A procedure to statistically evaluate agreement of differential expression for cross-species genomics

Stan Pounds¹, Cuilan Lani Gao¹, Robert A. Johnson², Karen D. Wright³, Helen Poppleton⁴, David Finkelstein⁵, Sarah E. S. Leary⁶ and Richard J. Gilbertson⁴

¹Department of Biostatistics, St Jude Children’s Research Hospital, Memphis, TN, USA, ²Center for Childhood Cancer, The Research Institute at Nationwide Children’s Hospital and The Ohio State University College of Medicine, Columbus, OH, USA, ³Department of Oncology, ⁴Department of Developmental Neurobiology, ⁵Department of Information Sciences, St Jude Children’s Research Hospital, Memphis, TN and ⁶Department of Hematology-Oncology, Seattle Children’s Hospital, Seattle, WA, USA

Associate Editor: David Rocke

ABSTRACT

Motivation: Animal models play a pivotal role in translation biomedical research. The scientific value of an animal model depends on how accurately it mimics the human disease. In principle, microarrays collect the necessary data to evaluate the transcriptomic fidelity of an animal model in terms of the similarity of expression with the human disease. However, statistical methods for this purpose are lacking.

Results: We develop the agreement of differential expression (AGDEX) procedure to measure and determine the statistical significance of the similarity of the results of two experiments that measure differential expression across two groups. AGDEX defines a metric of agreement and determines statistical significance by permutation of each experiment’s group labels. Additionally, AGDEX performs a comprehensive permutation-based analysis of differential expression for each experiment, including gene-set analyses and meta-analytic integration of results across studies. As an example, we show how AGDEX was recently used to evaluate the similarity of the transcriptome of a novel model of the brain tumor ependymoma to that of a subtype of the human disease. This result, combined with other observations, helped us to infer the cell of origin of this devastating human cancer.


Contact: stanley.pounds@stjude.org

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 4, 2011; revised on June 6, 2011; accepted on June 13, 2011

1 INTRODUCTION

Microarray technology has enabled researchers to simultaneously measure the expression of thousands of genes in a biological tissue specimen. These data are commonly used to compare gene expression among biological conditions within the same species.

For example, the experiments may compare the transcriptomes of tumors and host normal tissue or between tumors arising in the same tissue. However, translating this wealth of data into other experimental systems has proven difficult because of limitations in comparing transcriptome data generated with different microarray platforms or from different species.

To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
2.1 Differential expression of individual probe-sets for one experiment

Suppose an array measures the expression of \( j = 1, \ldots, m \) probe-sets for each of \( g = 1, 2 \) groups defined by distinct biological conditions. Now, for each probe-set \( j \), define

\[
d_j = \bar{x}_j - \bar{x}_g
\]

(1)
as the difference between the average \( \bar{x}_j \) log-signal for group 1 and the average \( \bar{x}_g \) for group 2. Let \( d = (d_1, \ldots, d_m) \) be the vector of \( d_j \) for all probe-sets \( j \).

For each probe-set \( j \), compute a \( P \)-value for \( d_j \) by permutation of the group labels (Good, 2010; Gadbury et al., 2003). Let \( b = 1, \ldots, B \) index a set of permutations of assignment of the group-labels to the arrays. For \( b = 1, \ldots, B \), let \( d_{jb} \) be the value of (1) obtained by permutation \( b \) of the group labels. For each \( d_j \), the permutation \( P \)-value is the proportion of permuted group labels that yield an absolute value of (1) than does the original assignment of group labels. Mathematically, the \( P \)-value for \( d_j \) is given by

\[
P(d_j) = \frac{1}{B} \sum_{b=1}^{B} I(|d_{jb}| \geq |d_j|)
\]

(2)
where \( d_{jb} \) is the value of \( d_j \) computed using the original group assignments, and \( I(\cdot) \) is the indicator function that equals 1 if the enclosed statement is true and equals 0 otherwise. In practice, all possible permutations are used whenever it is computationally feasible to do so. Otherwise, a large number of randomly selected permutations are utilized. In some applications, the permutations should be restricted to preserve important stratification features of the experimental design.

2.2 Differential expression of gene-sets for one experiment

Suppose the probe-sets are annotated according to membership in a gene-set \( s \). Let \( j_s = 1, \ldots, m_s \) index the \( m_s \) probe-sets that belong to gene-set \( s \). The gene-set differential expression statistic

\[
h_s = \frac{1}{m_s} \sum_{j_s} d_{js}
\]

(3)
is the average of the absolute value of the differential expression statistics of probe-sets belonging to gene-set \( s \).

The statistic defined in (3) is a function only of the data of genes that belong to the gene-set and thus is self-contained by Goeman and Bühlmann’s (2007) definition. Goeman and Bühlmann’s (2007) show that self-contained gene-set testing procedures have greater statistical power than competitive testing procedures that compare the differential expression results of genes belonging to a gene-set to those of genes that do not belong to a gene-set.

A \( P \)-value for \( h_s \) may be determined by permutation of the group labels (Barry et al., 2005). Let \( b = 1, \ldots, B \) index a set of permutations of the group labels. Let \( h_{sb} \) be the value of (3) obtained using the original group labels. Also, let \( h_{sb} \) be the value (3) obtained by permutation \( b = 1, \ldots, B \). The permutation \( P \)-value

\[
P(h_s) = \frac{1}{B} \sum_{b=1}^{B} I(|h_{sb}| \geq |h_s|)
\]

(4)
is the proportion of permuted assignments of labels that yield a greater value of (3) than does the original assignment of labels. The same set of permutations is used for all \( h_s \) for each of the differential expression statistics for individual probe-sets and each gene-set.

2.3 Meta-analytic integration of probe-set results

Now, suppose that data have been collected in two experiments that perform biologically analogous comparisons of expression across two biological conditions. Additionally, suppose that we wish to perform a meta-analysis that combines the analysis results to identify differentially expressed probe-sets. Without loss of generality, assume that the probe-sets from each study have been matched and ordered so that the index \( j = 1, \ldots, m \) corresponds to the probe-sets that query the same gene in both studies. Let \( d_{1j} \) and \( d_{2j} \) represent the vectors differential expression statistics computed for probe-set \( j \) in the analysis of data from studies 1 and 2, respectively. Also, let \( P(d_{1j}) \) and \( P(d_{2j}) \) be the \( P \)-values computed from \( B_1 \) and \( B_2 \) permutations for each study, respectively.

Now, define

\[
Z(d_{1j}) = \text{sign}(d_{1j})\Phi^{-1} \left( 1 - \frac{B_1 P(d_{1j}) + 0.5}{B_1 + 1} \right)
\]

and

\[
Z(d_{2j}) = \text{sign}(d_{2j})\Phi^{-1} \left( 1 - \frac{B_2 P(d_{2j}) + 0.5}{B_2 + 1} \right)
\]

where \( \Phi^{-1}(\cdot) \) is the quantile function of the standard normal distribution. Note that the definitions for the \( z \)-statistics incorporate information about the direction (sign) and significance \( (P \)-value) of differential expression from each study and include a correction for discrete permutation \( P \)-values by adding 0.5 to the numerator and adding 1 to the denominator of the fraction. If the null hypothesis is true for both studies, then the \( P \)-values are approximately uniform. Thus, the \( z \)-statistics are independent and each \( z \)-statistic approximately follows a standard normal distribution. Therefore, \( Z(d_{1j}) + Z(d_{2j}) \) is approximately normal with mean 0 and variance 2. Stouffer et al., (1949) show that a meta-analysis \( P \)-value is then

\[
P(d_{1j}, d_{2j}) = 2\Phi \left( -\frac{Z(d_{1j}) + Z(d_{2j})}{\sqrt{2}} \right)
\]

where \( \Phi(\cdot) \) is the cumulative distribution function of the standard normal distribution.

2.4 Meta-analytic integration of gene-set results

Now, suppose that (3) has been used to compute statistics \( h_{1s} \) and \( h_{2s} \) and their corresponding \( P \)-values \( P(h_{1s}) \) and \( P(h_{2s}) \) for differential expression analysis of gene-set \( s \) for study 1 and study 2. Define

\[
Z(h_{1s}) = \Phi^{-1} \left( 1 - \frac{B_1 P(h_{1s}) + 0.5}{B_1 + 1} \right)
\]

and

\[
Z(h_{2s}) = \Phi^{-1} \left( 1 - \frac{B_2 P(h_{2s}) + 0.5}{B_2 + 1} \right)
\]

as \( z \)-statistics that characterize the statistical significance of \( h_{1s} \) and \( h_{2s} \), respectively. Recall that \( h_{1s} \geq 0 \) and \( h_{2s} \geq 0 \) by definition (3). If the null hypothesis is true for both studies, then \( Z(h_{1s}) + Z(h_{2s}) \) approximately follows a normal distribution with mean 0 and variance 2. Therefore,

\[
P(h_{1s}, h_{2s}) = 1 - \Phi \left( \frac{Z(h_{1s}) + Z(h_{2s})}{\sqrt{2}} \right)
\]

is a meta-analysis \( P \)-value for the gene-set analysis that combines the analysis results from the two studies.

2.5 Agreement of differential expression

Now we describe how AGDEX evaluates the agreement of differential expression among probe-sets that belong to a gene-set \( s \) and evaluate its statistical significance by permutation. Again, suppose that data has been collected for two studies that each compare expression across two groups. For simplicity of notation, assume in this section that the data have already been limited to probe-sets that are members of gene-set \( s \) and that the order of these probe-sets has been matched across the two studies. Let \( j_s = 1, \ldots, m_s \) index the probe-sets. For simplicity of notation, the subscript \( s \) will be omitted from \( j_s \) and \( m_s \) in this section. For each probe-set \( j_s \), let \( d_{1s} \) and \( d_{2s} \) be the values of (1) from study 1 and study 2, respectively. Let \( d_1 = (d_{11}, \ldots, d_{1m}) \) and \( d_2 = (d_{21}, \ldots, d_{2m}) \).
The AGDEX procedure uses two statistics to measure the agreement of \( d_1 \) and \( d_2 \). The first statistic is the cosine of the angle between \( d_1 \) and \( d_2 \) in multivariate space (Fitzpatrick, 1996), i.e.

\[
\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{||d_1|| \times ||d_2||} \tag{5}
\]

The second statistic is the difference of proportions (dop). Formally, the second statistic is defined as

\[
dop(d_1, d_2) = \frac{1}{m} \sum_{j=1}^{m} \text{sign}(d_{1j}d_{2j}) \tag{6}
\]

which is the difference between the proportion of probe-sets that show agreement and the proportion that show disagreement in the direction of differential expression. Both statistics have range \([-1, +1]\) with +1 indicating the greatest agreement of differential expression results and -1 indicating the greatest disagreement of differential expression results.

A \( P \)-value can be computed for each statistic by permutation of the group labels for either study. Let \( a_0 \) be the value of the cosine or dop statistic obtained from the original assignment of group labels in both studies. Let \( n = 1, \ldots, B \) index permutations of the group labels from study 1 and let \( v = 1, \ldots, B \) index permutations of the group labels from study 2. Let \( d_{1v} \) be the vector of differential expression statistics obtained from permutation \( v \) of the group labels of Study 1. For each \( v \), let \( a_v \) be computed using \( d_{1v} \) and the observed \( d_2 \). Thus,

\[
P_1(a_0) = \frac{1}{B} \sum_{v=1}^{B} \text{I}(|a_v| \geq |a_0|) \tag{7}
\]

is a \( P \)-value obtained by permutation of the group labels in study 1. This \( P \)-value compares the observed value \( a_0 \) of the agreement statistic to the distribution of agreement statistics obtained under the null hypothesis that there is no differential expression in Experiment 1.

Similarly, for each \( v \), let \( d_{2v} \) be the vector of differential expression statistics obtained by permutation \( v \) of group labels for Study 2. Also, define \( a_2 \) as the value of the agreement statistic obtained using \( d_{2v} \). This yields

\[
P_2(a_0) = \frac{1}{B} \sum_{v=1}^{B} \text{I}(|a_v| \geq |a_0|) \tag{8}
\]

as a \( P \)-value for \( a_0 \) based on permutation of group assignments for Study 2.

### 2.6 Statistical Properties

For each study \( s = 1, 2 \) and each probe-set \( j = 1, \ldots, m \), let the expected value of the differential expression statistic be \( E(d_{1j}) = \delta_j \). Also, let \( d_{1} \) and \( d_{2} \) be the vectors of true difference between means for Experiments 1 and 2, respectively. Recall that \( d_1 \) and \( d_2 \) are independent because they are estimated from datasets collected in separate studies. Therefore,

\[
E(d_{1j}d_{2j}) = \sum_{j=1}^{m} E(d_{1j}d_{2j}) = \sum_{j=1}^{m} E(d_{1j})E(d_{2j}) = \delta_1 \delta_2 \tag{9}
\]

and thus the expected value of the numerator of the cosine statistic in (5) is zero if \( d_1 \) and \( d_2 \) are orthogonal. Similarly, the expected value of the dop statistic in (6) is zero if \( d_1 \) and \( d_2 \) are orthogonal. In the specific orthogonal case, none of the genes with differential expression in Study 1 are differentially expressed in Study 2 (\( d_{1j} = 0 \) for each \( j \) with \( \delta_j \neq 0 \)), the expected value of the dop statistic is 0 and the expected value of the cosine statistic is approximately zero. Consequently, for analyses with this type of orthogonal differential expression, the \( P \)-values are not likely to be significant because they compare the absolute value of the observed statistic to that of statistics obtained by permutation.

In non-orthogonal cases, the probability of a significant result increases with sample size of each group in each study. As the sample sizes increase, the variance of \( d_{2j} \) decreases for each \( s \) and \( j \) and thus the variances of the agreement statistics also decrease. The probability of a significant result also increases with the magnitude of

\[
\Delta = \frac{d_1 \cdot d_2}{||d_1|| \times ||d_2||} \tag{10}
\]

for the cosine agreement statistic and with the magnitude of

\[
\Delta' = \text{sign}(d_{1j}) \text{sign}(d_{2j}) \tag{11}
\]

for the dop statistic.

### 2.7 Statistical and biological interpretation

The statistical interpretation of AGDEX results must consider the magnitude and sign of each agreement statistic and the sample sizes of each group in each differential expression analysis comparison. Analyses with large sample sizes may have power to detect some subtle but biologically uninteresting agreements as statistically significant. In such cases, AGDEX will give small agreement statistics with small \( P \)-values. Conversely, analyses with small sample sizes may not have sufficient power to detect strong agreements as statistically significant. With very small sample sizes, it is impossible for permutation procedures to give very small \( P \)-values. For example, a 3 versus 3 comparison cannot yield a permutation \( P \)-value smaller than 1/10 (Gadbury et al., 2003; Pounds and Dyer 2008). In these cases, large agreement statistics with moderate \( P \)-values may be of biological interest. In some analyses, one comparison may have large sample size and the other comparison may have a small sample size. In these analyses, one may wish to focus attention on the \( P \)-values from permuting the labels of the study with the larger sample size. In analyses where both comparisons have large sample size, one may take the maximum of the four \( P \)-values (one \( P \)-value for each of two agreement statistics by permutation of group labels from each of two studies) to focus attention on the most robust results.

The biological interpretation of AGDEX results must consider that the parameters \( \Delta \) and \( \Delta' \) are functions of the differential expression status of \( j = 1, \ldots, m \) genes for each of two studies that each compare expression between two groups. A significant result may be driven by a subset of genes with statistically influential patterns of agreement. The genes with statistically influential data may not have the greatest biological relevance. Also, the \( P \)-values evaluate significance against the null hypothesis that there is no differential expression in the experiment for which the group labels are permuted. Permutation simulates values of a test statistic under the null hypothesis that group labels are random and arbitrary. Thus, significant agreement may not indicate biologically meaningful agreement. Therefore, a significant AGDEX result alone does not necessarily indicate that an animal model accurately recapitulates the biologically important characteristics of a human disease.

Nevertheless, a significant AGDEX result can help direct additional biological studies to more thoroughly evaluate the fidelity of an animal model. We used the AGDEX procedure to characterize the transcriptional similarity of novel brain tumors in mice with a series of types of human brain tumors. We found that one mouse tumor showed significant transcriptomic agreement with one subtype of human ependymoma (Johnson et al., 2010) and another mouse model showed significant transcriptomic agreement with one subtype of human medulloblastoma (Gibson et al., 2010). AGDEX did not find significant agreement with other human brain tumor subtypes. Thus,
As an example, we apply AGDEX to the data from our published study of the expression of murine neural stem/progenitor cells, murine brain tumors and the human brain tumor ependyoma (Johnson et al., 2010). The Affymetrix 430v2 array was used to profile the expression of 45 037 probe-sets for 13 brain tumors from mice and a collection of 179 normal mouse stem cells. Additionally, the Affymetrix U133+2 array was used to profile the expression of 54 613 probe-sets for 83 human ependymoma tumors. The human tumors were classified as belonging to the novel subgroup ‘D’ or another subgroup. The expression data were normalized with the MAS 5.0 algorithm. We downloaded the 1454 biological process gene-set definitions from the gene-set enrichment analysis website (www.broadinstitute.org/gsea) for the U133+2 array. We used the Affymetrix best-match dataset (available from www.affymetrix.com) to define 78 061 pairs of ortholog-matched probe-sets across the two arrays. We used the best-match dataset to define the gene-sets for the mouse array probe-sets. Adaptive permutation with \( B_{\text{min}} = 100 \) and \( B_{\text{max}} = 10 000 \) was used to compute \( P \)-values for gene-set statistics. \( P \)-values for individual probe-set statistics and the genome-wide dop and cosine statistics were determined using 10 000 permutations. Here, we report statistical summaries of the results and refer readers to our previous publication (Johnson et al., 2010) for more detailed results and biological interpretations of the findings.

### 2.8 Comparison with other procedures
The AGDEX procedure addresses a different biological question than other methods proposed to evaluate the transcriptional similarity of a model system (cell line or animal tissue) to a human tissue. Sandberg and Ernberg (2005) and Zheng-Bradley et al., (2010) apply dimension reduction techniques such as singular value decomposition or principal components analysis to expression data matrices that include arrays from both the human tissues and the model systems. The similarity of expression is measured by proximity of the different systems’ samples in the lower dimensional space. These methods are useful for describing similarities in a dataset that includes a very diverse set of biological systems (multiple tissue types and multiple species). However, these approaches do not formally evaluate statistical significance and their results depend on arbitrary user choices of tuning parameters such as the number of dimensions in the reduced dataset. Poisson and Ghosh (2007) propose a method to determine whether the profile of genes that differentiate between two model systems is useful for developing a predictor of an outcome for humans. In contrast, AGDEX addresses the specific biological question of whether the difference between tumor and control in mice is similar to the difference between tumor and control in humans.

### 2.9 Adaptive permutation to reduce computational burden
Permutation testing is a computationally intensive procedure. We have developed an adaptive permutation procedure that retains the statistical rigor of permutation testing while greatly reducing its computational cost for many genomics applications (Pounds et al., 2011). Adaptive permutation testing computes permutation statistics until observing a minimum of \( B_{\text{min}} \) permutation statistics that exceed the observed value of the test statistic or until a maximum of \( B_{\text{max}} \) permutation statistics have been computed. In this way, the number of permutations \( B \) is a random variable with a truncated negative binomial distribution. The success probability of the truncated negative binomial distribution is the exact \( P \)-value \( \pi \) based on all possible permutations and the number of required successes in \( B_{\text{min}} \). The expected number of permutations is approximately \( \min(B_{\text{min}}, B_{\text{max}})/\pi \). The effect is that a small number of permutations are performed for tests with large \( \pi \) (and thus not of much interest) and a very large number of permutations are performed to achieve the desired precision for tests with very small \( \pi \). In many genomics applications, adaptive permutation testing greatly reduces the computational effort because most tests have insignificant \( P \)-values.

AGDEX allows users the option to utilize adaptive permutation to compute the gene-set differential expression \( P \)-value of Section 2.2 and the agreement of differential expression \( P \)-values of Section 2.5 for gene-sets. For AGDEX, the adaptive permutation is modified so that at least \( B_{\text{min}} \) permutations yield extreme values for each of the 3 gene-set statistics: the gene-set differential expression statistic in (3), the cosine agreement statistic in (5) and the difference of proportions agreement statistic in (6).

### 3 RESULTS
As an example, we apply AGDEX to the data from our published study of the expression of murine neural stem/progenitor cells, murine brain tumors and the human brain tumor ependyoma (Johnson et al., 2010). The Affymetrix 430v2 array was used to...
We have developed the agreement of differential expression (AGDEX) procedure to integrate differential expression analysis results across two experiments that may utilize different platforms or even different species. The AGDEX procedure can be used to examine the transcriptional fidelity of animal models of human disease by evaluating the similarities of transcriptome-wide differential expression profiles across species. AGDEX may also be used to integrate differential expression results that use different platforms for the same species. In our studies, the AGDEX procedure enabled us to confirm the transcriptomic fidelity of newly generated mouse models and thereby infer the cell of origin of two distinct human brain cancers (Gibson et al., 2010; Johnson et al., 2010).

The AGDEX procedure also performs a comprehensive and statistically rigorous permutation-based analysis to address other biologically important questions. For each experiment, the AGDEX procedure performs differential expression analysis at the probe-set and gene-set levels. Additionally, the AGDEX procedure meta-analytically integrates differential expression results at the probe-set and gene-set levels to improve statistical power to identify genes or pathways that are important in one or both experiments. Finally, the AGDEX procedure also evaluates similarity of differential expression across the two experiments for predefined gene-sets.

The AGDEX procedure computes P-values by permuting assignments of group labels within each experiment. Permutation is computationally demanding but yet is critical to obtain an accurate measure of statistical significance, especially for identifying differentially expressed gene-sets and evaluating the agreement of differential expression. Some computationally simple approaches for gene-set analyses utilize statistical models of chance that inappropriately use final results for genes as the unit of analysis. For example, in our setting, one might use the binomial distribution to model the number of ortholog pairs that have the same sign for differential expression analysis in both experiments. Such an approach grossly exaggerates statistical significance by inappropriately using the number of ortholog pairs as the “sample size” and ignoring the correlation among genes. Permutation of group labels preserves correlation among genes by keeping gene expression values for each subject together during the random reassignment, is explicitly linked to the experimental design and uses the correct sample size by retaining the correct number of group labels and expression data vectors (Allison et al., 2006; Barry et al., 2005).

Nevertheless, the AGDEX procedure is computationally feasible. Our example analysis was completed on a desktop computer in 2.5 h. The sample size and number of genes and gene-sets are similar to those of other contemporary applications. Additionally, utilization of adaptive permutation will help reduce computational burden by performing fewer permutations for gene-sets with insignificant results (Pounds et al., 2011). This may be especially useful if the more statistically robust and computationally demanding distance-based statistic of Nettleton et al., (2008) for differential expression analysis of gene-sets is incorporated into later versions of AGDEX.

The example analysis also shows that the AGDEX procedure and our previously published results are robust against the strategy used to select the ortholog pairs that are included in the analysis. In our previously published study, the expression data were used to select which ortholog pairs from the Affymetrix best match definitions were retained for the AGDEX analysis. In this work, no such filtering was applied. The results of the analyses based on the cosine statistic are qualitatively similar to the previously published findings. The robustness of AGDEX results against filtering of ortholog pairs is
We also thank Lei Shi, Arzu Onar and Xueyuan Cao for reviewing the manuscript and providing helpful editorial advice.

**ACKNOWLEDGEMENTS**

We thank Shengping Yang and the staff of the Hartwell Center for Bioinformatics and Biotechnology for technical assistance. We also thank Lei Shi, Arzu Omar and Xueyuan Cao for reviewing the manuscript and providing helpful editorial advice.

**REFERENCES**


