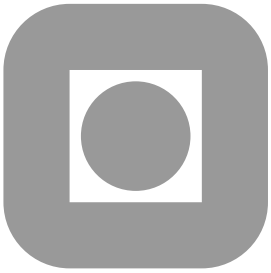
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Statistical Hypothesis Testing of Association Between Two Lists of Genes for a Given Gene Class

by

Clara-Cecilie Günther, Mette Langaas and Stian Lydersen

PREPRINT STATISTICS NO. 1/2006
DEPARTMENT OF MATHEMATICAL SCIENCES



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2006/S1-2006.pdf
Mette Langaas has homepage: http://www.math.ntnu.no/∼mettela
E-mail: Mette.Langaas@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

STATISTICAL HYPOTHESIS TESTING OF ASSOCIATION BETWEEN TWO LISTS OF GENES FOR A GIVEN GENE CLASS

CLARA-CECILIE GÜNTHER*, METTE LANGAAS* AND STIAN LYDERSEN#
* Department of Mathematical Sciences.
# Department of Cancer Research and Molecular Medicine.
The Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

MAY 2006

SUMMARY

Within functional genomics and systems biology, gene expression microarrays have become a valuable tool, and a first step towards arriving at a complete understanding of the systems biology of an organism is often to study lists of genes that are found to be statistically significantly differentially expressed between two conditions. To aid in the further interpretation of such findings *gene class testing* has become a popular and widely accepted analytical tool. Gene classes are often based on Gene Onlology categories. The focus of this report is on statistical hypothesis testing of association between two *intersecting* gene lists for a given class of genes, and we formally state the null and alternative hypotheses for comparing the lists. We develop two new statistical tests, the Unpooled Intersecting Asymptotic (UIA) and the Pooled Intersecting Asymptotic (PIA) test, and in addition we adapt the test of Leisering, Alonzo and Pepe (2000) for comparing two positive predictive values to the case of two intersecting gene lists. To compare the performance of the three tests we have conducted a simulation study.

1 BACKGROUND

When analysing data from high throughput technologies, like microarray experiments, a common aim is to arrive at lists of statistically signifi cantly differentially expressed genes between different situation. What may be more interesting is to understand which biological pathways that are active in the situations under study. To do this we may consider groups of genes instead of single genes. In this report we consider a group (which we call a class) of genes selected from a predefi ned set, i.e. using the Gene Ontology (GO) vocabulary, The Gene Ontology Consortium (2000). Gene Ontology is a vocabulary that describes gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

eGOn[1] (explore Gene Ontology) is a tool that facilitates use of biological background knowledge in analysis of genes selected from high throughput analysis like e.g microarray analysis. Lists of genes containing i.e. potentially differentially expressed genes are submitted through a web interface to

[1] eGOn is available from http://www.genetools.no and presented in Beisvåg, Jünge, Bergum, Jølsum, Lydersen, Günther, Ramampiaro, Langaas, Sandvik and Lægreid (2006).

the annotation database, and eGOn automatically translates the lists to GO-terms annotated to these genes. In addition to powerful graphical displays eGOn offers statistical hypothesis testing to assess the level of similarity between two different gene lists, list A and list B. We refer the reader to Khatri and Dragici (2005) for an overview of statistical tests implemented in different Gene Ontology tools.

This report will focus on statistical hypothesis testing of association between two dichotomized gene lists for a given class of genes. The results are also applicable to other situations, e.g. testing the positive predictive value of two binary diagnostic tests or testing the difference in the prevalence of a disease in two groups in the population.

2 THE NULL HYPOTHESIS

For a randomly chosen gene and a given gene class G, defi ne the following three events:

A = the gene is on list A (e.g. has responded to treatment A)

B = the gene is on list B (e.g. has responded to treatment B)

G = the gene is a member of gene class G.

The complementary event of an event E is denoted by E^*.

As an example, consider a microarray experiment where the objective of the study is to compare the differentially expressed genes from decidual and placental tissue between cases and controls where the cases are women with pre-eclampsia and the controls are healthy pregnant women, Eide, Rolfseng, Isaksen, Mecsei, Roald, Lydersen, Salvesen, Harsem and Austgulen (2006). In our example the list A would be the list of differentially expressed genes between cases and controls in decidual tissue while list B would be the differentially expressed genes between cases and controls in placental tissue. For the given gene class G, we are interested in investigating whether the probability of belonging to gene class G is different for genes on gene list A and genes on gene list B. For each gene on list A, there is a probability $P(G|A)$ of belonging to gene class G, and for each gene on list B, there is a probability $P(G|B)$ of belonging to gene class G. Under the null hypothesis these two probabilites are equal. We formulate the following null hypothesis and alternative hypothesis.

$$\begin{aligned} H_0 : P(G|A) = P(G|B) \text{ vs. } H_1 : P(G|A) \neq P(G|B) \\ H_0 : P(G|A) - P(G|B) = 0 \text{ vs. } H_1 : P(G|A) - P(G|B) \neq 0 \end{aligned} \tag{1}$$

Using the defi nition of conditional probability, the null hypotesis can be written equivalently as

$$\begin{aligned} \frac{P(A \cap G)}{P(A)} &= \frac{P(B \cap G)}{P(B)} \\ \frac{P(A|G)}{P(A)} &= \frac{P(B|G)}{P(B)} \\ \frac{P(A|G)}{P(B|G)} &= \frac{P(A)}{P(B)} \end{aligned}$$

This gives us the following additional interpretation. For a chosen gene class G, the ratio between the probability of membership on list A and membership on list B, is the same as the ratio between the probability of being a member of list A to the probability of being a member of list B for all genes.

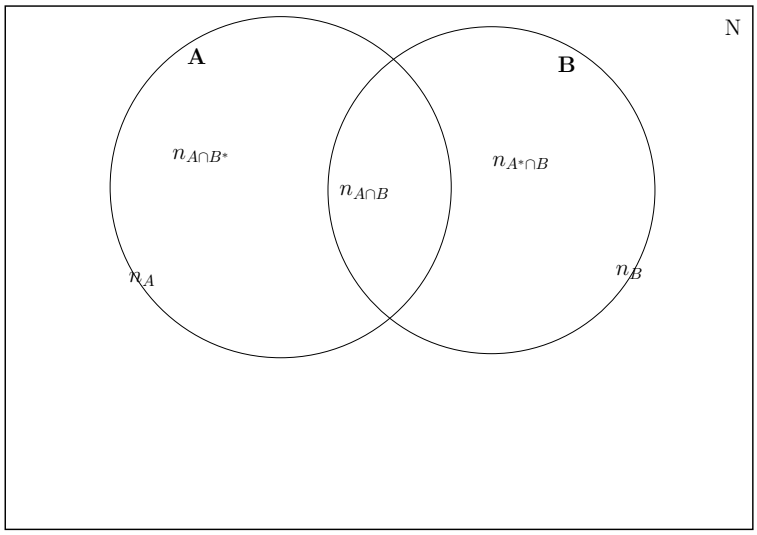


FIGURE 1: The number of genes that are on list A or B or on both lists

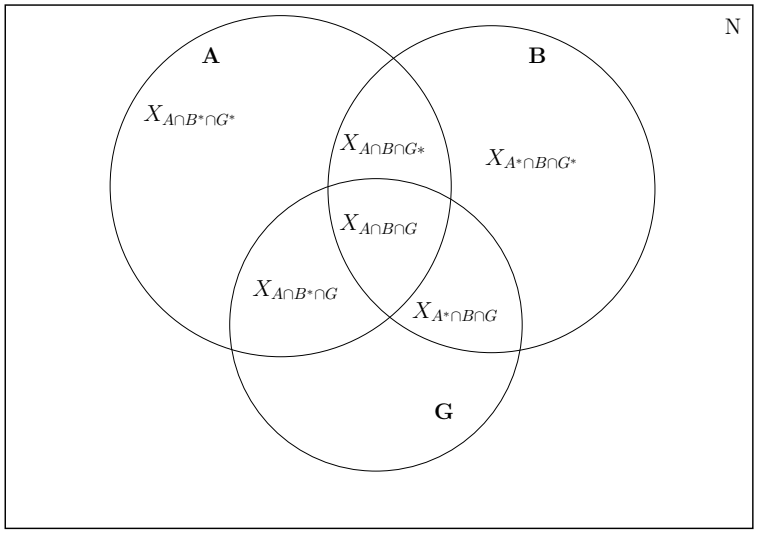


FIGURE 2: The events A, B and G, and random variables counting the number of genes in relevant subsets.

Considering the pre-eclampsia example again, we want to test whether the probability that a gene belongs to e.g. the gene class apoptosis given that it is on list A is equal to the probability that the gene belongs to the apoptosis gene class given that it is on list B.

Figure 1 shows the notation for the number of genes that are on list A, B or both. Figure 2 shows how the events A, B and G are related and the notation used for the number of genes in different subsets. Table 6 in Appendix A gives a more detailed explanation of the notation used further in this report.

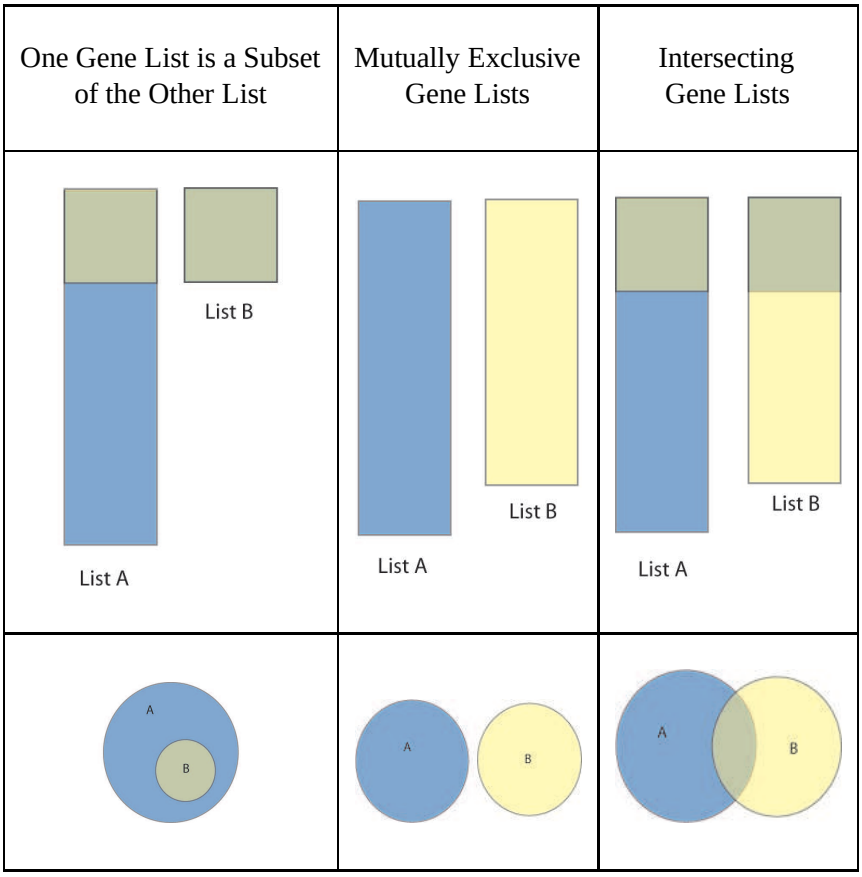


FIGURE 3: The three possible situations presented in the text: One Gene List is a Subset of the Other List, Mutually Exclusive Gene Lists and Intersecting Gene Lists

3 STATISTICAL TESTS

We consider two lists, A and B, and a chosen class G. Our aim is to test the null hypothesis (1). In the case of comparing two lists of genes, list A and list B, we assume that under the null hypothesis the genes on the lists (or subsets of the lists) act independently within each list. When we are comparing positive predictive values of two diagnostic tests the observational unit is the individual and we assume indepencence between the test results of different individuals. When testing for equality of the prevalence of a disease in two groups in the population, again the the observational unit is the individual and these are assumed independent.

Statistically we need to distinguish between three situations, to correctly handle the possible dependencies between gene lists A and B due to the fact that the same observational unit may present on both lists. An illustration of these situations is given in Figure 3.

- One Gene List is a Subset of the Other List: One of the two lists of genes compared is the list containing all genes present in that gene class in the full experiment. (e.g. all genes assayed on the chip in a microarray experiment) see the left column of Figure 3. For our pre-eclampsia example list A would be a list of all the genes investigated on the microarray chip and the

other gene list, list B, could be the differentially expressed genes between cases and controls in decidual tissue.

- Mutually Exclusive Gene Lists: Two gene lists, A and B, are compared, and there are no genes that are on both lists, e.g. A is a list of genes associated with up-regulation and B is a list of genes associated with down-regulation. In the pre-eclampsia example list A could be the list of differentially expressed genes that are up-regulated between the cases and the controls in decidual tissue, while list B contains the genes that are down-regulated between the cases and the controls in decidual tissue. This situation is illustrated in the middle column of Figure 3.
- Intersecting Gene Lists: Two gene lists, A and B, are compared, and there exist genes that are on both lists, e.g. A is a list of genes associated with treatment A, and B is a list of genes associated with treatment B, see the right column of Figure 3. Considering our pre-eclampsia example, list A could be the differentially expressed genes between cases and controls in the decidual tissue while list B would be the differentially expressed genes between cases and controls in the placental tissue.

3.1 Testing Independent Binomial Proportions

The first two situations presented in Figure 3, "One Gene List is a Subset of the Other List" and "Mutually Exclusive Gene Lists" have one important aspect in common; testing the null hypothesis in (1) is equivalent to testing the following hypothesis

$$H_0 : P(G|A \cap B^*) = P(G|B) \text{ vs. } H_1 : P(G|A \cap B^*) \neq P(G|B) \tag{2}$$

The reason for this is as follows:

- When list A and list B are mutually exclusive, the null hypothesis in (1) is clearly equivalent to hypothesis (2), since $A \cap B^* = A$ in this case.
- When list B is a subset of list A, then $A \cap B = B$ and a simple derivation of the equivalence of the null hypothesis in (1) and (2) is given in Appendix B.

This situation can be presented as in Table 1.

	Event	G	$G*$	Total
1	B	$X_{B\cap G}$	$X_{B\cap G^*}$	n_B
2	$A \cap B^*$	$X_{A\cap B^*\cap G}$	$X_{A\cap B^*\cap G^*}$	$n_{A\cap B^*}$
	$A \cup B$	$n_{(A\cup B)\cap G}$	$n_{(A\cup B)\cap G^*}$	$n_{A\cup B}$

TABLE 1: Crosstabulation of events for the situation when one gene list is a subset of the other or the gene lists are mutually exclusive.

Given n_B and $n_{A\cap B^*}$, the two random variables $X_{B\cap G}$ and $X_{A\cap B^*\cap G}$ are independent and binomially distributed:

$$X_{B\cap G} \sim \text{binomial}(n_B, P(G|B))$$
$$X_{A\cap B^*\cap G} \sim \text{binomial}(n_{A\cap B^*}, P(G|A \cap B^*)).$$

In the hypothesis (2) we are testing if two independent binomial proportions are equal. Common approaches are Pearson's asymptotic χ^2-test and Fisher's exact test for large and small samples, respectively, see for example Agresti (2002). Fisher's exact test is, however, conservative, and unconditional tests as well as conditional mid-p tests have been increasingly advocated lately, see for example Hirji (2006).

Fisher's exact test is based on the following observation. Under the null hypothesis (2) $X_{B\cap G}$ is hypergeometric distributed with parameters n_B, $n_{A\cup B}$ and $n_{(A\cup B)\cap G}$. That is,

$$P(X_{B\cap G} = x_{B\cap G}|n_B, n_{A\cup B}, n_{(A\cup B)\cap G}) = \frac{\binom{n_B}{x_{B\cap G}}\binom{n_{A\cap B^*}}{n_{(A\cup B)\cap G}-x_{B\cap G^*}}}{\binom{n_{A\cup B}}{n_{(A\cup B)\cap G}}}$$

In the situation where one of the gene lists is a subset of the other, e.g. $B \subset A$, then $n_{(A\cup B)\cap G} = n_{A\cap G}$ and when the two gene lists are mutually exclusive, $X_{A\cap B^*\cap G} = X_{A\cap G}$, since $A \cap B^* = A$

By conditioning on the fixed marginals n_B, $n_{A\cap B^*}$, $n_{(A\cup B)\cap G}$ and $n_{(A\cup B)\cap G^*}$ we can use Fisher's exact test to calculate the conditional p-value for the hypothesis (2). The p-value, P, is then the sum of hypergeometric probabilities $p(y)$ of all outcomes y of the random variable $X_{B\cap G}$ for all tables with the same marginals n_B, $n_{A\cap B^*}$ and $(X_{A\cap B^*\cap G} + X_{B\cap G})$ with probability less or equal to the observed hypergeometric probability,

$$P = \sum_{p(y)\leq P(x_{B\cap G})} p(y) \tag{3}$$

Instead of the conditional hypergeometric probability, we may use the Pearson's asymptotic χ^2 statistic, see Agresti (1996) Section 2.4.1, to calculate the p-value. Pearson's asymptotic χ^2 statistic is given by

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{n_{A\cup B}(X_{B\cap G}X_{A\cap B^*\cap G^*} - X_{B\cap G}X_{(A\cup B)\cap G})^2}{n_B n_{A\cap B^*} n_{(A\cup B)\cap G} n_{(A\cup B)\cap G^*}}$$

where O_{ij} is the observed frequency and E_{ij} is the expected frequency in cell ij)$i = 1, 2$ and $j = 1, 2$) in Table 1, and the other quantities are also found in Table 1. Under the null hypothesis (2), χ^2 is approximately χ^2_1 distributed. When one gene list is a subset of the other, e.g. $B \subset A$, then

$$\begin{aligned}\chi^2 = &\frac{(X_{B\cap G} - n_B \cdot k_1)^2}{n_B \cdot k_1} + \frac{(X_{A\cap B^*\cap G} - n_{A\cap B^*} \cdot k_1)^2}{n_{A\cap B^*} \cdot k_1} \\ &+ \frac{(X_{B\cap G^*} - n_B \cdot k_2)^2}{n_B \cdot k_2} + \frac{(X_{A\cap B^*\cap G^*} - n_{A\cap B^*} \cdot k_2)^2}{n_{A\cap B^*} \cdot k_2}\end{aligned}$$

where $k_1 = \frac{n_{A\cap G}}{n_A}$ and $k_2 = \frac{n_{A\cap G^*}}{n_A}$. When the gene lists are mutually exclusive, i.e. $A \cap B = \emptyset$ then

$$\begin{aligned}\chi^2 = &\frac{(X_{A\cap G} - n_A \cdot k_1)^2}{n_A \cdot k_1} + \frac{(X_{B\cap G} - n_B \cdot k_1)^2}{n_B \cdot k_1} \\ &+ \frac{(X_{A\cap G^*} - n_A \cdot k_2)^2}{n_A \cdot k_2} + \frac{(X_{B\cap G^*} - n_B \cdot k_2)^2}{n_B \cdot k_2}\end{aligned}$$

where $k_1 = \frac{n_{(A\cup B)\cap G}}{n_{A\cup B}}$ and $k_2 = \frac{n_{(A\cup B)\cap G}}{n_{A\cup B}}$.

The asymptotic p-value is the χ^2_1 right tail probability above the observed value of the Pearson's asymptotic statistic, χ^2_{obs}.

$$P = P(\chi^2_1 \geq \chi^2_{obs}) \tag{4}$$

3.2 Intersecting Gene Lists

For the Intersecting Gene List case the test of the previous sections are not valid. When $B \subset A$ or $A \cap B = \emptyset$ we refer the reader to Section 3.1.

3.2.1 Test Statistic

A natural estimator for $P(G|A)$ is $\frac{X_{A\cap G}/N}{n_A/N} = \frac{X_{A\cap G}}{n_A}$, and similarly for $P(G|B)$ is $\frac{X_{B\cap G}}{n_B}$. We thus start by looking at the statistic

$$D = \frac{X_{A\cap G}}{n_A} - \frac{X_{B\cap G}}{n_B}. \tag{5}$$

This statistic can be written as a linear combination of three random variables $X_{A\cap B\cap G}$, $X_{A\cap B^*\cap G}$ and $X_{A^*\cap B\cap G}$.

$$\begin{aligned} D = \frac{X_{A\cap G}}{n_A} - \frac{X_{B\cap G}}{n_B} &= \frac{X_{A\cap B\cap G} + X_{A\cap B^*\cap G}}{n_A} - \frac{X_{A\cap B\cap G} + X_{A^*\cap B\cap G}}{n_B} \\ &= (\frac{1}{n_A} - \frac{1}{n_B})X_{A\cap B\cap G} + \frac{X_{A\cap B^*\cap G}}{n_A} - \frac{X_{A^*\cap B\cap G}}{n_B} \end{aligned}$$

Given the marginals $n_{A\cap B}$, $n_{A\cap B}$ and $n_{A^*\cap B}$ (see Table 2), the following three random variables are independently and binomially distributed.

$$\begin{aligned} X_{A\cap B\cap G} &\sim \text{binomial}(n_{A\cap B}, P(G|A\cap B)) \\ X_{A\cap B^*\cap G} &\sim \text{binomial}(n_{A\cap B^*}, P(G|A\cap B^*)) \\ X_{A^*\cap B\cap G} &\sim \text{binomial}(n_{A^*\cap B}, P(G|A^*\cap B)) \end{aligned}$$

Since we are conditioning on $n_{A\cap B}$, $n_{A\cap B^*}$ and $n_{A^*\cap B}$, we are also conditioning on the sums $n_A = (n_{A\cap B} + n_{A\cap B^*})$ and $n_B = (n_{A\cap B} + n_{A^*\cap B})$. We may look at $X_{A\cap G} \sim \text{binomial}(n_A, P(G|A))$ and $X_{B\cap G} \sim \text{binomial}(n_B, P(G|B))$, but unfortunately $X_{A\cap G}$ and $X_{B\cap G}$ are not independent.

	Event	G	G^*	Total
1	$A\cap B$	$X_{A\cap B\cap G}$	$X_{A\cap B\cap G^*}$	$n_{A\cap B}$
2	$A\cap B^*$	$X_{A\cap B^*\cap G}$	$X_{A\cap B^*\cap G^*}$	$n_{A\cap B^*}$
3	$A^*\cap B$	$X_{A^*\cap B\cap G}$	$X_{A^*\cap B\cap G^*}$	$n_{A^*\cap B}$
	$A\cup B$	$n_{(A\cup B)\cap G}$	$n_{(A\cup B)\cap G^*}$	$n_{A\cup B}$

Table 2: Crosstabulation of events for intersecting gene lists

The mean and variance of our statistic, D, given in Equation (5), is under the null hypothesis (1) as follows:

$$\begin{aligned}
\mathrm{E}(\frac{X_{A\cap G}}{n_A} - \frac{X_{B\cap G}}{n_B}) &= \frac{n_A \cdot P(G|A)}{n_A} - \frac{n_B \cdot P(G|B)}{n_B} = P(G|A) - P(G|B) = 0 \\
\mathrm{Var}(\frac{X_{A\cap G}}{n_A} - \frac{X_{B\cap G}}{n_B}) &= (\frac{1}{n_A} - \frac{1}{n_B})^2 \mathrm{Var}(X_{A\cap B\cap G}) + \frac{\mathrm{Var}(X_{A\cap B^*\cap G})}{n_A^2} + \frac{\mathrm{Var}(X_{A^*\cap B\cap G})}{n_B^2} \\
&= (\frac{1}{n_A} - \frac{1}{n_B})^2 n_{A\cap B} \cdot P(G|A\cap B) \cdot [1 - P(G|A\cap B)] \\
&+ \frac{n_{A\cap B^*} \cdot P(G|A\cap B^*)[1 - P(G|A\cap B^*)]}{n_A^2} \\
&+ \frac{n_{A^*\cap B} \cdot P(G|A^*\cap B) \cdot [1 - P(G|A^*\cap B)]}{n_B^2}
\end{aligned}$$

Under the null hypothesis, using the central limit theorem,

$$Z_0 = \frac{\frac{X_{A\cap G}}{n_A} - \frac{X_{B\cap G}}{n_B} - 0}{\sqrt{\mathrm{Var}(\frac{X_{A\cap G}}{n_A} - \frac{X_{B\cap G}}{n_B})}} \tag{6}$$

is asymptotically standard normally distributed.

Unfortunately, in the denominator of Z_0 three unknown probabilities are present. To simplify the notation we define o_i, p_i, x_i, and n_i for $i = 1, ..., 3$ as presented in Table 3.

3.2.2 Unpooled Intersecting Asymptotic Test (UIA)

The natural estimators for each of the three probabilities $p_1 = P(G|A\cap B)$, $p_2 = P(G|A\cap B^*)$ and $p_3 = P(G|A^*\cap B)$, are as follows:

$$\hat{p}_1 = \frac{X_{A\cap B\cap G}}{n_{A\cap B}} = \frac{X_1}{n_1} \tag{7}$$

$$\hat{p}_2 = \frac{X_{A\cap B^*\cap G}}{n_{A\cap B^*}} = \frac{X_2}{n_2} \tag{8}$$

$$\hat{p}_3 = \frac{X_{A^*\cap B\cap G}}{n_{A^*\cap B}} = \frac{X_3}{n_3} \tag{9}$$

and the statistic D can then be written

$$D = \frac{X_1 + X_2}{n_1 + n_2} - \frac{X_1 + X_3}{n_1 + n_3}. \tag{10}$$

$A\cap B$	$A\cap B^*$	$A^*\cap B$
$o_1 = P(A\cap B)$	$o_2 = P(A\cap B^*)$	$o_3 = P(A^*\cap B)$
$p_1 = P(G\|A\cap B)$	$p_2 = P(G\|A\cap B^*)$	$p_3 = P(G\|A^*\cap B)$
$X_1 = X_{A\cap B\cap G}$	$X_2 = X_{A\cap B^*\cap G}$	$X_3 = X_{A^*\cap B\cap G}$
$n_1 = n_{A\cap B}$	$n_2 = n_{A\cap B^*}$	$n_3 = n_{A^*\cap B}$

Table 3: Definition of o_i, p_i, x_i, and n_i for $i = 1, ..., 3$, used to simplify the presentation.

Define $\widehat{\text{Var}}(D)$ to be $\text{Var}(D)$ inserted the estimators $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$. We define the test statistic Z_U from Equation (6) inserted $\widehat{\text{Var}}(D)$.

$$\begin{aligned} Z_U &= \frac{D-0}{\sqrt{\widehat{\text{Var}}(D)}} \\ &= \frac{\frac{X_1+X_2}{n_1+n_2} - \frac{X_1+X_3}{n_1+n_3}}{(\frac{1}{n_A}-\frac{1}{n_B})^2 n_1 \cdot \hat{p}_1 \cdot (1-\hat{p}_1) + \frac{1}{n_A^2} \cdot n_2 \cdot \hat{p}_2 \cdot (1-\hat{p}_2) + \frac{1}{n_B^2} \cdot n_3 \cdot \hat{p}_3 \cdot (1-\hat{p}_3)} \end{aligned} \tag{11}$$

For moderate to large samples sizes we expect Z_U to be approximately asymptotically standard normally distributed under the null hypothesis, and calculate the p-value of the test based on Z_U as

$$2 \cdot \Phi(-|Z_U|) \tag{12}$$

where Φ is the cumulative standard normal distribution.

3.2.3 POOLED INTERSECTING ASYMPTOTIC TEST (PIA)

Under the null hypothesis we have the following constraint on p_1, p_2 and p_3,

$$\begin{aligned} P(G|A) - P(G|B) &= \frac{P(A \cap G)}{P(A)} - \frac{P(B \cap G)}{P(B)} \\ &= \frac{P(A \cap B \cap G) + P(A \cap B^* \cap G)}{P(A)} - \frac{P(A \cap B \cap G) + P(A^* \cap B \cap G)}{P(B)} \\ &= \frac{p_1 \cdot P(A \cap B) + p_2 \cdot P(A \cap B^*)}{P(A)} - \frac{p_1 \cdot P(A \cap B) + p_3 \cdot P(A^* \cap B}{P(B)} \\ &= (\frac{1}{P(A)} - \frac{1}{P(B)}) \cdot P(A \cap B) \cdot p_1 + \frac{P(A \cap B^*)}{P(A)} \cdot p_2 - \frac{P(A^* \cap B)}{P(B)} \cdot p_3 \end{aligned}$$

To simplify the notation we define weights w_1 and w_3

$$w_1 = (\frac{P(A)}{P(B)} - 1) \cdot \frac{P(A \cap B)}{P(A \cap B^*)} \tag{13}$$

$$w_3 = \frac{P(A)}{P(B)} \cdot \frac{P(A^* \cap B)}{P(A \cap B^*)} \tag{14}$$

and using $P(G|A) - P(G|B) = 0$, we find the following formula for p_2.

$$p_2 = w_1 \cdot p_1 + w_3 \cdot p_3 \tag{15}$$

We may use the maximum likelihood method for estimating p_1, p_2, p_3 under the constraint (15).

The likelihood $L(p_1, p_3; x_1, x_2, x_3, n_1, n_2, n_3, P(A), P(B), P(A \cap B), P(A \cap B^*), P(A^* \cap B))$ is given as the product of three conditionally independent binomial probabilities, inserted the expression (15) for p_2. The log likelihood can be written as follows,

$$\begin{aligned} l(p_1, p_3) &= x_1 \cdot \log(p_1) + (n_1 - x_1) \cdot \log(1-p_1) \\ &+ x_2 \cdot \log(w_1 \cdot p_1 + w_3 \cdot p_3) + (n_2 - x_2) \cdot \log(1 - w_1 \cdot p_1 - w_3 \cdot p_3) \\ &+ x_3 \cdot \log(p_3) + (n_3 - x_3) \cdot \log(1-p_3) \end{aligned} \tag{16}$$

Unknown probabilies enter into the weights w_1 and w_3, and we use the natural estimators for $P(A)$, $P(B)$, $P(A \cap B)$, $P(A \cap B^*)$, $P(A^* \cap B)$ (given below) inserted into the weights in a numerical optimization of the log likelihood.

$$\begin{aligned}
\tilde{w}_1 &= (\frac{\widehat{P(A)}}{\widehat{P(B)}} - 1) \cdot \frac{\widehat{P(A \cap B)}}{\widehat{P(A \cap B^*)}} \\
&= (\frac{\frac{n_1+n_2}{N}}{\frac{n_1+n_3}{N}} - 1) \cdot \frac{\frac{n_1}{N}}{\frac{n_2}{N}} \\
&= (\frac{n_1+n_2}{n_1+n_3} - 1) \cdot \frac{n_1}{n_2} \\
\tilde{w}_3 &= \frac{\widehat{P(A)}}{\widehat{P(B)}} \cdot \frac{\widehat{P(A^* \cap B)}}{\widehat{P(A \cap B^*)}} \\
&= \frac{\frac{n_1+n_2}{N}}{\frac{n_1+n_3}{N}} \cdot \frac{\frac{n_3}{N}}{\frac{n_2}{N}} \\
&= \frac{n_1+n_2}{n_1+n_3} \cdot \frac{n_3}{n_2}
\end{aligned}$$

When $n_2 = 0$, $A \cup B = A$, i.e. $A \subset B$ and we refer to Section 3.1 and the tests for the situation when one gene list is a subset of the other. Inserting the estimators into the log likelihood and maximizing wrt. p_1 and p_3, we get the estimators $\tilde{p}_1$ and $\tilde{p}_3$. This log likelihood has no simple analytic solution, and we use numerical optimization to maximize the log likelihood. Here $\tilde{p}_1$ and $\tilde{p}_3$ can not be written in closed form, but $\tilde{p}_2$ is found by inserting $\tilde{p}_1$ and $\tilde{p}_3$ in equation (15).

We define $\widetilde{\text{Var}}(D)$ to be $\text{Var}(D)$ inserted the estimators $\tilde{p}_1$, $\widetilde{p}_2$, and $\widetilde{p}_3$, and using $\widetilde{\text{Var}}(D)$ in Equation (6), we define the test statistic Z_P,

$$\begin{aligned}
Z_P &= \frac{D - 0}{\sqrt{\widetilde{\text{Var}}(D)}} \\
&= \frac{\frac{X_1+X_2}{n_1+n_2} - \frac{X_1+X_3}{n_1+n_3}}{(\frac{1}{n_A} - \frac{1}{n_B})^2 \cdot n_1 \cdot \tilde{p}_1 \cdot (1-\tilde{p}_1) + \frac{1}{n_A^2} \cdot n_2 \cdot \tilde{p}_2 \cdot (1-\tilde{p}_2) + \frac{1}{n_B^2} \cdot n_3 \cdot \tilde{p}_3 \cdot (1-\tilde{p}_3)}
\end{aligned} \tag{17}$$

For moderate large sample sizes we expect Z_P to be approximately asymptotically standard normally distributed under the null hypothesis, and calculate the p-value of the test based on Z_P as

$$2 \cdot \Phi(-|Z_P|) \tag{18}$$

where Φ is the cumulative standard normal distribution.

3.2.4 Test of Leisenring, Alonzo and Pepe

A more general approach to testing (1) is the test of Leisering et al. (2000) (which we denote the LAP-test) for comparison of predictive values of two diagnostic tests, tests A and B, with respect to a disease G. Every individual may have 0, 1 or 2 positive tests, i.e. every gene may be on 0, 1 or 2 of the gene lists A and B. The individuals are indexed by i and the (multiple) observations for each individual are indexed by $j = 1, ..., n_i$.

Leisering et al. (2000) use three binary random variables, Y_{ij} that denotes disease status, Z_{ij} that indicates which test was used and X_{ij} that describes the outcome of the diagnostic test.[2]

$$Y_{ij} = \begin{cases} 0, & \text{nondiseased} \\ 1, & \text{diseased} \end{cases}$$

$$Z_{ij} = \begin{cases} 0, & \text{test 1} \\ 1, & \text{test 2} \end{cases}$$

$$X_{ij} = \begin{cases} 0, & \text{negative} \\ 1, & \text{positive} \end{cases}$$

To compare this to our situation, we have $n_i = 2$ results for each gene i, and for now let $j = 1$ be the result for test A and $j = 2$ be the result for test B. Then $Y_{i1} = Y_{i2}$ would denote whether the gene is a member of gene class G or not; $Z_{i1} = 0$ (for all i) since we assume that test A is performed for $j = 1$ and $Z_{i2} = 1$ (for all i) since we assume that test B is performed for $j = 2$. Finally $X_{i1} = 1$ if the gene is on list A and $X_{i2} = 1$ if the gene is on list B.

Leisering et al. (2000) aim at comparing the positive predictive values (PPV) for the two tests, where the positive predicted value is defi ned as P(disease | positive test).

The PPV for test 1 can be represented as $PPV_1 = P(Y_{ij} = 1|Z_{ij} = 0, X_{ij} = 1)$ and the PPV for test 3 as $PPV_3 = P(Y_{ij} = 1|Z_{ij} = 1, X_{ij} = 1)$. The null hypothesis tested is $H_0 : PPV_2 = PPV_2$, i.e. the probability that an individual is diseased given that test 1 is positive is equal to the probability that an individual is diseased given that test 2 is positive. In our setting this means that we are testing whether the probability that a gene is a member of gene class G given that it is on list A is equal to the probability that the gene is a member of gene class G given that it is on list B. Thus, we have the same null hypothesis as in (1).

Base on generalized estimation equations Leisering et al. (2000) fi t a generalized linear model and defi ne the following test statistic for large samples;

$$T_{PPV} = \frac{\left\{\sum_{i=1}^{N_p}\sum_{j=1}^{m_i} Y_{ij}(Z_{ij} - \bar{Z})\right\}^2}{\sum_{i=1}^{N_p}\left\{\sum_{j=1}^{m_i}(Y_{ij} - \bar{Y})(Z_{ij} - \bar{Z})\right\}^2} \tag{19}$$

where

$$\bar{Z} = \frac{\sum_{i=1}^{N_p} m_i Z_i Y_i}{\sum_{i=1}^{N_p} m_i}$$

which is the proportion of positive test 2's among all the tests and

$$\bar{Y} = \frac{\sum_{i=1}^{N_P} m_i Y_i}{\sum_{i=1}^{N_P} m_i}$$

is the proportion of diseased individuals among all the individuals. Further, m_i is the number of positive test results for individual i, N_P is the number of individuals with at least one positive test outcome.

[2] In Leisering et al. (2000) the letter D is used instead of Y, but to avoid confusion with our previous notation we have chosen to use Y.

In our situation

$$\bar{Z} = \frac{n_B}{n_A + n_B}$$

i.e. the proportion of genes on list B among all the genes, and

$$\bar{Y} = \frac{X_{A\cap G} + X_{B\cap G}}{n_A + n_B},$$

the proportion of the genes that are members of gene class G. m_i is the number of lists of which gene i is present, and N_P is the number of genes that are on at least one of the lists.

The test statistic in (19) is general and can be used even if it is possible that a gene may be in the gene class G for some tests and not for other tests. However, in our case this is not possible, a particular gene is either always in the gene class G or never. Then the test statistic can be simplified.

By defining $T_i = \sum_{j=1}^{m_i} Z_{ij}$, the number of positive test 2's individual i contributes to the analysis, the statistic can be written

$$T_{PPV} = \frac{\left\{\sum_{i=1}^{N_p} Y_i(T_i - m_i\bar{Z})\right\}^2}{\sum_{i=1}^{N_p}(Y_i - \bar{Y})^2(T_i - m_i\bar{Z})^2}$$

In our situation, T_i is the number of lists at which the genes on list B are present. The numerator is

$$(X_1 + X_3 - \frac{2X_1 + X_2 + X_3}{2n_1 + n_2 + n_3}(n_1 + n_3))^2$$

which we interpret as the number of genes on list B that are in gene class G subtracted an estimate for the probability of being in gene class G for all genes on the lists A and B multiplied by the number of genes on list B.

If we rearrange the terms in the numerator and cancel out the common factors in the numerator and denominator, the test statistic for large samples is given by

$$T_{PPV} = \frac{((n_1 + n_2)(X_1 + X_3) - (n_1 + n_3)(X_1 + X_2))^2}{f(X_1, X_2, X_3, n_1, n_2, n_3)} \tag{20}$$

where

$$
\begin{aligned}
&f(X_1, X_2, X_3, n_1, n_2, n_3) = \\
&X_1(n_2 - n_3)^2 \left(1 - \frac{2X_1 + X_2 + X_3}{2n_1 + n_2 + n_3}\right)^2 \\
&+X_2(n_1 + n_3)^2 \left(1 - \frac{2X_1 + X_2 + X_3}{2n_1 + n_2 + n_3}\right)^2 \\
&+X_3(n_1 + n_2)^2 \left(1 - \frac{2X_1 + X_2 + X_3}{2n_1 + n_2 + n_3}\right)^2 \\
&+(n_1 - X_1)(n_2 - n_3)^2 \left(\frac{2X_1 + X_2 + X_3}{2n_1 + n_2 + n_3}\right)^2 \\
&+(n_2 - X_2)(n_1 + n_3)^2 \left(\frac{2X_1 + X_2 + X_3}{2n_1 + n_2 + n_3}\right)^2 \\
&+(n_3 - X_3)(n_1 + n_2)^2 \left(\frac{2X_1 + X_2 + X_3}{2n_1 + n_2 + n_3}\right)^2
\end{aligned}
$$

We see that the numerator in (20) is somewhat similar to our D statistic given in 5. If we multiply D by $(n_1 + n_2)(n_1 + n_3)$ and then take the square, we obtain the numerator in (20).

The p-value of the LAP-test is calculated as:

$$P(\chi_1^2 \geq T_{PPV}) = 1 - F_{\chi_1^2}(T_{PPV}) \tag{21}$$

where $F_{\chi_1^2}$ is the cumulative χ^2 distribution with 1 degree of freedom.

3.2.5 TESTS BASED ON $A \cap B = \emptyset$

When one list is not a subset of the other list, the tests outlined in Section 3.1 is valid when $A \cap B = \emptyset$. When $A \cap B \neq \emptyset$, but the size of $A \cap B$ is small, we may regard these tests as approximate strategies. We will investigate the following two strategies further in the simulation study in Section 4.

- Genes in $A \cap B$ are deleted both from list A and list B, resulting in $A \cap B^*$ and $A^* \cap B$ taking the place of A and B in the test of Section 3.1. Fisher's exact test or Pearson's asymptotic χ^2 test is then used for the case of mutually exclusive gene lists. We call this strategy "Delete $A \cap B$". This strategy is to our knowledge used in the GO-tool FatiGO Al-Shahrour, Diaz-Uriarte and Dopazo (2004), implemented using Fisher's exact test.
- Another possible approach is to simply ignore the fact that there are genes that are on both lists and use Fisher's exact test or the Pearson's asymptotic χ^2 test in Section 3.1 as if the lists were mutually exclusive. We call this strategy "Ignore $A \cap B$".

4 SIMULATION STUDY

All analyses are performed using the R language, R Development Core Team (2005).

4.1 METHODS

In Section 3 we presented two new tests, the unpooled intersecting asymptotic test (UIA) and the pooled intersecting asymptotic test (PIA), for the situation of intersecting gene lists. In addition we adapted the Leisering et al. (2000) (LAP) test for comparing two positive predictive values to the case of two intersecting gene lists. To compare the performance of the three tests (UIA, PIA and LAP) we have conducted a simulation study. For comparison, we have also included both small sample and asymptotic versions of the tests (in the intersecting gene lists situation) presented in Section 3.2.5. The methods under study are listed in Table 4.

4.2 CASES

In the simulation study we have investigated 10 situations designed to be under the null hypothesis (1) and 9 situations under specifi c alternative hypotheses. These 19 cases are presented in Table 5.

For each situation six parameter values are set. These are

- $o_1 = P(A \cap B)$, $o_2 = P(A \cap B^*)$, and $o_3 = P(A^* \cap B)$,
- $p_1 = P(G|A \cap B)$, $p_2 = P(G|A \cap B^*)$, and $p_3 = P(G|A^* \cap B)$.

Here the parameters o_1, o_2, o_3 are the probabilies that a randomly selected gene is a member of the list $(A \cap B)$, $(A \cap B^*)$ or $(A^* \cap B)$, respectively. We refer to these probabilities as o-probabilites. And p_1, p_2, p_3 are the probabilies that a randomly selected gene is a member of gene class G, given that the gene is a member of $(A \cap B)$, $(A \cap B^*)$ and $(A^* \cap B)$, respectively. We refer to these probabilites as p-probabilites.

Each situation is labelled using a code with four slots:

- o-probabilities,
- p-probabilites,
- expected length of A and B list,
- expected values of X_1, X_2, X_3.

Test	Distribution	Section	P-value
Unpooled intersecting asymptotic test (UIA)	Asymptotic N	3.2.2	(12)
Pooled intersecting asymptotic test (PIA)	Asymptotic N	3.2.3	(18)
Leisenring, Alonzo, Pepe test (LAP)	Asymptotic χ^2_1	3.2.4	(21)
Delete $A \cap B$ test	Fisher, hypergeometric	3.2.5	(3)
Delete $A \cap B$ test	Asymptotic χ^2_1	3.2.5	(4)
Ignore $A \cap B$ test	Fisher, hypergeometric	3.2.5	(3)
Ignore $A \cap B$ test	Asymptotic χ^2_1	3.2.5	(4)

TABLE 4: Overview of the hypothesis testing strategies applied in the simulation study of Section 4.

Situations under the null hypothesis $P(G\|A) = P(G\|B)$											
No.	Code	o_1	o_2	o_3	p_1	p_2	p_3	$P(A)$	$P(B)$	$P(G\|A)$	$P(G\|B)$
1	e.e.600.30	0.010	0.010	0.010	0.100	0.100	0.100	0.020	0.020	0.100	0.100
2	e.e.300.15	0.005	0.005	0.005	0.100	0.100	0.100	0.010	0.010	0.100	0.100
3	e.e.60.3	0.001	0.001	0.001	0.100	0.100	0.100	0.002	0.002	0.100	0.100
4	b.e.600.45a135	0.005	0.015	0.015	0.300	0.300	0.300	0.020	0.020	0.300	0.300
5	b.e.600.15a45	0.005	0.015	0.015	0.100	0.100	0.100	0.020	0.020	0.100	0.100
6	b.b.600.45	0.005	0.015	0.015	0.300	0.100	0.100	0.020	0.020	0.150	0.150
7	b.b.600.75a27	0.005	0.015	0.015	0.500	0.060	0.060	0.020	0.020	0.170	0.170
8	u.b.10a30.1.8a1.2a7.2	0.00013	0.0002	0.001	0.450	0.200	0.277	0.0003	0.001	0.300	0.300
9	u.b.50a150.9a6a36	0.00067	0.0010	0.0043	0.450	0.200	0.277	0.002	0.005	0.300	0.300
10	u.b.250a750.45a30a180	0.0033	0.005	0.022	0.450	0.200	0.277	0.008	0.025	0.300	0.300
Situations under the alternative hypothesis $P(G\|A) \neq P(G\|B)$											
No.	Code	o_1	o_2	o_3	p_1	p_2	p_3	$P(A)$	$P(B)$	$P(G\|A)$	$P(G\|B)$
11	e.u.300.15a22.5	0.005	0.005	0.005	0.100	0.100	0.150	0.010	0.010	0.100	0.125
12	e.u.300.15a30	0.005	0.005	0.005	0.100	0.100	0.200	0.010	0.010	0.100	0.150
13	e.u.300.15a45	0.005	0.005	0.005	0.100	0.100	0.300	0.010	0.010	0.100	0.200
14	b.u.600.15a45a67.5	0.005	0.015	0.015	0.100	0.100	0.150	0.020	0.020	0.100	0.138
15	b.u.600.15a45a135	0.005	0.015	0.015	0.100	0.100	0.300	0.020	0.020	0.100	0.250
16	b.u.600.75a27a36	0.005	0.015	0.015	0.500	0.060	0.080	0.020	0.020	0.170	0.185
17	b.u.600.75a27a54	0.005	0.015	0.015	0.500	0.060	0.120	0.020	0.020	0.170	0.215
18	u.u.50a150.9a6a40	0.0007	0.001	0.0043	0.450	0.200	0.3077	0.0017	0.005	0.300	0.327
19	u.u.50a150.9a6a50	0.0007	0.001	0.0043	0.450	0.200	0.3846	0.0017	0.005	0.300	0.393

TABLE 5: Situations (cases) investigated under null hypothesis $P(G|A) = P(G|B)$ and the alternative hypothesis $P(G|A) \neq P(G|B)$.

The following shorthand-notation is used:

1. The first slot gives a letter describing the relationship between the probabilities o_1, o_2 and o_3, and the letter "e" denotes "equality", "b" denotes "balanced", i.e. $P(A) = o_1 + o_2 = P(B) = o_1 + o_3$, and "u" denotes "unbalanced".

2. The second slot gives a letter describing the relationship between the probabilities p_1, p_2 and p_3, and the letter "e" denotes "equality", and "b" denotes "balanced", i.e. $P(G|A) = P(G|B)$, and "u" denotes "unbalanced".

3. The third slot contains number(s) denoting the expected length of the A and B list (if the two numbers are not equal, they are separated by the letter "a").

4. The fourth slot contains number(s) denoting the the expected values of X_1, X_2, X_3.

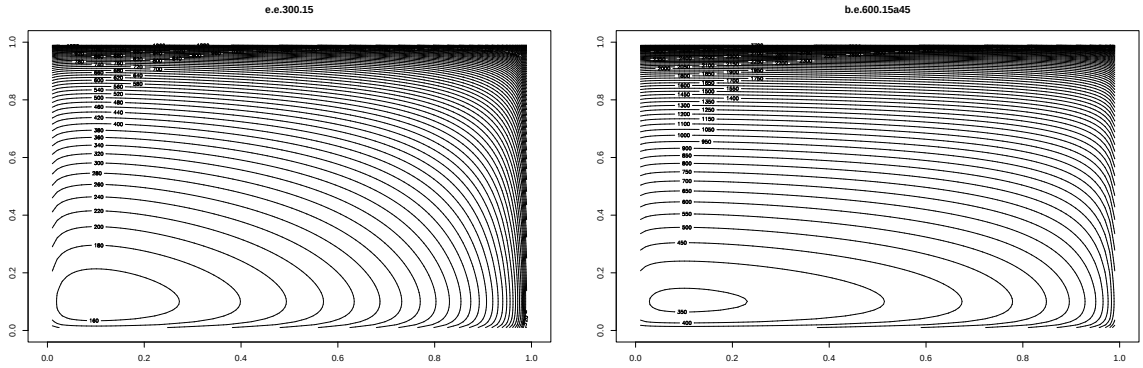
For the simulated situations under the null hypothesis there are four groups of combinations: e.e (equal o and p probabilities), b.e (balanced o probabilites and equal p probabilites), b.b (balanced o and p probabilites), u.b (unequal o-probabilites and balanced p probabilities). For the three first groups the expected length of the A and B lists are equal. For the simulated situations under the alternative hypothesis there are four groups of combinations: e.u (equal o probabilites and unequal p probabilities), b.u (balanced o probabilites and unequal p probabilites), b.u (balanced o probabilites and unequal p probabilites), u.u (unequal o and p probabilities). These situations are constructed by starting with one situation under the null hypothesis and changing the probability p_3.

In the first three situations under the null hypothesis (Case 1-3:"e.e") in Table 5, both the expected sizes of the subsets $(A \cap B)$, $(A \cap B^*)$ and $(A^* \cap B)$ are equal, and the expected sizes of the subsets $(A \cap B \cap G)$, $(A \cap B^* \cap G)$ and $(A^* \cap B \cap G)$ are equal. For Case 2 a contour plot of the log likelihood (16) inserted the expected values of n_i and X_i for $i = 1, ..., 3$, as functions of $p_1 = P(G|A \cap B)$ on the horizonal axis and of $p_3 = P(G|A \cap B^*)$ on the vertical axis, is found in the upper left panel of Figure 4. Cases 11-13 under the alternative hypothesis, are constructed from Case 2.

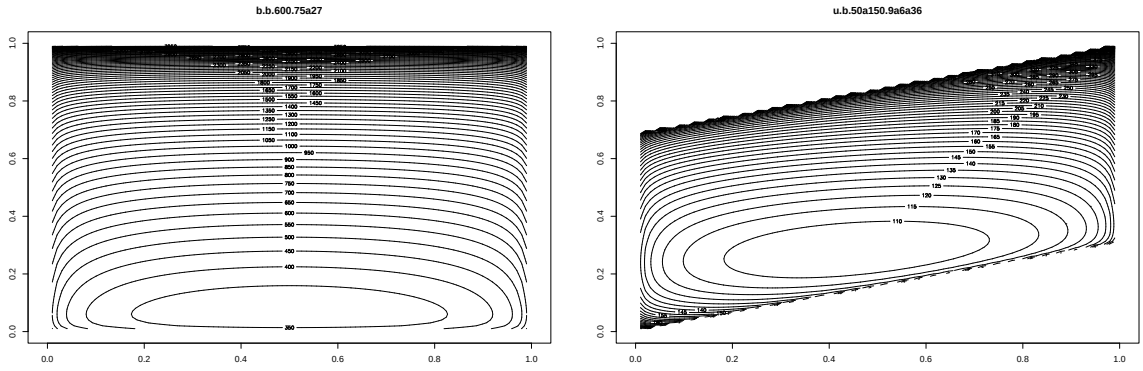
For the next two situations (Case 4-5:"b.e") in Table 5, the expected sizes of the subset $(A \cap B)$, $(A \cap B^*)$ and $(A^* \cap B)$ are not equal, but the expected sizes of the subsets A and B are equal. The expected sizes of the subsets $(A \cap B \cap G)$, $(A \cap B^* \cap G)$ and $(A^* \cap B \cap G)$ are equal. For Case 5 a contour plot of the log likelihood is found in the upper right panel of Figure 4. Cases 13-14 under the alternative hypothesis, are constructed from Case 5.

For the next two situations (Case 6-7:"b.b") in Table 5, the expected sizes of the subsets $(A \cap B \cap G)$, $(A \cap B^* \cap G)$ and $(A^* \cap B \cap G)$ are not equal, but the expected size of the subsets A and B are equal. The expected sizes of the subsets $(A \cap B \cap G)$, $(A \cap B^* \cap G)$ and $(A^* \cap B \cap G)$ are not equal, but the expected probabilities $P(G|A)$ and $P(G|B)$ are equal. For Case 7 a contour plot of the log likelihood is found in the lower left panel of Figure 4. Cases 15-16 under the alternative hypothesis, are constructed from Case 7.

In the last three situations (Case 8-10:"u.b") in Table 5, the expected sizes of the subsets A and B are not equal. For Case 9 a contour plot of the log likelihood is found in the lower right panel of Figure 4. Cases 17-18 under the alternative hypothesis, are constructed from Case 9.



Case 2 Case 5



Case 7 Case 9

FIGURE 4: Contour plot of the log likelihood for cases 2, 5, 7 and 9 inserted the expected values of n_i and X_i for $i = 1, ..., 3$, as functions of $p_1 = P(G|A \cap B)$ on the horizonal axis and of $p_3 = P(G|A \cap B^*)$ on the vertical axis.

4.3 SIMULATION ALGORITHM

The simulation study is devided into separate series of *unconditional* and *conditional* simulations. The two strategies differ only in how the lengths of the list $(A \cap B)$, $(A \cap B^*)$ and $(A^* \cap B)$ are generated. In the unconditional sampling these numbers are drawn from the multinomial distribution, while for the conditional simulations we use the expected values of the lengths of these lists. The following algorithm, with a unconditional or conditional choice at step 2, is used to generate one simulated data set.

1. $N = 30000$ genes are under study.

2. The number of genes on lists $(A \cap B)$, $(A \cap B^*)$ and $(A^* \cap B)$ are called $n_1 = n_{(A \cap B)}$, $n_2 = n_{(A \cap B^*)}$, and $n_3 = n_{(A^* \cap B)}$.

   UNCONDITIONAL: Let $n_4 = n_{(A^* \cap B^*)}$ denote the number of genes on the list $(A^* \cap B^*)$. The numbers n_1, n_2, n_3 and n_4 are drawn randomly using the multinomial distribution and the o-probabilities.

   $$(n_1, n_2, n_3, n_4) \sim \text{multinomial}(N, o_1, o_2, o_3, (1 - o_1 - o_2 - o_3))$$

   The value n_4 is not used further.

   CONDITIONAL: The values of n_1, n_2, and n_3 are set to their expected values in the multinomial model.

   $$n_i = o_i \cdot N \text{ for } i = 1, 2, 3$$

3. Given n_1, n_2, and n_3, the number of genes that are members of a given gene class G, are called X_1, X_2, and X_3, and are drawn using three independent binomial models and the p-probabilites.

   $$\begin{aligned} X_1 &\sim \text{binomial}(n_1, p_1) \\ X_2 &\sim \text{binomial}(n_2, p_2) \\ X_3 &\sim \text{binomial}(n_3, p_3) \end{aligned}$$

4. P-values for each of the methods in Table 4 are calculated based on $(X_1, n_1, X_2, n_2, X_3, n_3)$.

This algorithm is repeated $M = 30000$ times, to produce M p-values for both the conditional and the unconditional simulation strategies, for each situation in Table 5, and each statistical test presented in Table 4.

4.4 EVALUATION STRATEGIES

The test situations under the null hypothesis will be used to study if each test preserves the test size, i.e. that a specifi ed nominal signifi cance level equals the observed actual signifi cance level. We will look closely at signifi cance levels within the interval $[0.001, 0.1]$ and in particular to the values {0.001, 0.005, 0.01,0.05, 0.1}. The smallest values are motivated by the possible subsequent use of methods for controlling multiple testing error rates.

For each stategy, situation, test and selected value of the nominal signifi cance level α, let W be a random variable counting the number of p-values smaller than or equal to α. Then W is binomially

distributed with size M and probability α. An estimate of the significance level of the test, $\hat{\alpha}$ and a $100 \cdot (1-\gamma)\%$ confidence interval with limits $\hat{\alpha}_L$ and $\hat{\alpha}_U$, are given as, Agresti and Coull (1998).

$$\begin{aligned} \hat{\alpha} &= \frac{W}{M} \qquad (22) \\ \tilde{W} &= W + 2 \\ \tilde{M} &= M + 4 \\ \tilde{\alpha} &= \frac{\tilde{W}}{\tilde{M}} \\ \hat{\alpha}_L &= \tilde{\alpha} - z_{\frac{\gamma}{2}} \sqrt{\frac{\tilde{\alpha} \cdot (1-\tilde{\alpha})}{\tilde{M}}} \qquad (23) \\ \hat{\alpha}_U &= \tilde{\alpha} + z_{\frac{\gamma}{2}} \sqrt{\frac{\tilde{\alpha} \cdot (1-\tilde{\alpha})}{\tilde{M}}} \qquad (24) \end{aligned}$$

The alternative hypothesis situations will be used to assess the power of the tests. Then $\hat{\alpha}$ in Equation (22) is an estimate for the power of the test, i.e. $P(\text{reject } H_0 | H_1)$, at the specific value under H_1 defined by the situation under study.

We also look at the distribution of the calculated p-values for the continuous tests (not the two discrete tests). Under the null hypothesis this distribution (when the p-values are continuous) should be uniform. The distribution of the p-values under the alternative hypothesis has not been investigated theoretically for the different methods, but we expect that the distribution is skewed to the left and that most p-values are small.

4.5 RESULTS

For each situation and each method the following is investigated:

- estimated significance level (or power) for nominal levels in within the interval $[0.001, 0.1]$ and in particular at the values {0.001, 0.005, 0.01,0.05, 0.1},
- 95% confidence intervals for the significance level (or power) for the above nominal levels, and
- distribution of p-values.

The results from the simulation study are presented in Figures 6-9 and Tables 7-10 in Appendix C. A colour coded figure legend for the results from the simuation study is presented in Figure 5.

We summarize the results for the simuations under the null and alternative hypothesis:

- In the first three situations under the null hypothesis (Case 1-3:"e.e") the general finding is that in the conditional simulations the three intersecting tests UIA, PIA and LAP all perform satisfactory (correct test size, approximate uniform distribution of p-values), but that the tests based on $A \cap B = \emptyset$ are conservative (i.e. estimated significance level below the nominal level). The "Delete $A \cap B$" test using the Fisher-p-values is conservative, but using the Pearson's

asymptotic χ^2-p-value the test performs satisfactory. The "Ignore $A \cap B$" is by far the most conservative.

For the unconditional simulations the results are the same as for the conditional simulations for Cases 1-2, but for Case 3 (where the expected value for X_1, X_2 and X_3 is 3 and thus very small), the UIA test has higher observed significance level than the nominal level for significance levels above 0.02. The PIA, LAP and "Delete $A \cap B$" Pearson's asymptotic χ^2 tests have observed significance level below the nominal level for the interval $[0, 0.06]$ and observed significance level above the nominal level for levels above 0.06. Results for Case 2 (together with Case 11) are found in Table 7 and Figure 6 in Appendix C.

- The first three situations under the alternative hypothesis (Case 11-13:"e.u") are slightly modified versions of Case 2, where the probability of p_3 is changed. The general finding is that the UIA-test has the highest power, followed by the PIA-test and the "Delete $A \cap B$" Pearson's asymptotic χ^2 test, and then the LAP-test. The 'Ignore $A \cap B$' test has low power. For the conditional simulations the PIA, LAP and "Delete $A \cap B$" Pearson's asymptotic χ^2 tests have nearly identical power. Results for Case 11 (together with Case 2) are found in Table 7 and Figure 6 in Appendix C.

- For the Cases 4-5 ("b.e") the results from the conditional and unconditional situations show the same conclusions and are nearly numerically identical. Under the null hypothesis the three intersecting tests UIA, PIA and LAP all performs satisfactory (correct test size, approximate uniform distribution of p-values), and so does the "Delete $A \cap B$" test using the Pearson's asymptotic χ^2-p-value. The "Delete $A \cap B$" test using the Fisher-p-value is conservative, and the "Ignore $A \cap B$" is very conservative. Results for Case 5 (together with Case 14) are found in Table 8 and Figure 7 in Appendix C.

- The two situations in Case 14-15 ("b.u") are modifications of Case 5 (the probability of p_3 is changed). The findings are the same as for the Cases 11-13. Results for Case 14 (together with Case 5) are found in Table 8 and Figure 7 in Appendix C.

- For the Cases 6-7 ("b.b") under the null hypothesis the results differ for the unconditional and conditional simulations. Case 7 involves smaller sample sizes than Case 6, and here the results are more extreme than for Case 6. For the UIA and PIA tests, the observed significance levels are significantly higher than the nominal level in the unconditional simulations, and the LAP and the "Delete $A \cap B$" test using the Pearson's asymptotic χ^2-p-value have correct level. For the conditional simulations the UIA, PIA and "Delete $A \cap B$" Pearson's asymptotic χ^2-test have correct level, while the LAP-test has level below the observed level. Again the "Delete $A \cap B$" Fisher test is conservative and the "Ignore $A \cap B$" test is very conservative. Results for Case 7 (together with Case 16) are found in Table 9 and Figure 8 in Appendix C.

- The Cases 16 and 17 ("b.u") are modifications of Case 7. Since UIA and PIA have observed significance level slightly higher than the nominal level for the unconditional simulations in Case 7, they also have the highest power. Results for Case 16 (together with Case 7) are found in Table 9 and Figure 8 in Appendix C.

- For Case 8 ("u.b") all the intersecting tests have observed level slightly larger than nominal level, with PIA as the least extreme. The conditional simulations are much less extreme than the unconditional simulations. The "Delete $A \cap B$" and "Ignore $A \cap B$" tests are conservative.

Cases 9-10 ("u.b") have larger sample sizes than Case 8, and the (expected) lengths of lists A and B are very different (50 vs. 150 for Case 9 and 250 vs. 750 for Case 10). The "Ignore $A \cap B$" tests are conservative, while the "Delete $A \cap B$" tests do not preserve test size. The observed significance levels for the "Delete $A \cap B$" tests are very much higher than the nominal levels (for example for nominal level 0.01 the observed level is 0.02 in Case 9 and 0.24 in Case 10, and for nominal level 0.1 the observed level is 0.2 in Case 9 and 0.62 in Case 10 for the Pearson asymptotic test version of the unconditional simulations). For the intersecting tests the observed levels are slightly larger than the nominal levels for the unconditional simulations with UIA and PIA, while the LAP test preserves the test size. For the conditional simulations the UIA and PIA have still slightly higher observed than nominal level (not significant for PIA for Case 9 and not significant for UIA and PIA in Case 10). The LAP test has significantly lower observed level than nominal level for Case 9 and 10. Results for Case 9 (together with Case 18) are found in Table 10 and Figure 9 in Appendix C.

- Cases 18-19 ("u.u") are modifications of Case 9. Since the "Delete $A \cap B$" tests did not preserve the test size these tests also have by far the highest power. Results for Case 18 (together with Case 9) are found in Table 10 and Figure 9 in Appendix C.

4.6 Conclusions from the simulation study

For Cases 1,2,4 and 5 (equal p-probabilites) all the intersecting tests (UIA, PIA and LAP) were found to preserve the test size for both the conditional and unconditional simulations. For Case 3 (equal p-probabilites, but very small expected lengths of X_1, X_2, X_3) all intersecting tests performed satisfactory in the conditional simulations, but for the unconditional simulations none of the intersecting tests preseved the test size. For Case 6 (balanced o and p-probabilites, but also equal expected length of X_1, X_2 and X_3) the LAP test did not perform satisfactory for the conditional case, while the UIA and PIA test did not perform satisfactory for the unconditional case. For Cases 7-10 (either unbalanced o-probabilites and/or very different expected value of X_1, X_2 and X_3) the general finding is that only the PIA test preserves the test size for all the conditional simulations, while only the LAP test preserves the test size for all the unconditional simulations (except for Case 8 where none of the intersecting tests preserves the test size in the unconditional simulations).

Thus, for the intersecting test the LAP test gave the best results for the unconditional simulations and the PIA test gave the best results for the conditional simulations. This is not surprising, since the PIA test is developed conditionally on observed values of n_1, n_2 and n_3, while the LAP test is based on an unconditional model. We may therefore conclude that we suggest that the PIA test should used when the sampling strategy is conditional and the LAP test should be used when the sampling strategy is unconditional.

For the case of two intersecting gene lists, we would assume that both sampling strategies could be used. If e.g. lists of statistically significantly differentially expressed genes were based on a cut-off on p-values or adjusted p-values using multiple testing criterias, we would assume that the sampling strategy is unconditional. If instead "top 100" or "top 500" gene lists are produced (that is, cut-off not dependent on p-values but instead of a desired length of the gene lists) the sampling would be regarded as conditional.

For the tests based on $A \cap B = \emptyset$ the "Ignore $A \cap B$"-test did not preserve the test size in any of the case tested (conservative) and had also very low power. This test should not be used in for intersecting gene lists unless the intersection is close to the empty set. The "Delete $A \cap B$" Pearson's asymptotic

χ^2 test was found to perform satisfactory when the expected lengths of the lists A and B were equal, but did not preserve the test size when the expected lengths of the lists A and B were different. Thus, the "Delete $A \cap B$" Pearson's asymptotic χ^2 test may be used with caution only when the expected lengths of the lists A and B are equal.

For the tests based on $A \cap B = \emptyset$ we found that there were a substantial difference between using the Fisher's exact test and the Pearson asymptotic χ^2 test. For test statistics having a continuous distribution the distribution of the p-values under the null hypothesis is uniform. The expected value for the p-values are thus 0.5. But, for test statistics having discrete distribution the distribution of the p-values under the null hypothesis are not uniform and the expected value for the p-values are greater 0.5, Agresti (1996), p 43. For p-values that are based on discrete test statistics the average p-value under the null hypothesis tends to be too large. Comparing the average of the p-values for the "Delete $A \cap B$" Fisher test and the "Delete $A \cap B$" Pearson's asymptotic χ^2 test this was observed also in our simulation study. However, the proportion of p-values below nominal signifi cance levels presented in the Tables 7-10 were found to be smaller for Fisher's exact test than for the Pearson's asymptotic χ^2 test.

5 Conclusions and Further Work

In Allison, Cui, Grier and Sabripour (2006) one of the important "consensus points" presented within statistical inference is that gene class testing is desirable, and has become a popular and widely accepted analytical tool. However, one important problem they found with gene class testing is that the null and alternative hypotheses often are not formally defi ned or poorly defi ned. By formally stating the null and alternative hypotheses we have provided the researcher with an adequate and statistically valid starting point for their analyses.

The focus of this report has been on statistical hypothesis testing of association between two gene lists for a given class of genes. In the eGOn-tool, Beisvåg et al. (2006) three tests for comparing two gene lists are implemented. The Fisher's exact test is available in two versions for the "One Gene List is a Subset of the Other List" and "Mutually Exclusive Gene Lists" situations, as presented in 3.1. In the simulations study in Section 4 we found that the LAP test was found to perform the best among the intersecting gene lists test when the sampling strategy was unconditional, while the PIA test was the best for the conditional sampling strategy. We would assume that in most cases when comparing two gene lists the unconditional strategy is used. The LAP test is thus implmented in eGOn. No other GO-tool, to our knowledge, offers tests for the Intersecting Gene Lists situation. As noted in Section 4.6 the "Delete $A \cap B$" Pearson's asymptotic χ^2-test may be used (with caution) for the Intersecting Lists situation when the expected lengths of the list A and B are equal. The "Delete $A \cap B$" Fisher's exact test is implemented in FatiGO, Al-Shahrour et al. (2004).

Several interesting issues have not been adressed in this presentation, and may be explored in further work.

The motivation for developing test for the null hypothesis in Equation (1) was the case of testing for association between two intersecting gene reporter lists submitted to eGOn. Thus, the raw data underlying the statistical analyses producing the gene reporter lists are not submitted to eGOn. This means that eGOn may not offer permutation based methods for addressing the dependence structure between the genes. The statistical tests developed in this report are thus based on the assumption that under the null hypothesis the genes on the lists (or subsets of the lists in the intersecting gene lists

situations) act independently. This is also commonly assumed in other GO-tools. When testing for equality of the prevalence of a disease in two groups in the population, the observational unit is the individual and the assumption of independence of test results between individuals are in most cases not seen to be problematic. However, a possible extension of the methods developed in this report could be to look at different dependence between the observational units.

In this presentation we test the null hypothesis of association between gene lists in a given class of genes. However, when using Gene Ontology to select gene classes, we are interested in testing a hierarchy of gene classes. In eGOn the Benjamini and Hochberg step-up procedure, Benjamini and Hochberg (1995) for controlling the False Discovery Rate (FDR), is implemented to handle multiple testing. Other strategies with focus on the dependence between the gene classes in the GO-hierarchy may be investigated further, including the FDR-trees of Benjamini and Yekutieli (2003).

Other issues for further research are developing small sample tests for interesecting gene list (not asymptotic test) and to look at testing the null hypothesis of association of more than two gene list in a given gene class.

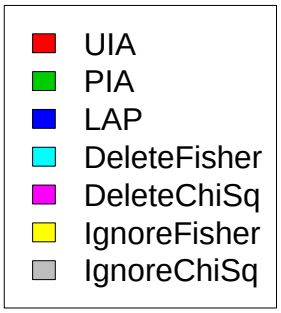


FIGURE 5: Colour codes for Figures 6-9 from the simulations study.

REFERENCES

AGRESTI, A. (1996). *An Introduction to Categorical Data analysis*, John Wiley & Sons, New York.

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd edn, Wiley.

AGRESTI, A. AND COULL, B. (1998). Approximate is better than "exact" for interval estimation of binomial proportions, *The American Statistician* pp. 119–126.

AL-SHAHROUR, F., DIAZ-URIARTE, R. AND DOPAZO, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, *Bioinformatics* 20(4): 578–580.

ALLISON, D. B., CUI, X., GRIER, P. AND SABRIPOUR, M. (2006). Microarray data analysis: from disarray to consolidation and consensus, *Nature Reviews Genetics* 7: 55–6.

BEISVÅG, V., JÜNGE, F. K. R., BERGUM, H., JØLSUM, L., LYDERSEN, S., GÜNTHER, C.-C., RAMAMPIARO, H., LANGAAS, M., SANDVIK, A. K. AND LÆGREID, A. (2006). Genetools - application for genomic functional annotation and statistical hypothesis testing, *Submitted*.

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* 57(1): 289–300.

BENJAMINI, Y. AND YEKUTIELI, D. (2003). Hierarchical fdr testing of trees of hypotheses, *Technical report*, Department of Statistics and OR, Tel Aviv University. http://www.math.tau.ac.il/ yekutiel/papers/treefd4.pdf.

EIDE, I. P., ROLFSENG, T., ISAKSEN, C. V., MECSEI, R., ROALD, B., LYDERSEN, S., SALVESEN, K. A., HARSEM, N. K. AND AUSTGULEN, R. (2006). Serious foetal growth restriction is associated with reduced proportions of natural killer cells in decidua basalis, *Virchows Arc* 448(3): 269–76.

HIRJI, K. F. (2006). *Exact analysis of discrete data*, Chapman & Hall.

KHATRI, P. AND DRAGICI, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics* 21(18): 3587–3595.

LEISERING, W., ALONZO, T. AND PEPE, M. S. (2000). Comparisons of Predictive Values of Binary Medical Diagnostic Tests for Paired Designs, *Biometrics* 56: 345–351.

R DEVELOPMENT CORE TEAM (2005). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

THE GENE ONTOLOGY CONSORTIUM (2000). Gene ontology: tool for the unification of biology., *Nature Genetics* 25: 25–29.

A NOTATION

The notation used in this report is presented in Figures 1 and 2, and in Table 6.

Notation	Explaination
N	the number of genes tested, i.e. the number of genes on a microarray slide where inference and annotation was possible.
n_A	the number of genes on list A, e.g. genes where event A has taken place.
n_B	the number of genes on list B, e.g. genes where event B has taken place.
$n_{G\cap(A\cup B)}$	the number of genes in gene class G that are on either gene lists A and/or B, e.g. genes on either of the gene lists A and/or B where event G has taken place.
$n_{A\cap B}$	the number of genes both on list A and B, e.g. genes where both events A and B have taken place.
$n_{A\cup B}$	the number of genes either on list A or B, e.g. genes where both of one of the events A and B have taken place.
$n_{A\cap B^*}$	the number of genes on list A, but not on list B, e.g. genes where both events A and B^* have taken place.
$n_{A^*\cap B}$	the number of genes on list B, but not on list A, e.g. genes where events A^* and B have taken place.
$X_{A\cap B\cap G}$	the number of genes in gene class G that are both on list A and B, e.g. genes where events A, B and G have all taken place.
$X_{A\cap B^*\cap G}$	the number of genes in gene class G that are on list A, but not on list B, e.g. genes where events A, B^* and G have all taken place.
$X_{A^*\cap B\cap G}$	the number of genes in gene class G that are on list B, but not on list A, e.g. genes where events A^*, B and G have all taken place.
$X_{A\cap G}$	the number of genes in gene class G that are on list A, e.g. genes where events A, and G have taken place.
$X_{B\cap G}$	the number of genes in gene class G that that are on list B, e.g. genes where events B and G have taken place.
$X_{A\cap B\cap G^*}$	the number of genes not in gene class G that are both on list A and B, e.g. genes where events A, B and G^* have all taken place.
$X_{A\cap B^*\cap G^*}$	the number of genes not in gene class G that that are on list A, but not on list B, e.g. genes where events A, B^* and G^* have all taken place.
$X_{A^*\cap B\cap G^*}$	the number of genes not in gene class G that are on list B, but not on list A, e.g. genes where events A^*, B and G^* have all taken place.
$X_{A\cap G^*}$	the number of genes not in gene class G that are on list A, e.g. genes where events A, and G^* have taken place.
$X_{B\cap G^*}$	the number of genes not in gene class G that that are on list B, e.g. genes where events B and G^* have taken place.

TABLE 6: Notation for the number of genes connected to different events.

B DERIVATION OF EQUIVALENCE OF NULL HYPOTHESES

If one gene list is a subset of the other, i.e. $B \subset A$, then we have the following equivalence of null hypotheses:

$$P(G|A) = P(G|B) \Leftrightarrow P(G|A \cap B^*) = P(G|B).$$

PROOF: Since $B \subset A$, then $B = A \cap B$ and $P(G|B) = P(G|A \cap B)$. The null hypothesis $H_0 : P(G|A) = P(G|B)$ can be written

$$P(G|A) = P(G|A \cap B)$$

and, thus G and B are conditionally independent given A. Hence, G and B^* are also conditionally independent given A, so $P(G|A \cap B) = P(G|A \cap B^*)$ and

$$P(G|A) = P(G|A \cap B^*).$$

ALTERNATIVE PROOF:

$$\begin{aligned}
P(G|A) &= P(G|B) \\
\frac{P(G \cap A)}{P(A)} &= P(G|B) \\
\frac{P(G \cap (A \cap B^*)) + P(G \cap (A \cap B))}{P(A)} &= P(G|B) \\
\frac{P(G|A \cap B^*)P(A \cap B^*) + P(G|A \cap B)P(A \cap B)}{P(A)} &= P(G|B) \\
\frac{P(G|A \cap B^*)P(A \cap B^*) + P(G|B)P(B)}{P(A)} &= P(G|B) \\
P(G|A \cap B^*)\frac{P(A \cap B^*)}{P(A)} &= P(G|B)[1 - \frac{P(B)}{P(A)} \\
P(G|A \cap B^*)\frac{P(A \cap B^*)}{P(A)} &= P(G|B)\frac{P(A \cap B^*)}{P(A)} \\
P(G|A \cap B^*) &= P(G|B)
\end{aligned}$$

C DETAILS OF THE RESULTS FROM THE SIMULATION STUDY

Case 2: e.e.300.15, with $P(G\|A) = P(G\|B) = 0.1$ ($p_1 = p_2 = p_3 = 0.1$)							
Unconditional simulations, $o_1 = o_2 = o_3 = 0.005$, $P(A) = 0.01$, $P(B) = 0.01$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0007	0.0006	0.0006	0.0005	0.0005	0.0000	0.0000
0.005	0.0048	0.0043	0.0042	0.0035	0.0044	0.0000	0.0000
0.01	0.0100	0.0094	0.0092	0.0074	0.0095	0.0001	0.0001
0.05	0.0516	0.0503	0.0500	0.0398	0.0499	0.0040	0.0052
0.1	0.1025	0.1006	0.1000	0.0807	0.1003	0.0151	0.0201
Conditional simulations, $n_1 = n_2 = n_3 = 150$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0011	0.0009	0.0009	0.0006	0.0009	0.0000	0.0000
0.005	0.0057	0.0050	0.0050	0.0030	0.0050	0.0001	0.0001
0.01	0.0110	0.0097	0.0097	0.0063	0.0097	0.0002	0.0004
0.05	0.0508	0.0489	0.0489	0.0323	0.0489	0.0034	0.0066
0.1	0.1031	0.0977	0.0975	0.0644	0.0977	0.0133	0.0199
Case 11: e.u.300.15a22.5, variant of case 2 with $P(G\|A) = 0.1$ and $P(G\|B) = 0.125$ ($p_1 = p_2 = 0.1$, $p_3 = 0.15$)							
Unconditional simulations, $o_1 = o_2 = o_3 = 0.005$, $P(A) = 0.01$, $P(B) = 0.01$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0230	0.0199	0.0194	0.0180	0.0205	0.0003	0.0005
0.005	0.0680	0.0633	0.0627	0.0542	0.0627	0.0045	0.0059
0.01	0.1056	0.0998	0.0992	0.0872	0.0999	0.0112	0.0135
0.05	0.2626	0.2580	0.2567	0.2288	0.2568	0.0762	0.0908
0.1	0.3776	0.3747	0.3733	0.3380	0.3739	0.1609	0.1833
Conditional simulations, $n_1 = n_2 = n_3 = 150$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0237	0.0199	0.0198	0.0151	0.0199	0.0006	0.0009
0.005	0.0686	0.0624	0.0624	0.0460	0.0624	0.0047	0.0066
0.01	0.1073	0.0974	0.0974	0.0728	0.0974	0.0095	0.0136
0.05	0.2661	0.2617	0.2615	0.2133	0.2617	0.0651	0.0916
0.1	0.3749	0.3735	0.3735	0.3094	0.3735	0.1403	0.1820

TABLE 7: Case 2 e.e.300.15, under the null hypothesis $P(G|A) = P(G|B)$, and Case 11 e.u.300.15a22.5 that is a modifi cation thereof, under the alternative hypothesis $P(G|A) \neq P(G|B)$.

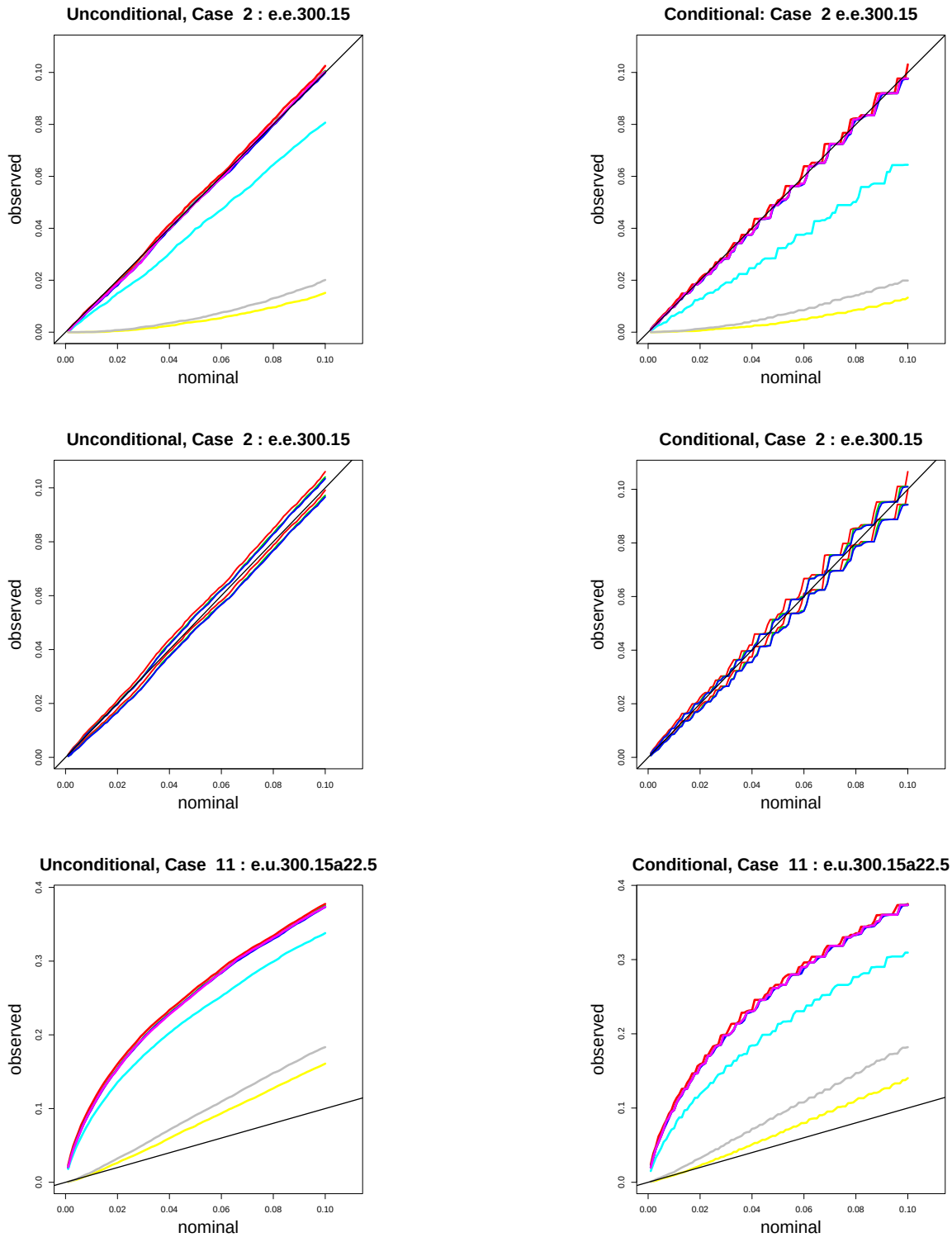


FIGURE 6: Unconditional simulations in the left column and conditional simulations in the right column. The top row shows nominal vs. observed significance levels, $\hat{\alpha}$ for the seven methods for Case 2. The middle row shows upper and lower confidence limits, $\hat{\alpha}_L$ and $\hat{\alpha}_U$, for the intersecting lists methods UIA, PIA and LAP, for Case 2. The bottom row shows the observed power at different significance levels for the seven methods for Case 11.

Case 5: b.e.600.15a45, with $P(G\|A) = P(G\|B) = 0.1$ ($p_1 = p_2 = p_3 = 0.1$)							
Unconditional simulations, $o_1 = 0.005$, $o_2 = o_3 = 0.015$, $P(A) = 0.02$, $P(B) = 0.02$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0008	0.0008	0.0007	0.0006	0.0008	0.0001	0.0001
0.005	0.0050	0.0048	0.0048	0.0042	0.0046	0.0009	0.0012
0.01	0.0098	0.0094	0.0095	0.0083	0.0095	0.0025	0.0029
0.05	0.0512	0.0507	0.0505	0.0442	0.0507	0.0202	0.0236
0.1	0.1030	0.1026	0.1023	0.0912	0.1025	0.0515	0.0595
Conditional simulations, $n_1 = 150$, $n_2 = n_3 = 450$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0012	0.0010	0.0010	0.0006	0.0010	0.0000	0.0001
0.005	0.0052	0.0052	0.0051	0.0038	0.0052	0.0007	0.0013
0.01	0.0109	0.0103	0.0103	0.0075	0.0103	0.0025	0.0034
0.05	0.0513	0.0500	0.0500	0.0388	0.0500	0.0186	0.0238
0.1	0.1020	0.1003	0.1003	0.0815	0.1003	0.0454	0.0577
Case 14 b.u.600.15a45a67.5: variant of case 5 with $P(G\|A) = 0.1$ and $P(G\|B) = 0.1375$ ($p_1 = p_2 = 0.1$, $p_3 = 0.3$)							
Unconditional simulations, $o_1 = 0.005$, $o_2 = o_3 = 0.015$, $P(A) = 0.02$, $P(B) = 0.02$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.1597	0.1549	0.1548	0.1454	0.1547	0.0688	0.0737
0.005	0.3018	0.2979	0.2974	0.2818	0.2971	0.1736	0.1857
0.01	0.3874	0.3832	0.3828	0.3644	0.3828	0.2467	0.2620
0.05	0.6253	0.6239	0.6234	0.6038	0.6244	0.5052	0.5253
0.1	0.7395	0.7384	0.7383	0.7215	0.7384	0.6429	0.6625
Conditional simulations, $n_1 = 150$, $n_2 = n_3 = 450$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.1579	0.1542	0.1542	0.1344	0.1542	0.0614	0.0730
0.005	0.3035	0.2980	0.2976	0.2670	0.2980	0.1620	0.1863
0.01	0.3890	0.3814	0.3814	0.3485	0.3814	0.2343	0.2640
0.05	0.6296	0.6271	0.6260	0.5874	0.6271	0.4882	0.5253
0.1	0.7385	0.7379	0.7379	0.7058	0.7379	0.6271	0.6616

TABLE 8: Case 5, b.e.600.15a45, under the null hypothesis $P(G|A) = P(G|B)$, and case 14 b.u.600.15a45a67.5 that is a modification thereof, under the alternative hypothesis $P(G|A) \neq P(G|B)$.

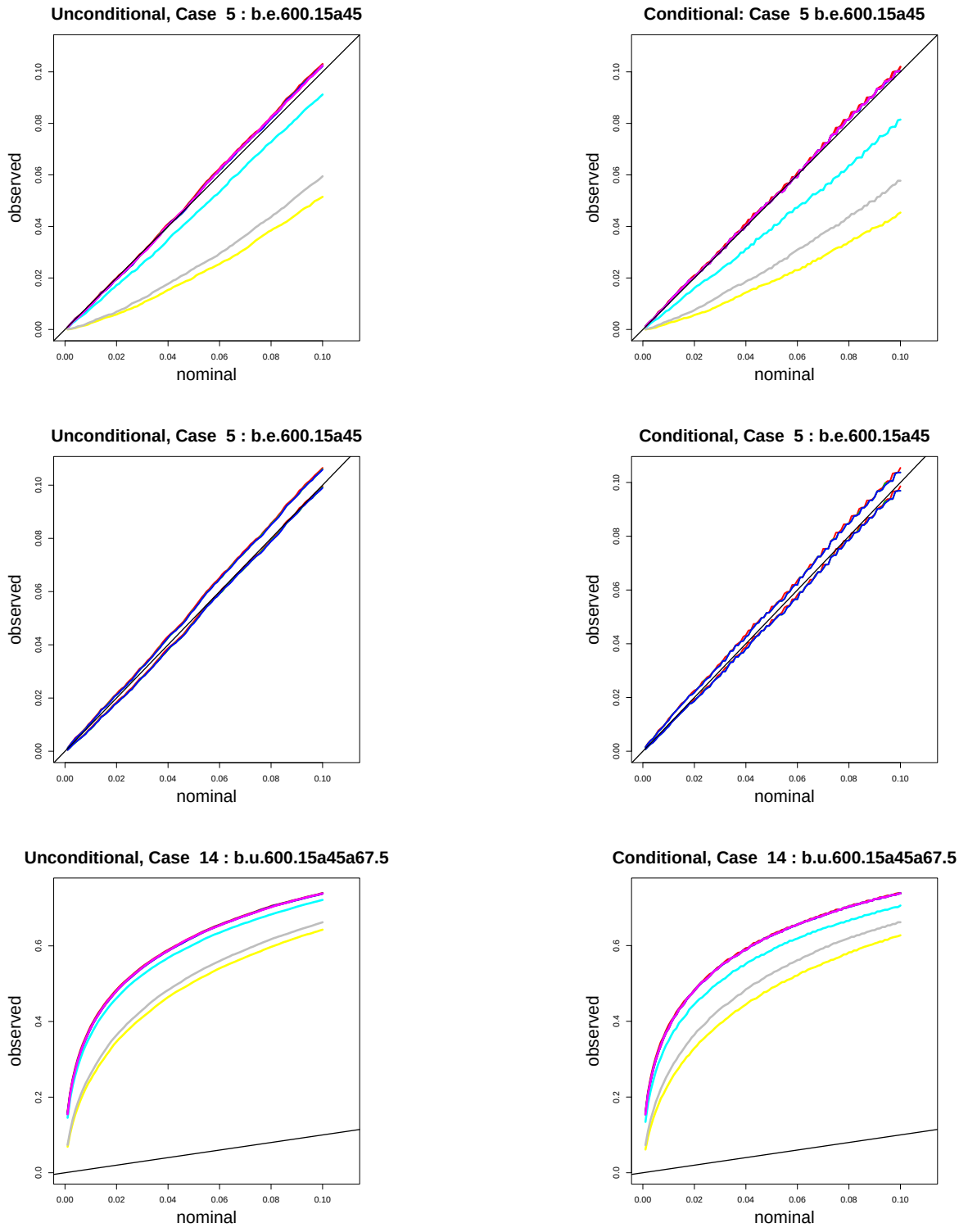


FIGURE 7: Unconditional simulations in the left column and conditional simulations in the right column. The top row shows nominal vs. observed significance levels, $\hat{\alpha}$ for the seven methods for Case 5. The middle row shows upper and lower confidence limits, $\hat{\alpha}_L$ and $\hat{\alpha}_U$, for the intersecting lists methods UIA, PIA and LAP, for Case 5. The bottom row shows the observed power at different significance levels for the seven methods for Case 14.

Case 7: b.b.600.75a27 with $P(G|A) = P(G|B) = 0.17$
($p_1 = 0.05$, $p_2 = p_3 = 0.06$)

Unconditional simulations, $o_1 = 0.005$, $o_2 = o_3 = 0.015$, $P(A) = 0.02$, $P(B) = 0.02$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0025	0.0022	0.0008	0.0009	0.0009	0.0000	0.0000
0.005	0.0104	0.0099	0.0044	0.0041	0.0051	0.0000	0.0000
0.01	0.0195	0.0188	0.0092	0.0088	0.0102	0.0000	0.0000
0.05	0.0750	0.0744	0.0495	0.0414	0.0491	0.0010	0.0013
0.1	0.1353	0.1345	0.0992	0.0853	0.0988	0.0050	0.0063
Conditional simulations, $n_1 = 150$, $n_2 = n_3 = 450$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0009	0.0009	0.0003	0.0008	0.0009	0.0000	0.0000
0.005	0.0043	0.0043	0.0018	0.0031	0.0043	0.0000	0.0000
0.01	0.0093	0.0089	0.0039	0.0061	0.0089	0.0001	0.0001
0.05	0.0488	0.0482	0.0289	0.0339	0.0482	0.0002	0.0004
0.1	0.1011	0.1011	0.0677	0.0727	0.1011	0.0018	0.0026

Case 14: b.u.600.75a27a36, variant of case 7 with $P(G|A) = 0.17$ and $P(G|B) = 0.185$
($p_1 = 0.05$, $p_2 = 0.06$, $p_3 = 0.08$)

Unconditional simulations, $o_1 = 0.005$, $o_2 = o_3 = 0.015$, $P(A) = 0.02$, $P(B) = 0.02$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0242	0.0234	0.0125	0.0143	0.0161	0.0000	0.0000
0.005	0.0643	0.0627	0.0397	0.0438	0.0488	0.0006	0.0007
0.01	0.0960	0.0947	0.0645	0.0691	0.0770	0.0014	0.0016
0.05	0.2383	0.2374	0.1874	0.1954	0.2147	0.0187	0.0212
0.1	0.3381	0.3372	0.2911	0.2953	0.3212	0.0540	0.0611
Conditional simulations, $n_1 = 150$, $n_2 = n_3 = 450$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0164	0.0158	0.0067	0.0116	0.0158	0.0000	0.0000
0.005	0.0500	0.0500	0.0291	0.0396	0.0500	0.0001	0.0001
0.01	0.0784	0.0775	0.0504	0.0618	0.0775	0.0003	0.0004
0.05	0.2210	0.2164	0.1698	0.1795	0.2164	0.0089	0.0129
0.1	0.3304	0.3304	0.2773	0.2820	0.3304	0.0363	0.0473

TABLE 9: Case 7, b.b.600.75a275, under the null hypothesis $P(G|A) = P(G|B)$, and case 16 b.u.600.75a27a36 that is a modifi cation thereof, under the alternative hypothesis $P(G|A) \neq P(G|B)$.

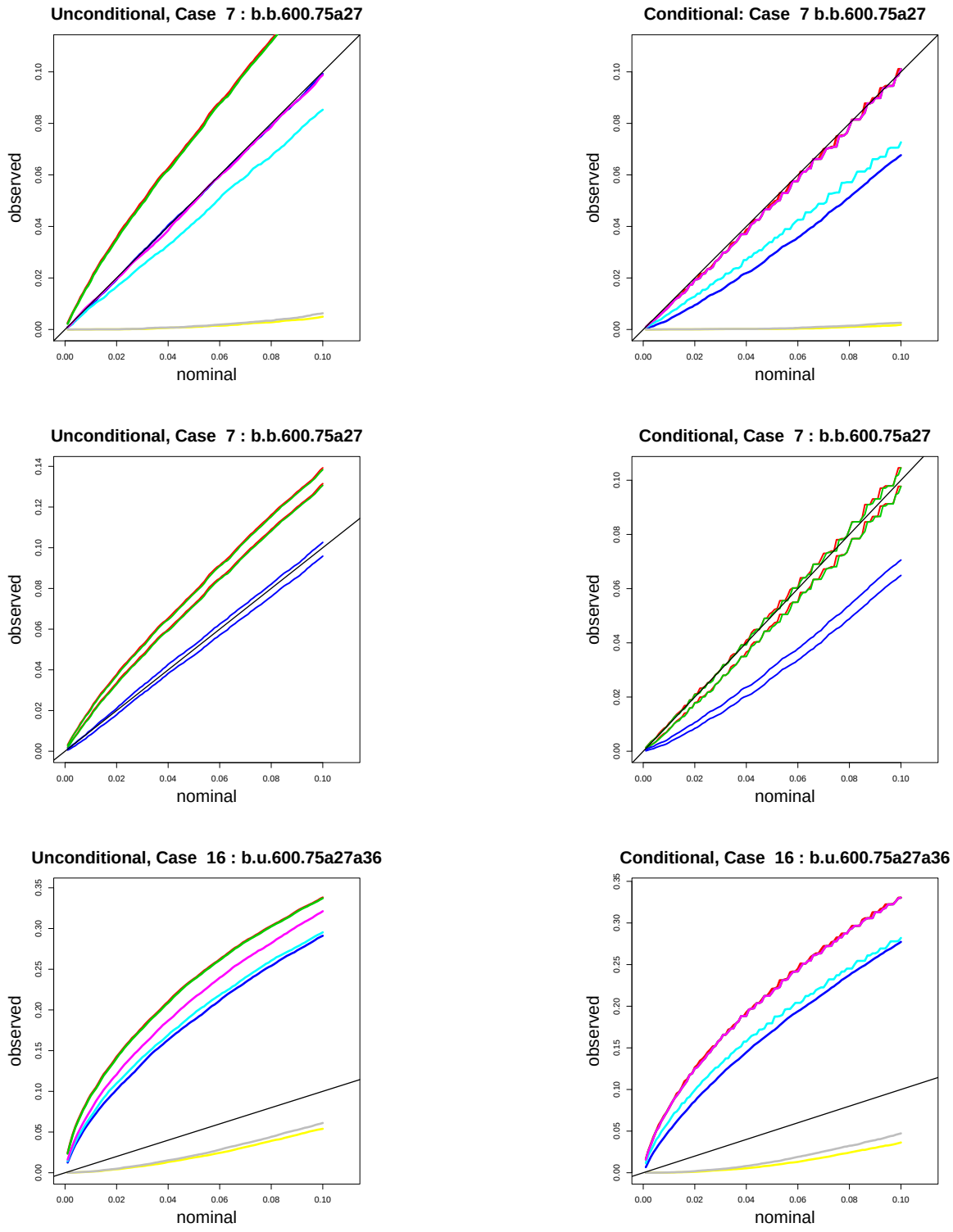


FIGURE 8: Unconditional simulations in the left column and conditional simulations in the right column. The top row shows nominal vs. observed significance levels, $\hat{\alpha}$ for the seven methods for Case 7. The middle row shows upper and lower confidence limits, $\hat{\alpha}_L$ and $\hat{\alpha}_U$, for the intersecting lists methods UIA, PIA and LAP, for Case 7. The bottom row shows the observed power at different significance levels for the seven methods for Case 16.

Case 9: u.b.50a150.9a6a36 with $P(G|A) = P(G|B) = 0.3$
($p_1 = 0.450$, $p_2 = 0.200$, $p_3 = 0.277$)

Unconditional simulations, $o_1 = 0.00067$, $o_2 = 0.001$, $o_3 = 0.0043$, $P(A) = 0.00167$, $P(B) = 0.005$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0036	0.0013	0.0014	0.0032	0.0016	0.0001	0.0001
0.005	0.0098	0.0068	0.0058	0.0142	0.0126	0.0010	0.0014
0.01	0.0171	0.0127	0.0106	0.0288	0.0267	0.0022	0.0031
0.05	0.0641	0.0573	0.0506	0.1026	0.1206	0.0176	0.0231
0.1	0.1197	0.1114	0.1022	0.1725	0.2127	0.0428	0.0558
Conditional simulations, $n_1 = 20$, $n_2 = 30$, $n_3 = 130$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0026	0.0006	0.0011	0.0043	0.0018	0.0001	0.0001
0.005	0.0085	0.0052	0.0045	0.0137	0.0122	0.0005	0.0007
0.01	0.0156	0.0105	0.0088	0.0283	0.0256	0.0014	0.0021
0.05	0.0590	0.0530	0.0474	0.1012	0.1239	0.0152	0.0203
0.1	0.1103	0.1032	0.0944	0.1733	0.2082	0.0395	0.0532

Case 18: u.u.50a150.9a6a40, variant of case 10 with $P(G|A) = 0.3$ and $P(G|B) = 0.3267$
($p_1 = 0.450$, $p_2 = 0.200$, $p_3 = 0.385$)

Unconditional simulations, $o_1 = 0.00067$, $o_2 = 0.001$, $o_3 = 0.0043$, $P(A) = 0.00167$, $P(B) = 0.005$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0073	0.0013	0.0031	0.0085	0.0054	0.0002	0.0002
0.005	0.0202	0.0073	0.0114	0.0331	0.0304	0.0015	0.0016
0.01	0.0316	0.0158	0.0201	0.0558	0.0551	0.0036	0.0043
0.05	0.0959	0.0725	0.0767	0.1729	0.2018	0.0265	0.0332
0.1	0.1585	0.1348	0.1371	0.2701	0.3166	0.0610	0.0765
Conditional simulations, $n_1 = 20$, $n_2 = 30$, $n_3 = 130$							
level	UIA	PIA	LAP	DeleteFisher	DeleteChiSq	IgnoreFisher	IgnoreChiSq
0.001	0.0071	0.0013	0.0024	0.0093	0.0049	0.0001	0.0001
0.005	0.0184	0.0072	0.0104	0.0312	0.0290	0.0012	0.0014
0.01	0.0288	0.0140	0.0181	0.0558	0.0569	0.0024	0.0029
0.05	0.0888	0.0657	0.0698	0.1668	0.2036	0.0254	0.0295
0.1	0.1485	0.1272	0.1267	0.2727	0.3110	0.0526	0.0661

TABLE 10: Case 9 u.b.50a150.9a6a36, under the null hypothesis $P(G|A) = P(G|B)$, and case 18 u.u.50a150.9a6a40 that is a modifi cation thereof, under the alternative hypothesis $P(G|A) \neq P(G|B)$.

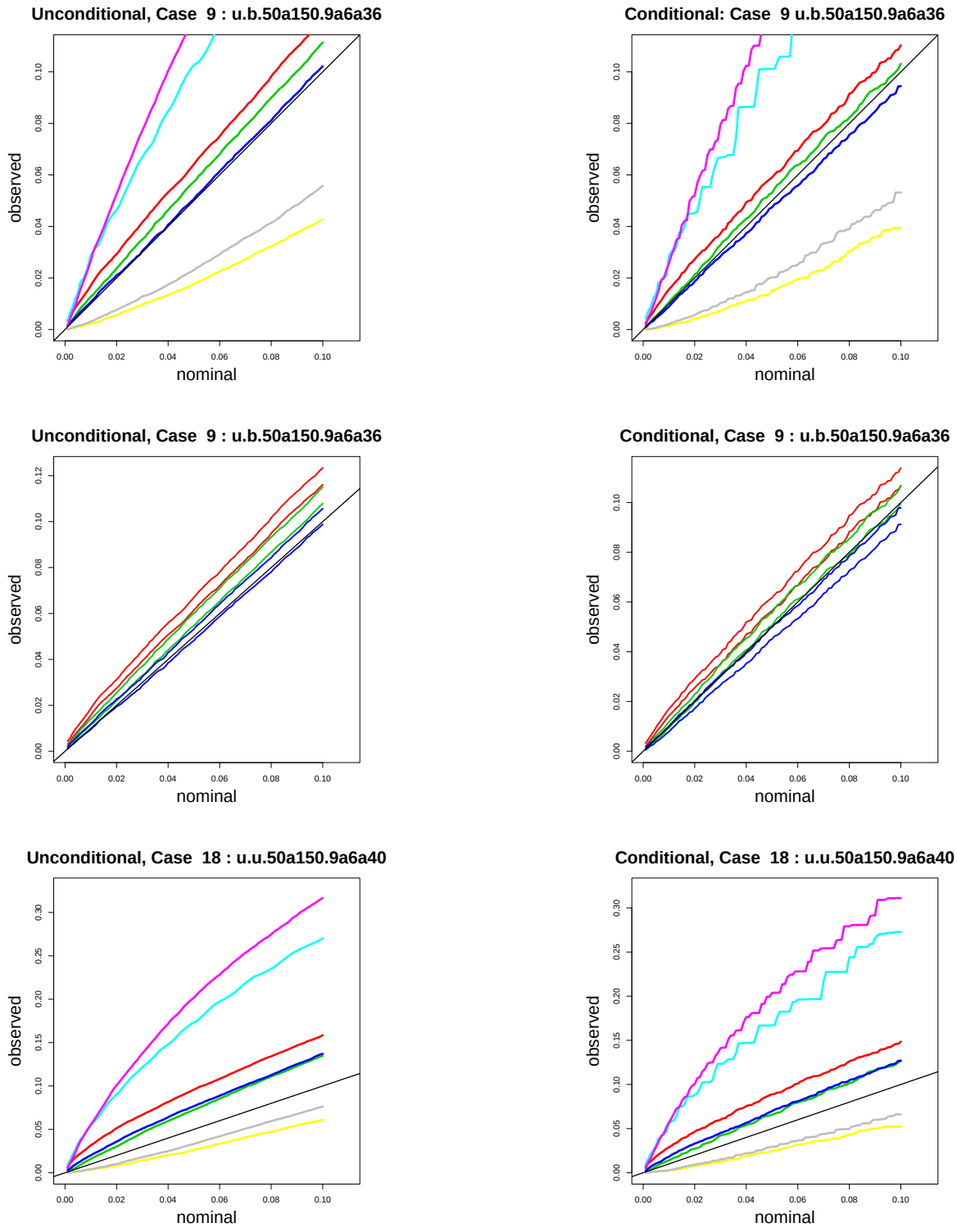


FIGURE 9: Unconditional simulations in the left column and conditional simulations in the right column. The top row shows nominal vs. observed significance levels, $\hat{\alpha}$ for the seven methods for Case 9. The middle row shows upper and lower confidence limits, $\hat{\alpha}_L$ and $\hat{\alpha}_U$, for the intersecting lists methods UIA, PIA and LAP, for Case 9. The bottom row shows the observed power at different significance levels for the seven methods for Case 18.