Assessing Equivalent and Inverse Change in Genes Between Diverse Experiments

Lisa Neums
University of Kansas Medical Center
Devin C. Koestler
University of Kansas Medical Center
Qing Xia
University of Kansas Medical Center
Jinxiang Hu
University of Kansas Medical Center
Shachi Patel
University of Kansas Medical Center
Shelby Bell-Glenn
University of Kansas Medical Center
Dong Pei
University of Kansas Medical Center
Bo Zhang
University of Kansas Medical Center
Samuel Boyd
University of Kansas Medical Center
Prabhakar Chalise
University of Kansas Medical Center
Jeffrey A. Thompson (jthompson21@kumc.edu)
University of Kansas Medical Center

Research Article

Keywords: differential gene expression, equivalence test, study comparison

Posted Date: November 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1005370/v1

Assessing Equivalent and Inverse Change in Genes between Diverse Experiments

Lisa Neums[1,2], Devin C. Koestler[1,2], Qing Xia[1,2], Jinxiang Hu[1,2], Shachi Patel[1,2], Shelby Bell-Glenn[1,2], Dong Pei[1,2], Bo Zhang[1], Samuel Boyd[1,2], Prabhakar Chalise[1,2], Jeffrey A. Thompson[1,2]

[1]Department of Biostatistics & Data Science, University of Kansas Medical Center, 3901 Rainbow Blvd. Kansas City, KS 66160, USA

[2]University of Kansas Cancer Center, 8919 Parallel Parkway, Suite 326, Kansas City, KS 66112, USA

Corresponding author:

Jeffrey A. Thompson

Phone: (913) 588-8970

E-mail: jthompson21@kumc.edu

Abstract

Background: It is important to identify when two exposures impact a molecular marker (e.g., a gene's expression) in similar ways, for example, to learn that a new drug has a similar effect to an existing drug. Currently, statistically robust approaches for making comparisons of equivalence of effect sizes obtained from two independently run treatment versus control comparisons have not been developed.

Results: Here, we propose two approaches for evaluating the question of equivalence between effect sizes of two independent studies: a bootstrap test of the Equivalent Change Index (ECI), which we previously developed, and performing Two One-Sided t-Tests (TOST) on the difference in log-fold changes directly. The ECI of a gene is computed by taking the ratio of the effect size estimates obtained from the two different studies, weighted by the maximum of the two p-values and giving it a sign indicating if the effects are in the same or opposite directions, whereas TOST is a test of whether the difference in log-fold changes lies outside a region of equivalence. We used a series of simulation studies to compare the two tests on the basis of sensitivity, specificity, balanced accuracy, and F1-socre. We found that TOST is not efficient for identifying equivalently changed gene expression values (F1-score = 0) because it is too conservative, while the ECI bootstrap test shows good performance (F1-score = 0.96). Furthermore, applying the ECI bootstrap test and TOST to publicly available microarray expression data from pancreatic cancer of tumor tissue and peripheral blood mononuclear cells (PBMC) showed that, while TOST was not able to identify any equivalently or inversely changed genes, the ECI bootstrap test identified genes associated with pancreatic cancer, intestinal cancer, and other disease types associated with pancreatic cancer.

Conclusion: A bootstrap test of the ECI is a promising new statistical approach for determining if two diverse studies show similarity in the differential expression of genes and can help to identify genes which are similarly influenced by a specific treatment or exposure.

Keywords: differential gene expression, equivalence test, study comparison

Background

In whole genome differential gene expression studies, the difference in the expression of thousands of genes among groups (e.g. treatment group vs control group) is investigated with the motivation of finding underlying mechanisms of different conditions such as cancer or to observe treatment effects (1). An emerging use for gene expression data is to identify genes that are affected in similar or opposing ways across different studies. There are various reasons for doing so, including validating the results of a study (2), finding common underlying mechanisms of a disease (3, 4), or investigating similar treatment effects of different drugs (5). One of the important challenges in comparing studies is variable study conditions such as technology, environment, and personnel (6). For these reasons, gene expression levels are challenging to directly compare across studies. However, the gene expression levels themselves are typically not of primary interest, but instead, the focus is typically on the nature of the change in gene expression between treatment or disease conditions (1). Therefore, there is a need for statistical methods that can be used to validate the similarity in expression changes across studies.

Given the lack of statistical methods that can test such equivalence hypotheses, researchers currently employ naïve methods for determining if treatments have similar effects on gene expression. The most-commonly employed method focused on intersection, which simply involves finding the intersection of differentially expressed genes across studies, without determining the probability of such intersections occurring by chance. One example is the e-cigarette study from Shen et al. (4), in which it investigated if similar pathways were enriched in cigarette smokers. In the process of analyzing their results, they observed that the same or different genes were differentially expressed at several time points of their study and used this result as well as a subsequent gene enrichment study to form their conclusion. Another example is a study from L. E. Blake et al. (3), which compares gene expression across tissues in

different animal species. Again, they define genes of interest as being both equivalently changed in the same direction and statistically significant in a pairwise comparison of tissue types. However, because this naïve approach is not statistically motivated, it is prone to false positives, i.e. declaring a gene equivalently changed when the differential expression of the gene between two studies is in the same direction only by chance. This could lead to a large number of candidate genes requiring validation, which could be time-consuming and cost prohibitive (7, 8).

In this study, we introduce and compare two statistical tests to test the hypothesis of equivalent change in gene expression between two studies. An advantage to testing for an equivalent change in differential gene expression is that the log2-fold changes are likely to be far more comparable across studies than the gene expression itself due to variable study conditions such as technology, environment, and personnel (9, 10). It should be noted that these tests work equally well to find significant opposing changes in gene expression which, for example, could be used to suggest treatments that might reverse changes associated with a disease or to identify genes affected in opposite ways in a gene knockout and overexpression experiment. The first test is an adaptation of the two one-sided t-tests (TOST) applied to the fold-changes from a differential gene expression analysis. Although traditionally the TOST approach has been used to establish bioequivalence of drugs (e.g. to approve generics) (11), it has been adapted to several other cases (12-14), including finding equivalently expressed genes in the same study (15). Nevertheless, it has never been adapted to differential gene expression analysis. The second test uses the Equivalent Change Index (ECI), introduced by Thompson and Koestler (16), together with a bootstrap procedure to calculate confidence intervals. In the following, we will explain the mechanisms of both approaches followed by a simulation study to compare and assess the performance of the two methods. This is followed by a real data application using different publicly available gene expression studies of pancreatic cancer to demonstrate if the tests are able to identify biologically plausible results.

Results

Simulation

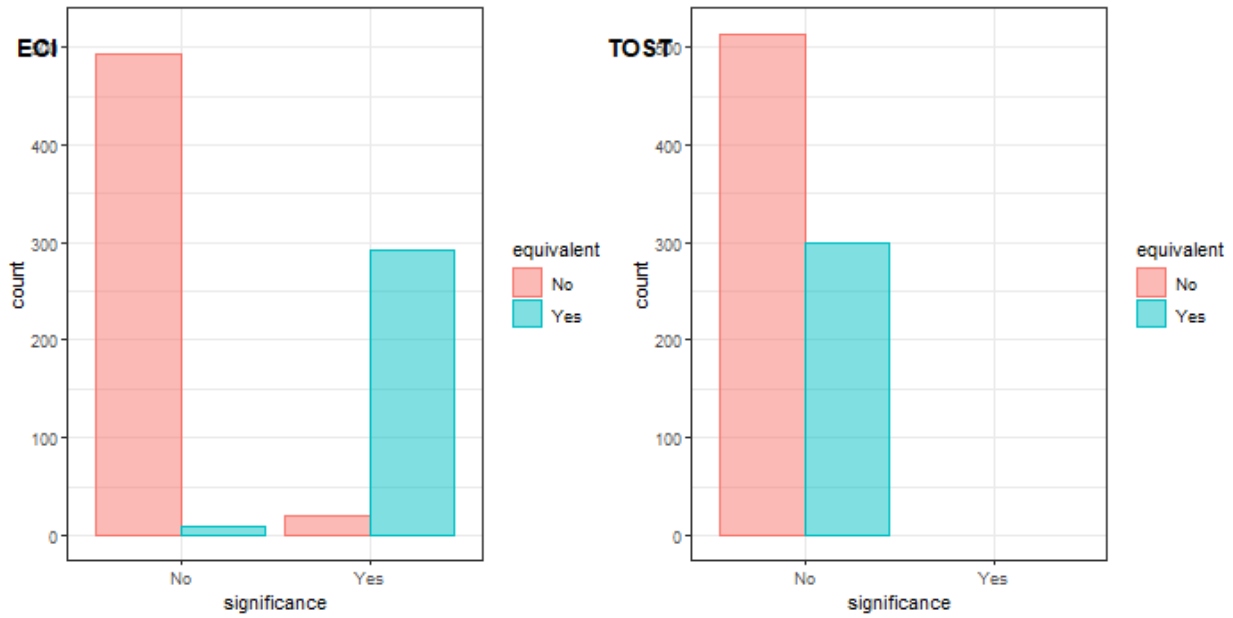
Simulation 1 was comprised of simulated study 1 and simulated study 2. Each simulated study contained a simulated treatment and control. The two simulated studies were created such that 30% of the genes were equivalently changed with varying degrees of equivalence. Figure 1 and Table 1 show the decision of significance of the two tests (ECI bootstrap test and TOST) for the example simulation mentioned above. As can be observed, for a sample size of the comparator groups (e.g., case versus control) of 20 people in each of the control group and treatment group, the ECI bootstrap test is able to identify the majority of equivalently changed genes (in this example sensitivity = 0.97) but also falsely identifies not equivalently changed genes (specificity < 1). TOST on the other hand, does not misclassify any non-equivalently changed genes but fails to identify any of the equivalently changed genes.

Table 1: Performance measurements for one single simulation with group size 20.

Performance	**ECI bootstrap test**	**TOST**
Sensitivity	0.973	0
Specificity	0.963	1
Balanced accuracy	0.968	0.5
F1-score	0.956	0

Legend: Decisions of equivalence are made using the q-value of 0.05 for ECI bootstrap test and TOST. The TOST is too conservative to identify any equivalently changed genes, while the ECI bootstrap test shows good performance.

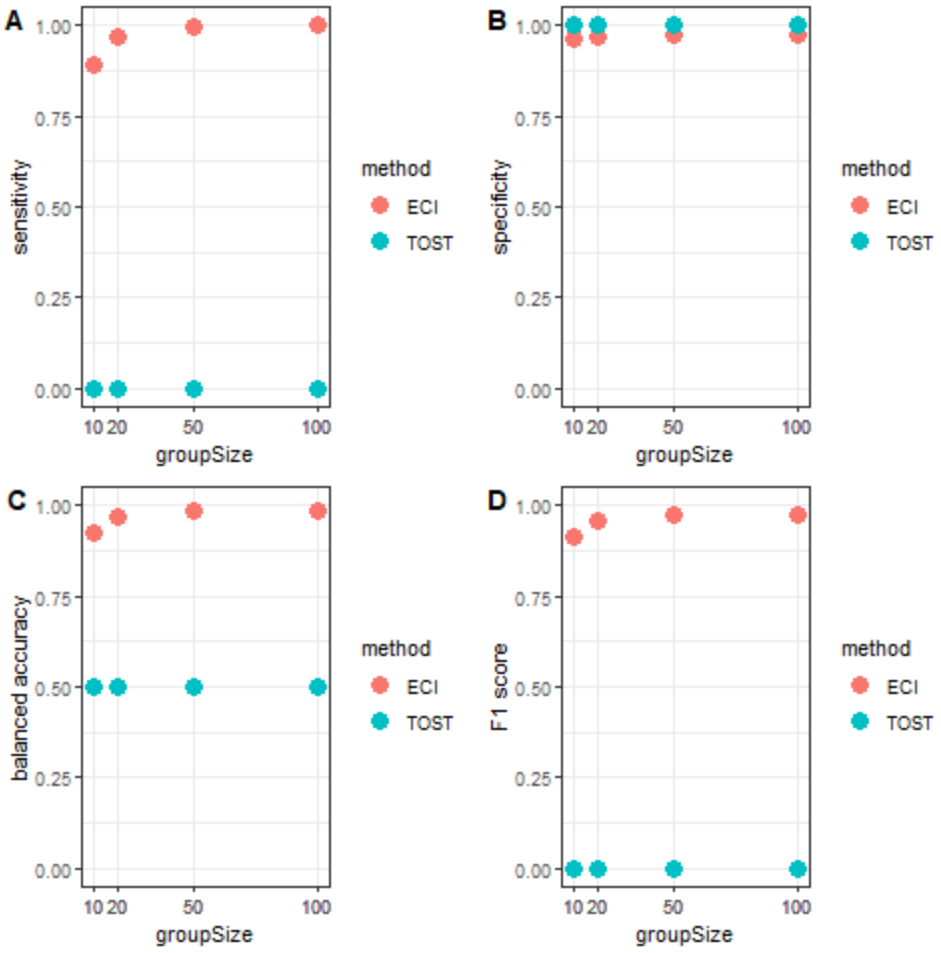
Figure 1: Example plots of decisions based on the q-value by the two tests of one simulation iteration.



Legend: Left is the significance decision of the ECI bootstrap test, right is the significance decision of the of TOST.

We repeated the previously described simulation 1000 times and calculated the average of sensitivity, specificity, balanced accuracy, and F1 score for different sample sizes of the comparator groups, namely 10, 20, 50 and 100 (Figure 2). As can be seen, the sensitivity, balanced accuracy and F1 score for TOST are close to zero irrespective of group size and the specificity is close to 1 for all group sizes. The ECI bootstrap test shows that all performance metrics are dependent on the group size where a small group size leads to worse performance and a large group size enhances performance. Nevertheless, the performance of the ECI bootstrap test is overall very good. For a group size of 10, the balanced accuracy is at average 0.938 and the F1 score is at average 0.922 (Supplemental table T1).

Figure 2: Performance metrics for ECI bootstrap test and two one-sided t-test



Legend: A) sensitivity, B) specificity), C) balanced accuracy, D) F1 score. The ECI bootstrap test shows increasing performance with increasing group size and has overall a good performance. The TOST shows only in specificity a better performance than the ECI bootstrap test and shows no performance for all the other measurements.

Biological Data

We used 3 publicly available data sets that were created to study pancreatic cancer. Two of the datasets contained gene expression measurements in pancreatic tumor tissue with adjacent normal tissue as control (DT1 and DT2) and the third data set contained gene expression measurements from peripheral blood mononuclear cells (PBMCs) with gene expression data from gender, age, and habits matched healthy patients as controls (hereafter referred to as the PBMC data set).

We performed both bootstrap ECI tests and TOST on the biological data. Using TOST, none of the genes were significantly equivalently changed between any of the studies. While using bootstrap ECI, the comparison of the two tumor tissue data sets led to the identification of 74 genes which were equivalently changed (Supplemental table T2). Using the DO database, we found no associated disease description with the identified genes. When we use the DGN database, multiple disease descriptions were related to cancer such as gastric adenocarcinoma (p-adj = 0.0474), or noninfiltrating intraductal carcinoma (p-adj =0.0297) (Supplemental table T5). Belonging to the top 10 equivalently changed genes are genes with known or investigated tumor relevance such as ASAP2 (17), NRP2 (18), and MUC13 (19), which are being investigated as drug targets for pancreatic cancer, and FAT1 (20), which is a known tumor suppressor.

We identified 203 equivalently changed genes and 107 inversely changed genes by comparing tumor tissue data set 1 with the PBMC data set (Supplemental table T3). Some of the identified genes are associated with pancreatic cancer (p-adj =8.3e-03), pancreatic carcinoma (p-adj =9.6e-03), and pancreatic ductal adenocarcinoma (p-adj =9.6e-03) when using the DO database (Supplemental table T6). When using the DGN database, multiple disease descriptions associated with pancreatic cancer were identified such as pancreatic ductal adenocarcinoma (p-value = 1.3e-07), and adenocarcinoma of pancreas (p-value = 8.8e-04) (Supplemental table T7).

The comparison of tumor tissue data set 2 with the data set from PBMC identified 8 equivalently changed genes and 7 inversely changed genes. Using the DO database as well as the DGN database resulted in no disease descriptions found to be associated with the equivalently changed.

Discussion

In this work, we compared two options to test for equivalently changed genes between two studies, namely the proposed ECI bootstrap test and TOST. We were able to show that the ECI bootstrap

performs well in identifying equivalently changed genes with respect to balanced accuracy and F1, which were both close to 0.9, while maintaining a high specificity. Furthermore, we found that TOST greatly underperformed the ECI bootstrap test with balanced accuracy close to 0.5 and F1 score close to 0. The reason for TOST's performance is related to the pooled variance of the effect sizes. We used standard deviations from a real-world study and a margin of error $\Delta = 0.25$ and performed a t-test with three degrees of freedom, therefore, the t-test statistics have to exceed an absolute value of 2.353 to reject the null hypothesis. For a case of perfect equivalence, the test statistic for non-inferiority would be $\frac{-0.25}{s} < -2.353$. The pooled standard deviation must be smaller than 0.106 to be able to reject the null hypothesis in the case of perfect equivalence but, as shown in Supplemental figure F2, this is not the case for any of the pooled standard deviations of the tested genes.

By using a threshold of 0.05 for the q-value we were able to identify several equivalently changed genes between two pancreatic cancer tumor tissue studies where most of the genes were related to cancer progression. This shows that the ECI bootstrap test met our expectation of identifying equivalently changed genes in studies of the same disease type and, furthermore, is able to identify equivalently changed genes which are functionally related to the disease type and could lead to more reproducible or robust results. Additionally, we were able to show that several genes in non-cancer tissue (peripheral blood) of patients of pancreatic cancer showed equivalently changed behavior to genes of tumor tissue of the same disease which implies the systemic impact of cancer. Those results may open the way for identifying reliable blood markers for cancer, and for new investigative approaches into the field of systemic changes of gene expression of cancer patients that may be addressed in future studies. It must be mentioned that the difference in the number of identified genes could be due to the group sizes of the data sets.

Limitations of this study include the differing nature of cases and controls in some datasets, the different group sizes, and the lack of a gold standard in biological data.

Conclusion

In this study we demonstrated the use of the ECI bootstrap test in a setting of differentially expressed genes to provide researchers with a statistical approach to identify genes which are similarly influenced by a specific treatment or exposure. Furthermore, statistically identified equivalently changed genes reduces the cost for validating those genes and offers the option of identifying new possible treatment targets.

In addition, it is also possible, due to the non-parametric nature of the bootstrap test and the lack of assumptions on the ECI value, to adapt the ECI bootstrap test to other options such as methylation data or other types of 'omics data. In future studies we want to investigate the effectiveness of bootstrap ECI on other types of data sets.

Methods

In the following, we give a definition of the Equivalent Change Index (ECI), the ECI bootstrap test, and the Two One-Sided t-Tests (TOST). Furthermore, we explain the design of the simulation study and give information about the publicly available data sets used in our real data analysis. For the purpose of this study, we need an effect size $\hat{\beta}_i$, the standard deviation (TOST) of $\hat{\beta}_i$, and a measurement of statistical significance for $\hat{\beta}_i$ (ECI bootstrap test) for each differentially expressed gene i. We decided to use the log2 foldchange (log2 FC) as effect size, although other measures could also be used, for example the standardized mean difference (SMD).

Equivalent Change Index (ECI)

The Equivalent Change Index (ECI), proposed by Thompson and Koestler (16), is a measure of the degree of equivalent or inverse change of attributes of the same type across two diverse studies. The ECI λ_i of a gene i is calculated as a ratio of the minimum and maximum of the absolute effect sizes $\hat{\beta}_{ik}$ from the two studies ($k = [1,2]$) multiplied by a sign, where the sign reflects whether the differential gene expression of the two studies was in the same direction (positive sign) or opposite direction (negative sign). Furthermore, the ECI is weighted by the maximum of the p-values p_{ik} of the two effect sizes.

$$\lambda_i = \frac{\mathrm{sgn}(\hat{\beta}_{i1} \times \hat{\beta}_{i2})\min(|\hat{\beta}_{i1}|, |\hat{\beta}_{i2}|)}{\max(|\hat{\beta}_{i1}|, |\hat{\beta}_{i2}|)} \cdot (1 - \max(p_{i1}, p_{i2}))$$

λ_i is in the range of [-1,1], where -1 indicates that the effect size was exactly opposite between the two studies and 1 indicates that the effect sizes between the two studies were identical. Hence, λ_i indicates the degree of equivalence or inverseness of the expression of a gene compared between two separate experiments.

ECI bootstrap test

In this section we describe our proposed bootstrap procedure to use the ECI statistic to test for equivalence. We first calculate the ECI values for all genes using the function getECI() from the R package ECEA (16) with the log2 FC as effect size and corresponding p-values as measurement of statistical significance (see Figure 3). To obtain a test measure for equivalent change, we proceed as follows:

- create bootstrap samples within treatment and control groups for each study
- recalculate the differential gene expression for each study separately, e.g., the log2 FC for each gene along with its p-value

- recalculate ECI values between studies for each gene

We repeated this procedure 1000 times.

Confidence Interval

We used a 95% bias-corrected and accelerated (BCa) bootstrap interval as a confidence interval for the ECI values since it corrects for both bias and skewness (21, 22). The BCa assumes that the data are independent and identically distributed and is calculated as $\mathrm{CI}\left[\hat{\beta}_{\mathrm{L}}^{*}, \hat{\beta}_{\mathrm{U}}^{*}\right]$, where $\hat{\beta}_{\mathrm{L}}^{*}$ denotes the L^{th} quantile and $\hat{\beta}_{\mathrm{U}}^{*}$ denotes the U^{th} quantile, i.e. lower limit and upper limit. The indices j and k are defined as $\mathrm{L} = \mathrm{a}_1 \cdot \mathrm{B}$ and $\mathrm{U} = \mathrm{a}_2 \cdot \mathrm{B}$, where B is the number of bootstrap samples, here $\mathrm{B} = 1000$, and a_1 and a_2 are defined as:

$$\mathrm{a}_1 = \Phi\left(\hat{\mathrm{z}}_0 + \frac{\hat{\mathrm{z}}_0 + \mathrm{z}^{(\alpha/2)}}{1 - \hat{\mathrm{a}}(\hat{\mathrm{z}}_0 + \mathrm{z}^{(\alpha/2)})}\right)$$

$$\mathrm{a}_2 = \Phi\left(\hat{\mathrm{z}}_0 + \frac{\hat{\mathrm{z}}_0 + \mathrm{z}^{(1-\alpha/2)}}{1 - \hat{\mathrm{a}}(\hat{\mathrm{z}}_0 + \mathrm{z}^{(1-\alpha/2)})}\right)$$

Here, Φ is the standard normal cumulative distribution function, $\mathrm{z}^{(\alpha/2)}$ is the $100\mathrm{x}\left(\frac{\alpha}{2}\right)$th percentile of a standard normal distribution, and $\hat{\mathrm{z}}_0$ and $\hat{\mathrm{a}}$ are the bias-corrections and acceleration factors, respectively. The bias-correction factor is computed as the inverse of the standard normal cumulative distribution function of the proportion of bootstrap effect sizes $\hat{\beta}^{*}$ smaller than the original effect size $\hat{\beta}$:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I_{[\hat{\theta}^{*}<\hat{\theta}]}\right)$$

The acceleration factor uses “leave one out” (jackknife) resampling as follows:

$$\hat{\mathrm{a}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\hat{\beta}_{(.)} - \hat{\beta}_{(-\mathrm{i})}\right)^3}{6\left[\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\hat{\beta}_{(.)} - \hat{\beta}_{(-\mathrm{i})}\right)^2\right]^{3/2}}$$

Where, $\hat{\beta}_{(-i)}$ is an effect size of a jackknife sample and $\hat{\beta}_{(.)}$ is the average of the effect sizes of all jackknife samples.

We used the bias corrected confidence interval to test the null hypothesis that the ECI value is not different from zero, in other words, there is no equivalent or inverse change between the effect sizes of the attributes of the two studies. We reject the null hypothesis when the confidence interval of the ECI values does not include zero.

P-value (approximate)

According to Asparouhov and Muth´en (23), assuming that $\hat{a} = 0$, we have

$$a_1 = \Phi\left(\hat{z}_0 + \hat{z}_0 + z^{(\alpha/2)}\right) = \Phi\left(2\hat{z}_0 + z^{(\alpha/2)}\right)$$

$$a_2 = \Phi\left(\hat{z}_0 + \hat{z}_0 + z^{(1-\alpha/2)}\right) = \Phi\left(2\hat{z}_0 + z^{(1-\alpha/2)}\right)$$

When we concentrate only on positive ECI values, we would reject the null hypothesis of non-equivalence when $\hat{\beta}^*_L > 0$ or L > M where

$$M = \sum_{b=1}^{B} I_{[\hat{\theta}^* < 0]}$$

In the special case of p-value = 0.05 (L = M) we have

$$0.5 = 2 \cdot 0.025 = 2 \cdot \Phi(-1.96)$$

$$= 2 \cdot \Phi(-2\hat{z}_0 + 2\hat{z}_0 - 1.96)$$

$$= 2 \cdot \Phi\left(-2\hat{z}_0 + \Phi^{-1}\left(\Phi(2\hat{z}_0 - 1.96)\right)\right)$$

$$= 2 \cdot \Phi\left(-2\hat{z}_0 + \Phi^{-1}(a_1 \cdot B/B)\right)$$

$$= 2 \cdot \Phi\left(-2\hat{z}_0 + \Phi^{-1}(L/B)\right)$$

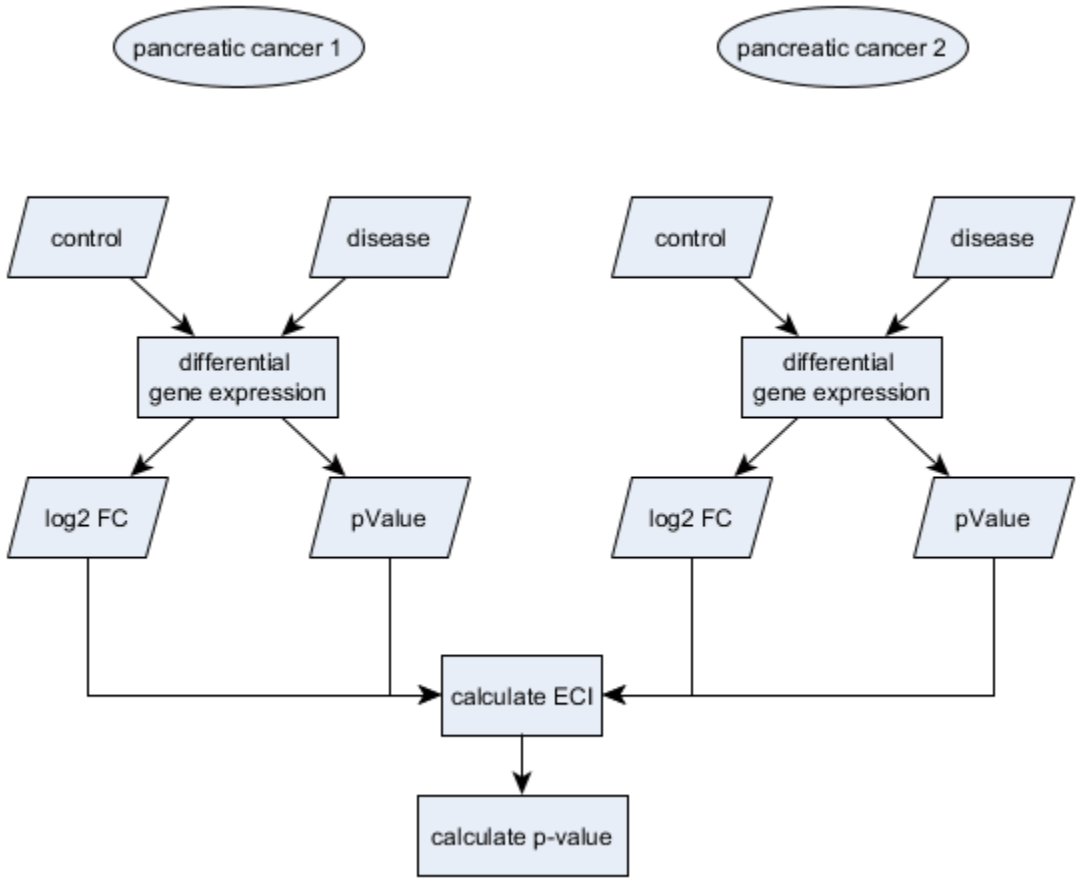
Therefore, the two-sided p-value can be computed, approximately, by

$$p - value = 2 \cdot \Phi\left(-2\hat{z}_0 + \Phi^{-1}(M/B)\right)$$

where M is the number of ECI values, which are smaller than zero. In case of the ECI value being smaller than zero we compute

$$p - value = 1 - 2 \cdot \Phi\left(-2\hat{z}_0 + \Phi^{-1}(M/B)\right)$$

Figure 3: Workflow to compare the differential gene expression of genes of two different studies using the Equivalent Change Index (ECI).



Two one-sided t-tests

What follows is a brief description of how the TOST method was applied to these data. Let Δ represent some fold change which is considered to be unimportant. To test for equivalent change, we use a null hypothesis of non-equivalence versus an alternative hypothesis of equivalence or inverseness. In particular, we must perform two hypothesis tests, namely:

$$H_{01}: \beta_{i1} - \beta_{i2} \geq \Delta \text{ vs } H_{a1}: \beta_{i1} - \beta_{i2} < \Delta$$

and,

$$H_{02}: \beta_{i1} - \beta_{i2} \leq -\Delta \text{ vs } H_{a2}: \beta_{i1} - \beta_{i2} > \Delta$$

where β_{ik} is the effect size of gene i of study k. The t-test statistics for the two tests are:

$$t_i^{*1} = \frac{\hat{\beta}_{i1} - \hat{\beta}_{i2} - \Delta}{s_i}$$

and,

$$t_i^{*2} = \frac{\hat{\beta}_{i1} - \hat{\beta}_{i2} + \Delta}{s_i}$$

Here, s_i is the pooled standard deviation of gene i of the two studies. Only when the p-value for test one and the p-value for test two are smaller than the significance level of $\alpha = 0.05$, is a gene considered to be equivalently changed. Likewise, to test for inverse change, we use the same test statistics except that we multiply $\hat{\beta}_{i1}$ by -1. To perform the t-tests, we used 3 degrees of freedom since we are performing the tests on 4 groups, namely control and treatment groups from one of the studies (indicated as study 1) represented by the effect size $\hat{\beta}_{i1}$ and control and treatment groups from the other study (indicated as study 2) represented by the effect size $\hat{\beta}_{i2}$.

Multiple testing

We performed a test of equivalence for each gene. To adjust the p-value for the multiple testing we used the adjusted false discovery rate (FDR) approach by Benjamini and Hochberg (24), which can be used to test how many of the accepted results are false. Here, the FDR correction $q_i = p_i N/i$, where p_i is the ith p-value in a sorted list of ascending p-values and N is the total number of p-values, is the ratio of expected false positives and the total number of accepted positives. To adjust for the non-monotony of the q_i value we replace the q_i value with the lowest q_i value among all q_f, where $f \geq i$.

Simulation

We conducted a simulation study to compare the two different tests for equivalent change. This simulation study consisted of two simulated studies of differential gene expression, for which we aimed to test equivalent change of gene expression. The two simulated studies are constructed so that 30% of the differentially expressed genes of simulated study 1 are equivalently changed between the two studies. The simulation process has two levels. First, we simulated descriptive features of each simulated study (mean and standard deviation), which we used in the second level to draw random gene expression values for each sample. We then used the simulated studies to performed differential gene expression and the equivalence testing as described above. In the following we will explain how each study is set up step by step.

Simulation

1. *Simulated study 1.* The first stage is to simulate a study with two groups. The group is denoted by k, with controls as $k = 0$ and cases or treatment as $k = 1$. The simulation derives certain values from a reference study of pancreatic cancer (see section Biological Data) with case control data, including samples that were either tumor or tumor adjacent normal tissue (GSE16515), as described below.

a. The mean expression for gene g in the simulated control group $(k = 0)$ is drawn from a gamma distribution with scale and shape parameters extracted from the reference dataset using the function egamma() of the R package EnvStats (25):

$$m_g \sim \Gamma(\alpha = 8.24, \beta = 0.66)$$

b. For each gene g from the reference dataset we obtained the mean difference in expression, denoted by δ_g.

c. Genes with $-1 \geq \delta_g \leq 1$ are removed. For this simulation we are only interested in genes with a difference between groups.

d. From the filtered down gene set we obtained a new gene set by sampling with replacement, where each gene g has:

   i. The standard deviation for gene g in group k of the reference dataset, denoted by s_{gk}

   ii. The mean difference δ_g

e. The genes were divided into three subgroups: equivalently changed genes (f=1), non-equivalently changed genes 1 (f=2), and non-equivalently changed genes 2 (f=3). The non-equivalently changed genes were divided into two subgroups to have the differential expression of those genes be balanced between the two studies: One half of the genes is differentially expressed in one study while non-differentially expressed in the other study and vice versa.

f. The expression value for gene g, observation i, group k is denoted x_{gik} and is drawn from a truncated normal distribution:

$$x_{gik} \sim N(\gamma_g, s_{gk}); 0 < x_{gik}$$

With

$$\gamma_g = \begin{cases} m_g + k\delta_g & if\ f = 1 \\ m_g & if\ f = 2 \\ m_g + k\delta_g & if\ f = 3 \end{cases}$$

g. With the simulated control group and treatment group, we calculated the differential expression for each gene as described in section "Differential Gene Expression".

2. *Simulated study 2.* The next stage is to simulate gene expression from a second, similar study. At this point, we will determine genes which are equivalently changed across the two studies. We will choose 30% of the differentially expressed genes (abs(log2FC) >1 and p-value < 0.05) from simulated study 1 to be equivalently changed.
    a. The mean expression for a gene in the simulated control group is equal to the mean expression in simulated study 1.
    b. Equivalently changed genes are further simulated to not always be perfectly equivalently changed (even on average). This is done by having a modifier for the change in gene g, denoted θ_g and is in [1,2.5].
    c. For simulated study 2, the expression value for gene g, observation j, group k is then drawn from a truncated normal distribution:

$$y_{gik} \sim N(\gamma_g, s_{gk}); 0 < y_{gik}$$

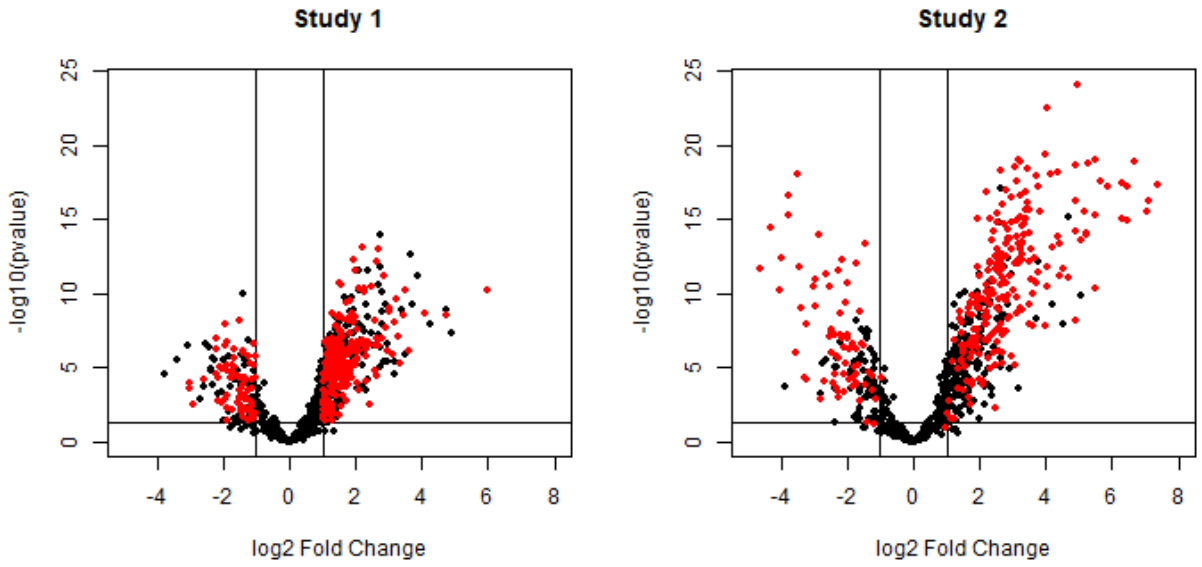
With

$$\gamma_g = \begin{cases} m_g + k\theta_g\delta_g & if\ f = 1 \\ m_g + k\delta_g & if\ f = 2 \\ m_g & if\ f = 3 \end{cases}$$

We used the pairing of s_{gk} and δ_g to be able to simulate the fact that differentially expressed genes with low difference in means between the treatment groups often have small standard deviations.

In Figure 4 can be seen the distribution of differential gene expression of the two simulated studies of one iteration of the simulation. The degree of equivalence of change in gene expression between the two studies for one iteration of the simulation is visualized in Supplemental Figure F1.

Figure 4: Example volcano plots of the two simulated gene expression studies of one simulation iteration.



Legend: Marked red are all genes set to be equivalent between the two studies. Here, the differential expression of the genes to be set equivalently changed between the two simulated studies must pass a threshold (abs(log2FC) >1 and p-value < 0.05) in simulated study 1 but the differential expression of the same genes is altered in simulated study 2 to mirror equivalent change and thus can fail the threshold for differential gene expression.

Performance Measures

To assess the performance of each test we used sensitivity, specificity, balanced accuracy and F1 score, which is the harmonic mean of precision and sensitivity. Furthermore, we tested the performance of the ECI bootstrap test when we impose thresholds such that only when the absolute ECI value exceeds the

threshold is it considered significant. We set the thresholds in a range from 0 to 0.9 in steps of 0.1 and investigated the performance according to the F1 score using a study set up with equivalently changed genes specified and the false positive rate (FPR) using a study set up with no equivalently changed genes specified.

Biological Data

The biological data analysis is based on publicly available microarray expression data of pancreatic cancer from peripheral blood mononuclear cells (PBMC) (Accession # GSE74629) and tumor tissue (Accession # GSE16515 and GSE22780) (26-28) available from the NCBI-GEO database (29, 30) (see Table 2). The first pancreatic tumor tissue data set (DT1) has 16 samples in the control group and 36 samples in the disease group. The second pancreatic data set (DT2) only has 8 samples in both the treatment and the control group. The data set from PBMC has similar group sizes as DT1 (control:14, disease: 36). The control data for the PBMC study were from gender, age, and habit matched healthy participants and the control data for the two disease tissue studies are from adjacent normal tissue of the cancer patients. After reading in the raw files, the expression data were normalized using the package affy (31) for Affymetrix microarray data, or the package limma (32) for Illumina and Agilent microarray data in the R statistical environment.

Table 2: Specifics for the different pancreatic cancer studies used in this project.

Tissue type	**Accession #**	**# tumor**	**# control**
Disease	GSE16515	36	16
Disease	GSE22780	8	8
PBMC	GSE74629	36	14

We expect to see equivalently changed genes, the majority of them changed in similar manners, between those studies because they are from the same type of disease.

Differential Gene Expression

We used the package limma to perform differential gene expression analysis. The model was fitted by using the functions lmfit() and eBayes(), which uses moderated t-statistics for ranking the genes. The standard deviation (sd) for gene g was be extracted from the output by multiplying the square root of the unscaled covariance ϑ_{gj} and the posterior residual variance $\tilde{s}_g^2$, $\mathrm{sd} = \sqrt{\vartheta_{gj}} \cdot \tilde{s}_g$ as recommended by GK Smyth (33).

Disease Enrichment Analysis

To investigate the association of equivalently changed genes with disease types we used the R package DOSE (34), which calculates a p-value using the hypergeometric distribution to determine whether the number of genes grouped to be associated with a disease is larger than expected. We used the function enrichDO(), which supports Disease Ontology (DO) data (35), and enrichDGN, which supports DisGeNET (36), with a minGSSize of 5 and a q-value cut off of 0.05.

List of abbreviations

- TOST – Two One-Sided t-Tests
- ECI - Equivalent Change Index
- DO - Disease Ontology
- DGN – DisGeNET
- DT1 - first pancreatic tumor tissue data set
- DT2 - second pancreatic tumor tissue data set

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets analyzed during the current study are available in the NCBI repository, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16515, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22780, and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74629.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by a CTSA grant from NCATS awarded to the University of Kansas for Frontiers: University of Kansas Clinical and Translational Science Institute (# UL1TR002366) and used the Quantitative 'Omics Core (QOC) and the National Cancer Institute (NCI) Cancer Center Support Grant P30CA168524, the Kansas IDeA Network of Biomedical Research Excellence Bioinformatics Core, supported by the National Institute of General Medical Science award P20 GM103418, and the Kansas Institute for Precision Medicine COBRE, supported by the National Institute of General Medical Science award P20 GM130423.

The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NCATS. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contribution

LN and JT created the new ECI bootstrap test and developed the design of the simulation study. LN analyzed the data and wrote the manuscript. JT interpreted the biological data. All authors read, reviewed, and approved the final manuscript.

Acknowledgements

We acknowledge the contribution of Mihaela Sardiu, Lynn Chollet Hinton, Nanda Kumar Yellapu, Emily Nissen, Whitney Shae, and Jonah Aponsah in reviewing the paper.

References

1. Gilbert S. Developmental Biology. 6th edition. Sunderland (MA): Sinauer Associates; 2000.

2. Fu Y, Thomas A, Gasior D, Harper J, Gay A, Jones C, et al. A comparison of shared patterns of differential gene expression and gene ontologies in response to water-stress in roots and leaves of four diverse genotypes of Lolium and Festuca spp. temperate pasture grasses. Plos One. 2021;16(4):e0249636.

3. Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, et al. A comparison of gene expression and DNA methylation patterns across tissues and species. Genome Res. 2020;30(2):250-62.

4. Shen Y, Wolkowicz MJ, Kotova T, Fan L, Timko MP. Transcriptome sequencing reveals e-cigarette vapor and mainstream-smoke from tobacco cigarettes activate different gene expression profiles in human bronchial epithelial cells. Sci Rep. 2016;6:23984.

5. Hollenbach PW, Nguyen AN, Brady H, Williams M, Ning Y, Richard N, et al. A comparison of azacitidine and decitabine activities in acute myeloid leukemia cell lines. Plos One. 2010;5(2):e9001.
6. Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends Biotechnol. 2017;35(6):498-507.
7. Provenzano M, Mocellin S. Complementary techniques: validation of gene expression data by quantitative real time PCR. Adv Exp Med Biol. 2007;593:66-73.
8. Garrido J, Aguilar M, Prieto P. Identification and validation of reference genes for RT-qPCR normalization in wheat meiosis. Sci Rep. 2020;10(1):2726.
9. Rudy J, Valafar F. Empirical comparison of cross-platform normalization methods for gene expression data. BMC Bioinformatics. 2011;12:467.
10. Walsh CJ, Hu P, Batt J, Santos CC. Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. Microarrays (Basel). 2015;4(3):389-406.
11. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15(6):657-80.
12. Dixon PM, Saint-Maurice PF, Kim Y, Hibbing P, Bai Y, Welk GJ. A Primer on the Use of Equivalence Testing for Evaluating Measurement Agreement. Med Sci Sport Exer. 2018;50(4):837-45.
13. Leichsenring F, Abbass A, Driessen E, Hilsenroth M, Luyten P, Rabung S, et al. Equivalence and non-inferiority testing in psychotherapy research. Psychol Med. 2018;48(11):1917-9.
14. Wu LJ, Gander PH, van den Berg M, Signal TL. Equivalence Testing as a Tool for Fatigue Risk Management in Aviation. Aerosp Med Hum Perf. 2018;89(4):383-8.
15. Qiu J, Cui X. Evaluation of a statistical equivalence test applied to microarray data. J Biopharm Stat. 2010;20(2):240-66.
16. Thompson JA, Koestler DC. Equivalent change enrichment analysis: assessing equivalent and inverse change in biological pathways between diverse experiments. Bmc Genomics. 2020;21(1).

17. Fujii A, Masuda T, Iwata M, Tobo T, Wakiyama H, Koike K, et al. The novel driver gene ASAP2 is a potential druggable target in pancreatic cancer. Cancer Sci. 2021;112(4):1655-68.

18. Wang L, Wang L, Wang S, Zhou Z, Liu Z, Xu P, et al. N2E4, a Monoclonal Antibody Targeting Neuropilin-2, Inhibits Tumor Growth and Metastasis in Pancreatic Ductal Adenocarcinoma via Suppressing FAK/Erk/HIF-1alpha Signaling. Front Oncol. 2021;11:657008.

19. Nishii Y, Yamaguchi M, Kimura Y, Hasegawa T, Aburatani H, Uchida H, et al. A newly developed anti-Mucin 13 monoclonal antibody targets pancreatic ductal adenocarcinoma cells. Int J Oncol. 2015;46(4):1781-7.

20. Peng Z, Gong Y, Liang X. Role of FAT1 in health and disease. Oncol Lett. 2021;21(5):398.

21. Jung K, Lee J, Gupta V, Cho G. Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation. Front Psychol. 2019;10.

22. Wicklin R. The bias-corrected and accelerated (BCa) bootstrap interval: SAS blogs; 2017 [Available from: https://blogs.sas.com/content/iml/2017/07/12/bootstrap-bca-interval.html.

23. Tihomir Asparouhov aBMe. Bootstrap P-value Computation 2021 [Available from: http://www.statmodel.com/download/FAQ-Bootstrap%20-%20Pvalue.pdf.

24. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J R Stat Soc B. 1995;57(1):289-300.

25. Millard SP. EnvStats: An R Package for Environmental Statistics}. New York: Springer; 2013.

26. Pei H, Li L, Fridley BL, Jenkins GD, Kalari KR, Lingle W, et al. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. Cancer Cell. 2009;16(3):259-66.

27. Li L, Zhang JW, Jenkins G, Xie F, Carlson EE, Fridley BL, et al. Genetic variations associated with gemcitabine treatment outcome in pancreatic cancer. Pharmacogenet Genomics. 2016;26(12):527-37.

28. Ellsworth KA, Eckloff BW, Li L, Moon I, Fridley BL, Jenkins GD, et al. Contribution of FKBP5 genetic variation to gemcitabine treatment and survival in pancreatic adenocarcinoma. Plos One. 2013;8(8):e70216.

29. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207-10.

30. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-5.

31. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15.

32. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

33. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.

34. Yu G, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015;31(4):608-9.

35. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40(Database issue):D940-6.

36. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database-Oxford. 2015.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaldatanew.docx