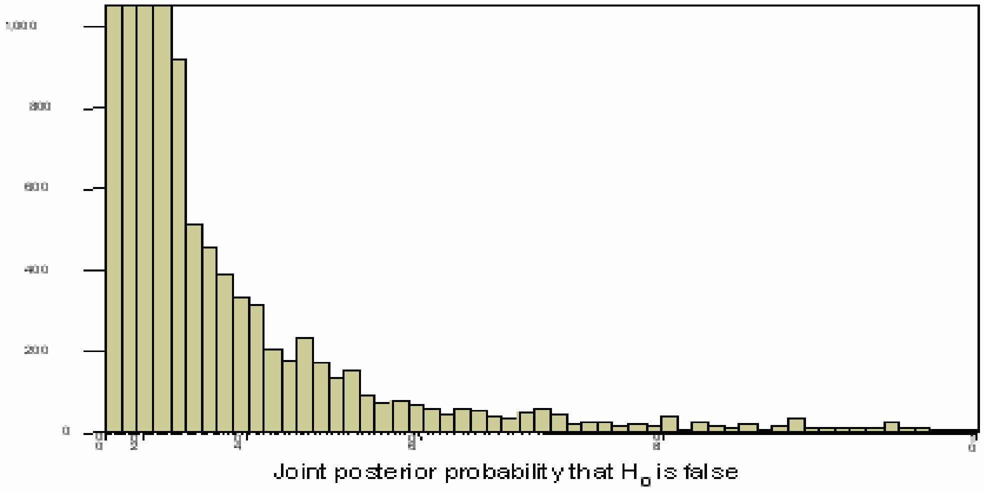


Figure 3. Implementation of IUTs via posterior probabilities by the mixture model. The height represents the frequency of the joint posterior probability that the compound null hypothesis is false.