Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 127–135

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Original Research

CDS: A Fold-change Based Statistical Test for Concomitant Identification of Distinctness and Similarity in Gene Expression Analysis

Nicolas Tchitchek [1], José Felipe Golib Dzib [1], Brice Targat [1], Sebastian Noth [1], Arndt Benecke [1,2,*], Annick Lesne [1,3]

[1] Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette 91440, France
[2] Centre National de la Recherche Scientifique, USR3078, Bures-sur-Yvette 91440, France
[3] Laboratoire de Physique Théorique de la Matière Condensée, CNRS UMR7600, Université Pierre et Marie Curie-Paris 6, Paris 75005, France

Received 9 May 2012; revised 8 June 2012; accepted 10 June 2012
Available online 25 June 2012

Abstract

The problem of identifying differential activity such as in gene expression is a major defeat in biostatistics and bioinformatics. Equally important, however much less frequently studied, is the question of similar activity from one biological condition to another. The fold-change, or ratio, is usually considered a relevant criterion for stating difference and similarity between measurements. Importantly, no statistical method for concomitant evaluation of similarity and distinctness currently exists for biological applications. Modern microarray, digital PCR (dPCR), and Next-Generation Sequencing (NGS) technologies frequently provide a means of coefficient of variation estimation for individual measurements. Using fold-change, and by making the assumption that measurements are normally distributed with known variances, we designed a novel statistical test that allows us to detect concomitantly, thus using the same formalism, differentially and similarly expressed genes (http://cds.ihes.fr). Given two sets of gene measurements in different biological conditions, the probabilities of making type I and type II errors in stating that a gene is differentially or similarly expressed from one condition to the other can be calculated. Furthermore, a confidence interval for the fold-change can be delineated. Finally, we demonstrate that the assumption of normality can be relaxed to consider arbitrary distributions numerically. The Concomitant evaluation of Distinctness and Similarity (CDS) statistical test correctly estimates similarities and differences between measurements of gene expression. The implementation, being time and memory efficient, allows the use of the CDS test in high-throughput data analysis such as microarray, dPCR, and NGS experiments. Importantly, the CDS test can be applied to the comparison of single measurements ($N = 1$) provided the variance (or coefficient of variation) of the signals is known, making CDS a valuable tool also in biomedical analysis where typically a single measurement per subject is available.

Keywords: Statistical test; Fold-change; Distinctness; Similarity; Gene expression; Single measurement; Patient study

Introduction

The problem of identifying differentially expressed genes has been widely studied [1]. Considering two different biological conditions, one aims to decide which genes are differentially expressed from one biological condition to the other, each composed of one or several gene expression measurements. RNA quantification, which is being used in transcriptome analysis here will serve as an instance representative of any type of high-throughput quantification of cellular components such as DNA, RNA, protein, or metabolites, as the underlying problem of identifying statistically significant changes remains similar independent of the nature of the experiment. Therefore, all of what follows similarly applies to proteome or other measurements.

* Corresponding author.
E-mail: arndt@ihesfr (Benecke A).

1672-0229/$ - see front matter
http://dx.doi.org/10.1016/j.gpb.2012.06.002

For the sake of simplicity, we will only continue to discuss the case of gene expression investigations. First attempts to tackle the question of differential quantities did not involve statistics and genes having expression levels differing by more than an arbitrary cut-off fold-change value were considered to be differentially expressed [2,3]. Although the identification of statistically differentially expressed genes has been widely covered [1], the identification of similarly expressed genes has been far less studied. This is surprising, for several reasons. (i) Statistical measures for similarity are an important tool in establishing reproducibility and thus track technical and biological variation. (ii) In relative quantification, such as microarray experiments, where no absolute numbers of, *e.g.*, transcripts is established, a defining procedure for what is considered similar, or unchanged, expression would in turn also provide a sound basis for defining what is to be considered different. (iii) Finally, especially in the case of biomedical studies on human subjects and patients, the question of genes with conserved expression across different biological conditions is of similar importance to the one of change [4].

When reasoning in a statistical manner, assumptions can generally be made that gene expression measurements are normally distributed. The simplest statistical method for detecting differentially expressed genes is the two-sample *t*-test [5]. The two-sample *t*-test allows us to formulate statements concerning the difference between the means of two normally distributed variables with the assumption that the variances are unknown. On the other hand, the two-sample *z*-test allows us to formulate statements concerning the difference between the means of two normally distributed variables with the assumption that the variances are known. However as this assumption can only be made with a large sample of independent records or with additional information about the variances, the two-sample *t*-test is more often used in the identification of differentially expressed genes. Different variants of the two-sample *t*-test can be classified in two groups: (i) methods such as the two-sample *t*-test with relative thresholds [6] carrying out local adjustments to account for biologically meaningful differences, and the Significance Analysis of Microarrays method [7] that uses a gene-specific correction; and (ii) jointly global and local methods such as the B-statistic [8] and the regularized two-sample *t*-test [9]. In addition to simple fold-change or *t*-test-like methods, another approach is to consider the statistical properties of the ratio of means of the two biological conditions sampled. Based on the previous work [10], Chapman [11] proposed for the first time a statistical test in this direction. Recent methods (*e.g.*, [12,13]) extended this approach by considering confidence intervals for the statistic of the ratio of the two means used in hypothesis testing. When comparing different methods for differential expression detection, among the desirable characteristics that a method should have are reproducibility and control of type I and type II errors. Not all of the existing methods necessarily combine both characteristics [14]. Another way of comparing different methods is to measure their false positive and false negative rates [15].

Assume two sets of gene expression measurements obtained from two different biological conditions (**Figure 1**). By initially making the assumptions that the gene measurements are normally distributed with known variances, we represent the fold-change as the tangent of θ in Figure 1A. Having two biological conditions we can expect different scenarios. If both biological conditions have a small variance within biological replicates and then show differential expression, then methods should detect them as signifi-

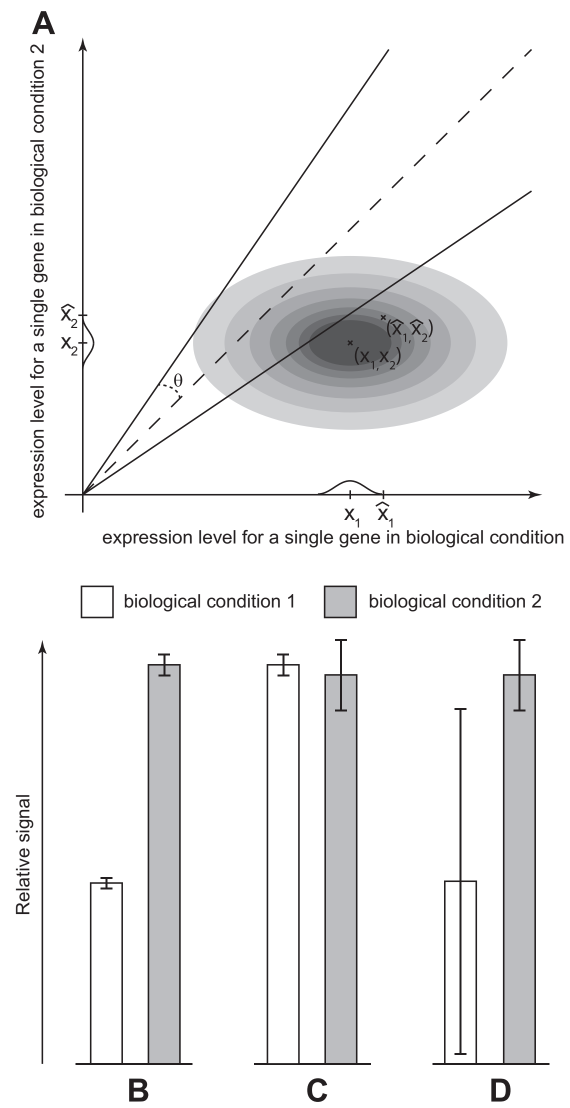

Figure 1 Graphical representation of the problematic and encountered scenarios

A. Expression signals of a single gene in two different biological conditions, with normal distributions having the parameters x_1 and x_2 (mean values) and σ_1 and σ_2 (variances). The fold-change criteria defining the difference or similarly is represented with a conic section defined by parameter θ. The problem is to determine the value of (x_1, x_2) having the values of estimators $(\widehat{x_1}, \widehat{x_2})$. **B.** Potential scenario for the statistical test for differential expression and low variability. **C.** Potential scenario of having low variability and similarly expressed genes. **D.** Potential scenario for the statistical test for no statistical significance and high variabilities.

cantly statistically differentially expressed (Figure 1B). Ideally, the same metric would provide for detecting similar expression across biological conditions (Figure 1C) when they present small variability. However, when variability is high, methods should indicate no statistical significance neither for similarity nor for difference (Figure 1D).

We describe here a statistical test, CDS for Concomitant identification of Distinctness and Similarity, which allows: (i) obtaining statements on the fold-change rather than on the difference between the mean expression levels; (ii) providing an estimate of the variance together with the signal; (iii) obtaining bounds on the fold-change, both in case of differentially expressed genes and similarly expressed genes. CDS can thereby be used for single measurements of biological conditions ($N = 1$), provided an estimate of the variance is available.

Statistical approach

Test formulation

Let X be a random variable following the given distribution D_x with unknown parameter x, and let $\hat{x}$ be an estimator of the parameter x from a sample of independent observations of X. Let H_0 be a null hypothesis and H_A an alternative hypothesis, and let R_0 and R_A be two regions (we use the term region as a synonym of set), such as:

$$\begin{cases} H_0 : x \in R_0 \\ H_A : x \in R_A \end{cases}$$

Let $\widehat{R}_0$ be the rejection region of H_0 such that H_0 is rejected if and only if (iff) $\hat{x} \in \widehat{R}_0$, and let $\widehat{R}_A$ be the rejection region of H_A such that H_A is rejected iff $\hat{x} \in \widehat{R}_A$.

The probability of type I error, which is the probability of making an error of rejecting the null hypothesis H_0 when it is actually true, is then defined by:

$$Prob(H_0 \text{ rejected}|H_0 \text{ true}) \Longleftrightarrow Prob(\hat{x} \in \widehat{R}_0|x \in R_0)$$

The probability of type II error, which is the probability of making an error of rejecting the alternative hypothesis H_A when it is actually true, is then defined by:

$$Prob(H_A \text{ rejected}|H_A \text{ true}) \Longleftrightarrow Prob(\hat{x} \in \widehat{R}_A|x \in R_A)$$

For any regions $(R, \widehat{R}) \in \{(R_0, \widehat{R}_0), (R_A, \widehat{R}_A)\}$, it can be noticed that we have:

$$Prob(\hat{x} \in \widehat{R}|x \in R) = \sup_{x \in R} Prob(\hat{x} \in \widehat{R}|x)$$

In plain words, $Prob(\hat{x} \in \widehat{R}|x \in R)$ is the probability that the estimated value belongs to $\widehat{R}$ knowing that the actual value of the parameter is x. Controlling $\sup_{x \in R_0} Prob(X \in \widehat{R}_0|x)$ and $\sup_{x \in R_A} Prob(X \in \widehat{R}_A|x)$ is hence equivalent to control probabilities of making type I and type II errors in worst cases.

Let Q_0 and Q_A be these two probabilities such as:

$$Q_0(R_0, \widehat{R}_0) = \sup_{x \in R_0} Prob(\hat{x} \in \widehat{R}_0|x) \tag{1}$$

$$Q_A(R_A, \widehat{R}_A) = \sup_{x \in R_A} Prob(\hat{x} \in \widehat{R}_A|x) \tag{2}$$

The above definitions can be exploited in three different ways. First, given regions R_0 and R_A defined by a null hypothesis H_0 and an alternative hypothesis H_A, and given the estimator $\hat{x}$ defining rejection regions $\widehat{R}_0$ and $\widehat{R}_A$ such that $\hat{x} \in \widehat{R}_0 \cap \widehat{R}_A$, the probabilities of making type I and II errors can be calculated (more precisely, upper bounded) using Eqs. (1) and (2). Second, given the estimator $\hat{x}$ defining rejection regions $\widehat{R}_0$ and $\widehat{R}_A$ such that $\hat{x} \in \widehat{R}_0 \cap \widehat{R}_A$ and given a confidence level α, a confidence interval for x can be obtained by delimiting regions R_0 and R_A such that $Q_0(R_0, \widehat{R}_0) = Q_A(R_A, \widehat{R}_A) = \alpha$. Then, it will be stated with a confidence level α that $x \in (R_0 \cup R_A)^c$ (complement of $R_0 \cup R_A$). Third, given regions R_0 and R_A defined by a null hypothesis H_0 and an alternative hypothesis H_A, and given ε a maximal tolerance for probability of making type I and type II errors, rejection regions $\widehat{R}_0$ and $\widehat{R}_A$ can be delimited such that $Q_0(R_0, \widehat{R}_0) = Q_A(R_A, \widehat{R}_A) = \epsilon$. Then, H_0 will be rejected iff $\hat{x} \in \widehat{R}_0$, and H_A will be rejected iff $\hat{x} \in \widehat{R}_A$, with at most a probability ε of making an error.

Formulation of fold-change statements

Let X_1 be a random variable following a normal distribution $X_1 : \mathcal{N}(x_1, \sigma_1^2)$ and $X_2 : \mathcal{N}(x_2, \sigma_2^2)$ with $cov(X_1, X_2) = 0$. Let s_1 be a sample from X_1 of size n_1 and empirical mean $\widehat{x_1^{obs}}$, and s_2 be a sample from X_2 of size n_2 and empirical mean $\widehat{x_2^{obs}}$. Furthermore, assume that σ_1 is known, and σ_2 is known (we will discuss this aspect later in detail). Consider the samples s_1 and s_2 as two sets of expression measurements of a specific gene of interest in two different biological conditions. Formulating statistical statements about the fold-change between the means x_1 and x_2 using the above described statistical approach leads to adequately define regions R_0, R_A, $\widehat{R}_0$ and $\widehat{R}_A$. In order to formulate fold-change statements between the means x_1 and x_2 of the two normal distributions, regions R_0, R_A, $\widehat{R}_0$ and $\widehat{R}_A$ have to be defined using conic sections C_θ such as:

$$C_\theta = \left\{ (a, b) \in \mathbb{R}^2, \quad \tan\left(\frac{\pi}{4} - \theta\right) < \frac{a}{b} < \tan\left(\frac{\pi}{4} + \theta\right) \right\}$$

where $0 \leq \theta \leq \frac{\pi}{4}$ is an angle on each side of the first diagonal.

Moreover, means x_1 and x_2 must be controlled to avoid a negative contribution of the distributions to the fold-change. As only positive values of means have to be taken into account, regions R_0 and R_A must be curtailed from zero, and regions $\widehat{R}_0$ and $\widehat{R}_A$ must be curtailed from $\widehat{x}_1^{obs}$ and $\widehat{x}_2^{obs}$. We will henceforth alleviate the notations and use $\widehat{x}_1$(resp. $\hat{x}_2$) for the value $\hat{x}_1^{obs}$ (resp. $\hat{x}_2^{obs}$) of the estimator of the mean x_1(resp.x_2), as computed from the data sample.

Then, regions R_0, R_A, $\widehat{R}_0$, and $\widehat{R}_A$ are defined such as:

$$R_0(\theta_0) = \left\{(a,b) \in \mathbb{R}^2, (a,b) \in C_{\theta_0} \text{ and } 0 < a \text{ and } 0 < b\right\}$$

$$R_A(\theta_A) = \left\{(a,b) \in \mathbb{R}^2, (a,b) \notin C_{\theta_A} \text{ and } 0 < a \text{ and } 0 < b\right\}$$

$$\widehat{R}_0(\widehat{\theta}_0, \widehat{x}_1, \widehat{x}_2) = \left\{(a,b) \in \mathbb{R}^2, (a,b) \notin C_{\widehat{\theta}_0} \text{ and } \widehat{x}_1 < a \text{ and } \widehat{x}_2 < b\right\}$$

$$\widehat{R}_A(\widehat{\theta}_A, \widehat{x}_1, \widehat{x}_2) = \left\{(a,b) \in \mathbb{R}^2, (a,b) \in C_{\widehat{\theta}_A} \text{ and } \widehat{x}_1 < a \text{ and } \widehat{x}_2 < b\right\}$$

Figure 2 illustrates the definition of regions R_0 (Figure 2A), R_A (Figure 2B), $\widehat{R}_0$ (Figure 2C), and $\widehat{R}_A$ (Figure 2D) with arbitrary parameters.

Probabilities Q_0 and Q_A described in Eqs. (1) and (2) with the above defined regions are then defined by:

$$Q_0(R_0(\theta_0), \widehat{R}_0(\widehat{\theta}_0, \widehat{x}_1, \widehat{x}_2)) = \sup_{(x_1,x_2)\in R_0(\theta_0)} Prob((Y_1, Y_2) \in \widehat{R}_0(\widehat{\theta}_0, \widehat{x}_1, \widehat{x}_2 | x_1, x_2)) \quad (3)$$

$$Q_A(R_A(\theta_A), \widehat{R}_A(\widehat{\theta}_A, \widehat{x}_1, \widehat{x}_2)) = \sup_{(x_1,x_2)\in R_A(\theta_A)} Prob((Y_1, Y_2) \in \widehat{R}_A(\widehat{\theta}_A, \widehat{x}_1, \widehat{x}_2 | x_1, x_2)) \quad (4)$$

with $Y_1 : \mathcal{N}\left(x_1, \sigma_1^2 = \frac{\sigma_1^2}{n_1}\right)$, $Y_2 : \mathcal{N}\left(x_2, \sigma_2^2 = \frac{\sigma_2^2}{n_2}\right)$, and $cov(Y_1, Y_2) = 0$.

As explained above, the above definitions can be exploited in three different ways. First, given two angles θ_0 and θ_A that are relevant to assess the similarity and the distinctness between x_1 and x_2, and given $\widehat{x}_1$ and $\widehat{x}_2$ defining the rejection regions such as $\widehat{R}_0(\hat{\theta}, \widehat{x}_1, \widehat{x}_2)$ and $\widehat{R}_A(\hat{\theta}, \widehat{x}_1, \widehat{x}_2)$ with $\hat{\theta} = |\arctan(\widehat{\frac{x_2}{x_1}}) - \frac{\pi}{4}|$, the probabilities of making type I and II errors can be calculated using Eqs. (3) and (4). To formulate fold-change statements, angles θ_0 and θ_A are defined as $\theta_0 = \arctan(f_{C_0}) - \frac{\pi}{4}$ and

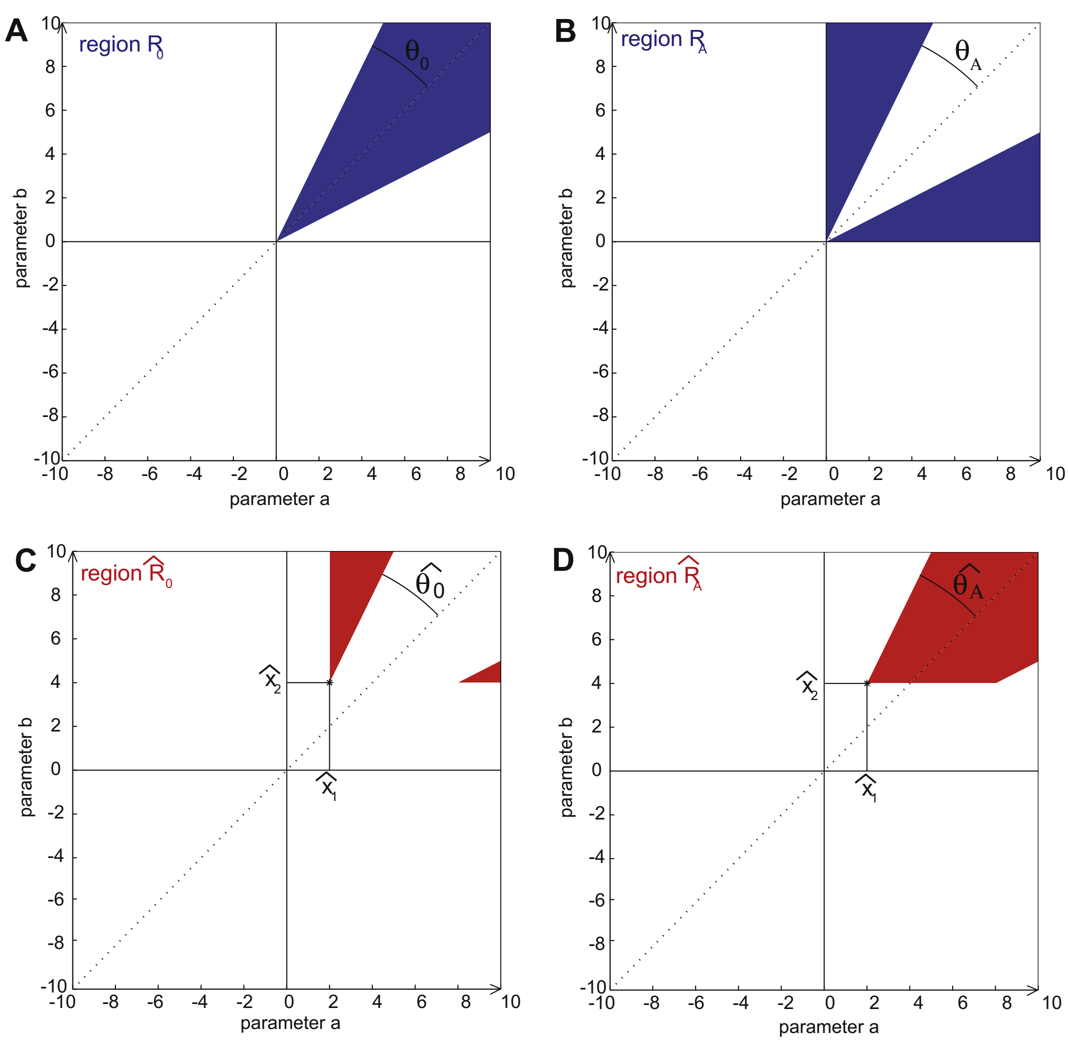

Figure 2 Representation of the different regions R_0, R_A, $\widehat{R}_0$ and $\widehat{R}_A$

A. Region R_0 shown in blue with $\theta_0 = \arctan(2) - \frac{\pi}{4}$. **B.** Region R_A shown in blue with $\theta_A = \arctan(2) - \frac{\pi}{4}$. **C.** Region $\widehat{R}_0$ shown in red with $\widehat{\theta}_0 = \arctan(2) - \frac{\pi}{4}$, $\widehat{x}_1 = 2$ and $\widehat{x}_2 = 4$. **D.** Region $\widehat{R}_A$ shown in red with $\widehat{\theta}_A = \arctan(2) - \frac{\pi}{4}$, $\widehat{x}_1 = 2$ and $\widehat{x}_2 = 4$.

$\theta_A = \arctan(f_{C_A}) - \frac{\pi}{4}$ where $f_{C_0} \geqslant 1$ and $f_{C_A} \geqslant 1$ are two fold-change values that are relevant to assess the similarity and the distinctness between x_1 and x_2. Q_0 (resp. Q_A) will then give the probability of making an error when stating that two genes are differentially (resp. similarly) expressed. Second, given $\widehat{x}_1$ and $\widehat{x}_2$ (computed from the data samples) defining the rejection regions such as $\widehat{R_0}(\hat{\theta}, \widehat{x_1}, \widehat{x_2})$ and $\widehat{R_A}(\hat{\theta}, \widehat{x_1}, \widehat{x_2})$ with $\hat{\theta} = \left|\arctan\left(\frac{\widehat{x_2}}{\widehat{x_1}}\right) - \frac{\pi}{4}\right|$, and given a confidence level α, a confidence interval for $\theta = \left|\arctan\left(\frac{x_2}{x_1}\right) - \frac{\pi}{4}\right|$ can be obtained by delimiting regions $R_0(\theta_0)$ and $R_A(\theta_A)$ such that $Q_0(R_0(\theta_0), \widehat{R_0}(\widehat{\theta_A}, \widehat{x_1}, \widehat{x_2})) = Q_A(R_A(\theta_A), \widehat{R_A}(\widehat{\theta_A}, \widehat{x_1}, \widehat{x_2})) = \alpha$. Then, it will be stated with a confidence level α that $\theta \in (R_0(\theta_0) \cup R_A(\theta_A))^c$ which corresponds to the state that $\theta_0 < \theta < \theta_A$. By denoting f_C the fold-change between x_1 and x_2, this is equivalent to state with a confidence level α that $f_{C_0} < f_C < f_{C_A}$ where $f_{C_0} = \tan(\theta_0) + \frac{\pi}{4}$, $f_C = \tan(\theta) + \frac{\pi}{4}$, and $f_{C_A} = \tan(\theta_A) + \frac{\pi}{4}$. Third, given two angles θ_0 and θ_A (*i.e.*, fold-changes, see above) that are relevant to assess the similarity and the distinctness between x_1 and x_2, and given ε a maximal tolerance for the probability of making type I and type II errors, rejection regions $\widehat{R_0}(\widehat{\theta_0}, \widehat{x_1}, \widehat{x_2})$ and $\widehat{R_A}(\widehat{\theta_A}, \widehat{x_1}, \widehat{x_2})$ can be delimited. However, as those regions are defined by three parameters, their delimitation is more complicated to define than for the regions R_0 and R_A. Also, as they are not essential for our question, we will not focus here on their delimitation.

Test behavior and biological application

Test behavior

Let us have three different situations as displayed in **Figure 3** represented as bar charts: a gene showing statistically significantly differential expressions (Figure 3A), another gene whose expression does not differ statistically significantly from one to another condition (Figure 3B), and finally a situation where a given gene cannot be said to be statistically significantly differentially nor similarly expressed due to the variability of its expression levels (Figure 3C).

As previously explained, Q_0 is the probability of making an error in stating that a certain gene is differentially expressed between the two biological conditions. Lower values (close to zero) of Q_0 indicate then dissimilarities in terms of gene expression as in Figure 3A ($Q_0 = 0.01$) as opposed to the cases presented in Figure 3B and C. Similarly, Q_A is the probability of making an error in stating that a certain gene is similarly expressed between the two biological conditions. The situation displayed in Figure 3B ($Q_A = 0.03$) can be considered as statistically significant as opposed to the cases presented in Figure 3A and C. Moreover, values such as the ones in the example of Figure 3C are associated neither with similarity nor distinctness from a statistical point of view.

In summary, our examples suggest three typical situations when comparing the expression levels of a certain gene between two different biological conditions that our statistical test can detect.

Biological application

In order to illustrate the behavior of our statistical test in a biological application, we use a dataset coming from transcriptome microarray studies of adrenal cancer [16–19]. This dataset is composed of 3 different biological conditions: (i) adrenal cortex carcinoma (ACC), 33 samples (ii) adrenal cortex adenoma (ACA), 22 samples and (iii) normal adrenal cortex (NAC) that serves as control, 10 samples [18]. The insulin-like growth factor (IGF) signaling system was identified as being one of the most dominantly altered in ACC in the form of greatly increased expression of IGF2 [17]. In a subsequent study [18], 10 genes associated with the cancer phenotype are identified. Steroid signaling is associated with ACA since the activation of this pathway is needed for different hormone production. We estimate the evolution of the values for Q_0 and Q_A as we vary the fold-change parameter f_{C_0} and f_{C_A}, respectively. For example, considering the differences between ACC and NAC (*i.e.*, the malignancy profile), we computed several subtraction profiles as displayed for Q_0 (Figure 3D) and Q_A (Figure 3E). As expected, Q_0 is more restrictive as f_{C_0} increases and conversely when we increase f_{C_A} the value of Q_A is more permissive. The results obtained here for differentially expressed genes is presented in **Figure 4**. A summary of the number of differentially expressed genes is displayed in a Venn Diagram (Figure 4A). The advantage of our method is that we can extend our scope by looking at cases other than simply differential expression amongst the different biological conditions. For instance, we can consider similar expression in one of the comparisons (Figure 4B) or even in two of them (Figure 4C). Each of these possibilities give us different insights. Among the 114 genes differentially detected, the collagen type I – alpha 1 gene (COL1A1) has been identified as present in the adrenal cancer malignancy (Figure 4D). In particular, we detected the gene encoding secreted phosphoprotein 1 (SPP1), which is present in carcinoma and control samples, with little variation while displaying a large variability in the adenoma conditions. This can be explained since this dataset has adenomas that produce different hormones all synthesized from cholesterol (steroids), which contribute to the variability of this gene (Figure 4E). We can predict that gene Interleukin-1 alpha (IL-1a) is similarly expressed among carcinoma samples but is not relevant as a malignancy marker since the expression is not statistically consistent with the other two biological conditions (Figure 4F).

Variance estimation

The CDS statistical test described here is based on the assumption of known variances of the signals. This

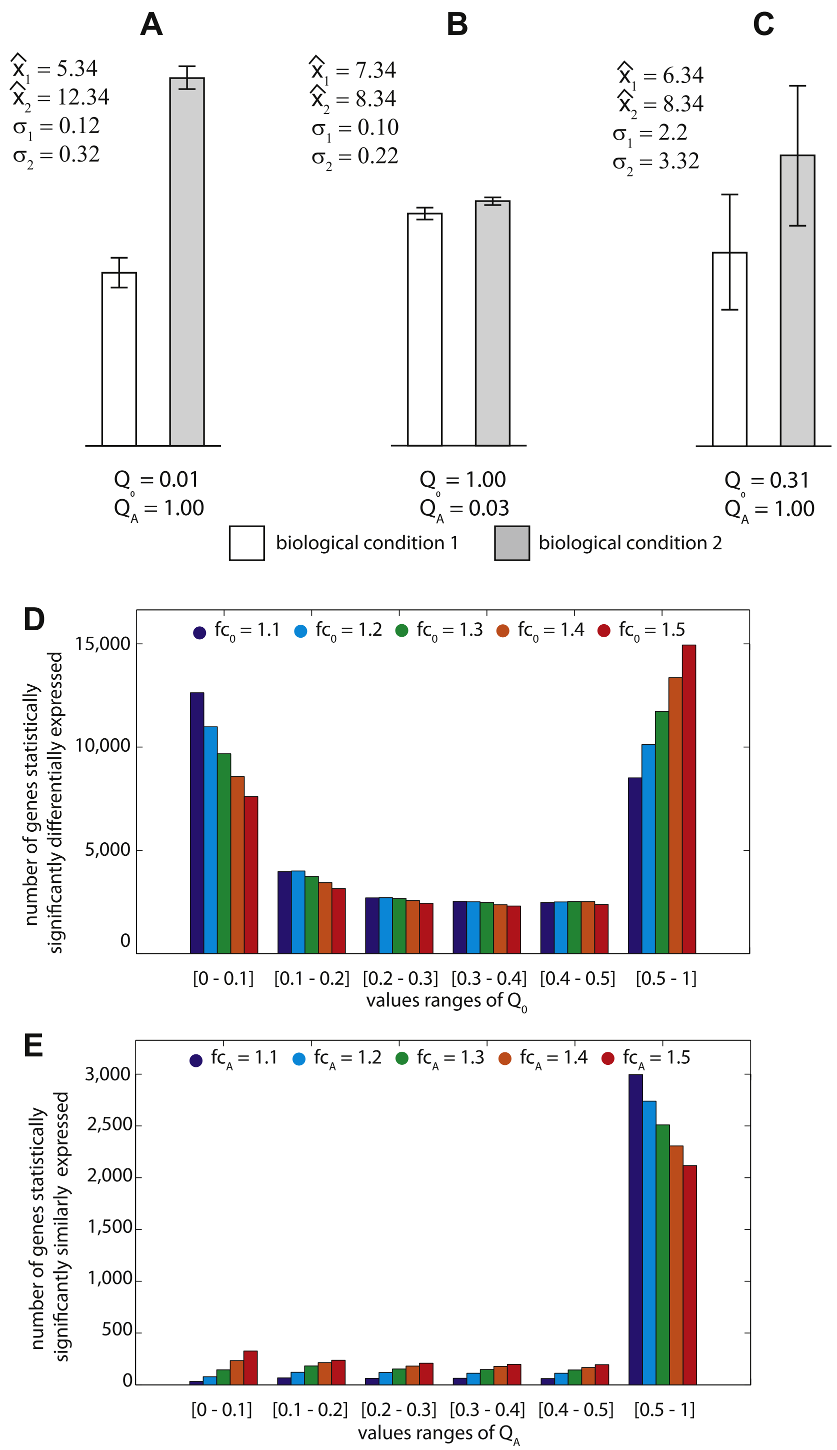

Figure 3 Test behavior validation

In silico simulations using standard normal distributed data with parameters x_1, x_2, σ_1, σ_2 capture 3 different situations shown in **A–C. A.** Case of differentially expressed gene having a significant Q_0 value ($Q_0 < 0.05$) but an high Q_A value. **B.** The opposite case being statistically similar ($Q_A < 0.05$). **C.** Case where neither Q_0 nor Q_A display statistical significance. Our method is tested on a real biological dataset (panels **D** and **E**) showing the correct behavior. **D.** Values of Q_0 are directly changing as a function of the f_{C_0} parameter. As we increase the f_{C_0} parameter, the values of Q_0 are higher. This means that the more we increase the f_{C_0} parameter the less Q_0 values we have for a given bin of the Q_0 histogram. **E.** Values of Q_A are inversely changing as a function of the f_{C_A} parameter. As we increase the f_{C_A} parameter the values of Q_A are lower. This means that the more we increase the f_{C_A} parameter the more Q_A values we have for a given bin of the Q_A histogram.

assumption is reasonable in cases where the technology itself provides direct estimates of the variance as is the case for dPCR and certain NGS applications using recall chemistry. Furthermore, modern microarray platforms provide coefficient of variation estimates which can be used as proxies for variance [20]. Another most important case is

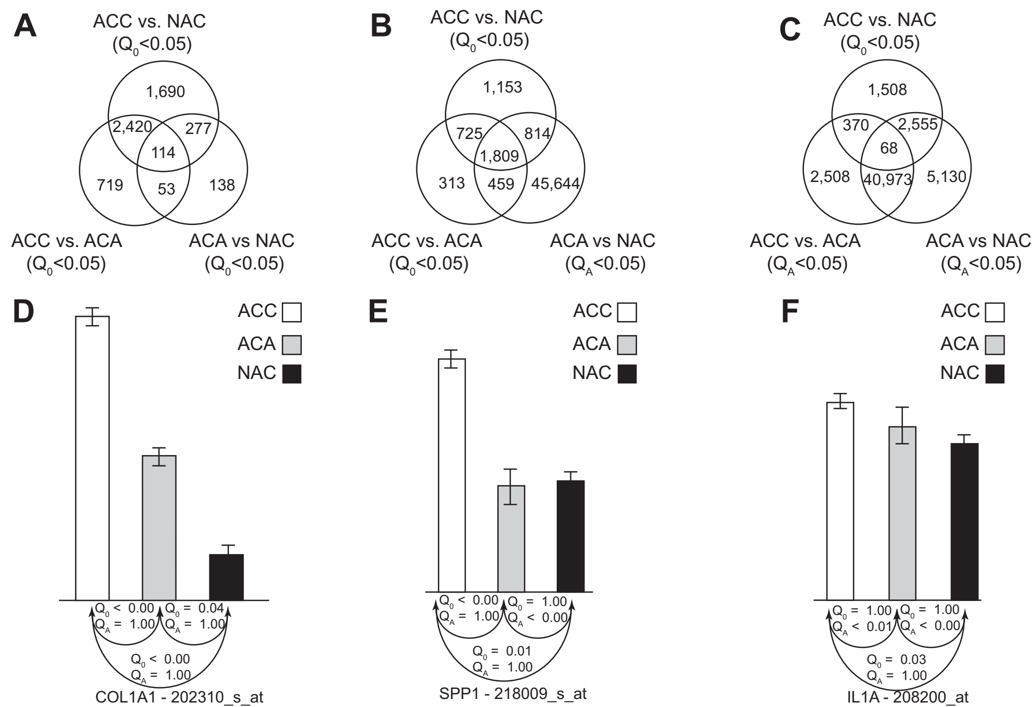

Figure 4 Experimental validation of our CDS statistical test

Venn Diagrams of the differentially expressed genes when comparing 3 different biological conditions are shown in panel **A–C. A.** Comparing differential expression across the three comparisons is the usual case. With the CDS method we can capture more cases, for instance shown in panel B and C. **B.** Comparing differential expression in two subtractions and similarity in one subtraction. **C.** Comparing one differential expression and two similarity expressions. Panels **D–F.** Examples of genes detected using both Q_0 and Q_A values issued from our method. **D.** Difference in the three biological conditions. **E.** Similarity between two biological conditions. **F.** Similarity in two comparisons and difference among the three biological conditions.

the often encountered scenario of biomedical investigations where a large number of individual measurements are available (*e.g.*, a single recording per patient or subject). Computing the biological variations from the entire cohort of samples can then allow us to compare individual measurements amongst each other with the CDS statistical test.

Multiple testing

The CDS statistical test can and should be combined with false-positive discovery rates or similar corrections when used in a serial manner. We have used successfully both FDR and pFDR methods [21,22]. Note, that the data presented here were not subjected to multiple testing correction as they only serve to demonstrate the applicability of the CDS method.

Conclusion

The CDS statistical test is suitable for quantitatively checking statements, typically to determine confidence intervals, about the fold-change between the means of two normally distributed variables, under assumptions that the variances are known. Applied to the identification of differentially and similarly expressed genes in the context of microarray measurements, this statistical test correctly identified genes of interest in benchmark situations and also gave confidence intervals of the fold-change. Moreover, this statistical test can be used for any -omics data as long as the similarity or distinctness between two signals is measured by the fold-change and the required assumptions are fulfilled. Even if in the present case, assumptions have been made that gene expression measurements are distributed according to normal distributions with known variances, the principle of the test remains valid for other distributions and it can be numerically implemented. Indeed, Monte Carlo simulations can be performed to estimated probabilities $Prob\left((Y_1, Y_2) \in \widehat{R_0}\left(\widehat{\theta_0}, \widehat{x_1}, \widehat{x_2}\right)\right)$ and $Prob\left((Y_1, Y_2) \in \widehat{R_A}\left(\widehat{\theta_A}, \widehat{x_1}, \widehat{x_2}\right)\right)$ when explicit forms cannot be obtained easily. Finally, when variances of the normal distributions are not supposed to be known but have to be estimated from the samples, Student t-distributions can be used instead of the normal distributions.

Methods

Explicit forms of probabilities

The explicit forms $Prob\left((Y_1, Y_2) \in \widehat{R_0}\left(\widehat{\theta_0}, \widehat{x_1}, \widehat{x_2}\right) | x_1, x_2\right)$ and $Prob\left((Y_1, Y_2) \in \widehat{R_A}\left(\widehat{\theta_A}, \widehat{x_1}, \widehat{x_2}\right) | x_1, x_2\right)$ have been obtained by applying affine transformations to the bivariate normal distribution (Y_1, Y_2) in order to make the integration region rectangular and then easily computable using the standard bivariate normal complementary cumulative distribution function. These explicit forms are given in Supplementary materials (http://cds.ihes.fr) in Eqs. (5) and (6).

Type I and type II risks upper bounds computation

It is notable that supremums of Eqs. (3) and (4) are reached on boundaries of regions $R_0(\theta_0)$ and $R_A(\theta_A)$, meaning on lines $y = \arctan\left(\frac{\pi}{4} + \theta_0\right)$ and $y = \arctan\left(\frac{\pi}{4} - \theta_0\right)$ for Q_0, and on lines $y = \arctan\left(\frac{\pi}{4} + \theta_A\right)$ and $y = \arctan\left(\frac{\pi}{4} - \theta_A\right)$ for Q_A. Albeit mathematically defined, as in (3) and (4), the computation of these probabilities begged for a numerical estimation given the complexity of the explicit form of their first and second derivatives. In this line of thought, we use numerical methods to obtain the maximum values of the probabilities considering a finite number of instances of the probability distribution functions (as opposed to the exact functions from the mathematical definition) and we evaluated them over a finite interval in the parameter space (as opposed to the infinite interval assumed in the mathematical definition).

Confidence interval computation

As Q_0 increases (respectively Q_A increase) with θ_0 (resp. θ_A), this delineation can done by performing a binary search of the angle θ_0 (resp. θ_A) from $\hat{\theta} = \left|\arctan\left(\frac{\hat{x_2}}{\hat{x_1}}\right) - \frac{\pi}{4}\right|$ to 0 (resp. to $\frac{\pi}{4}$) until $Q_0\left(R_0(\theta_0), \widehat{R_0}\left(\widehat{\theta_0}, \hat{x_1}, \hat{x_2}\right)\right) = \alpha$ (resp. $Q_A\left(R_A(\theta_A), \widehat{R_A}\left(\widehat{\theta_A}, \hat{x_1}, \hat{x_2}\right)\right) = \alpha$) is reached.

Implementation

This statistical test has been implemented in Java and it is possible to compute $Q_0\left(R_0(\theta_0), \widehat{R_0}\left(\widehat{\theta_0}, \hat{x_1}, \hat{x_2}\right)\right)$ and $Q_A\left(R_A(\theta_A), \widehat{R_A}\left(\widehat{\theta_A}, \hat{x_1}, \hat{x_2}\right)\right)$ as well as the confidence intervals for a set of 30,000 values in a few minutes. The computational speed allows us to imagine using this test for the analysis of NGS data. It may be interesting to notice here that some thought can be stated regarding the nature of the distributions to be used for the analysis of NGS. Indeed, in contrast to data from microarrays where the values are continuous signals, the measured values are discrete, and thus the use of discrete distributions like the negative binomial distribution can be interesting for better modeling of assumptions. An R implementation of the CDS statistical test is available at http://cds.ihes.fr.

Data processing

The transcriptome data discussed have first been published in [17], and are available from GEO [23] under Accession No. GSE10927, and mace (http://www.mace.ihes.fr) under Accession No. 2651913582. Data were log-transformed, subjected to an additional round of quality control [24,25], and normalized using NeONORM [26] for subtraction profiling. No multiple testing correction was performed so as to retain the original P values.

Authors' contributions

NT, JFGD, SN, AB and AL performed test formulation. NT, JFGD, BT and SN carried out implementation, and NT, BT and AB performed testing. NT, JFGD, AB and AL wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgments

This work was made possible through funds from the Centre National de la Recherche Scientifique, the Agence Nationale pour la Recherche (Grant No. ANR-07-PHYSIO-013-01), the Fondation pour la Recherche sur l'Hypertension Artérielle (Grant No. AO 2007), the Agence Nationale de Recherches sur le SIDA et les hépatites virales (ANRS) and the Genopole Evry (all awarded to AB). JFBG was recipient of a CONACYTMexico PhD Fellowship (Grant No. 207676/302245).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.06.002.

References

[1] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002;18:546–54.
[2] Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Opt 1997;2:364–74.
[3] DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–60.
[4] Barry WT, Kernagis DN, Dressman HK, Griffis RJ, Hunter JD, Olson JA, et al. Intratumor heterogeneity and precision of microarray-based predictors of breast cancer biology and clinical outcome. J Clin Oncol 2010;28:2198–206.
[5] Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 2003;4:210.
[6] McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 2009;25:765–71.
[7] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21.
[8] Lunnstedt I, Speed T. Replicated microarray data. Stat Sin 2002;12:31–46.

[9] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. Bioinformatics 2001;17:509–19.

[10] Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. Ann Math Stat 1945;16:243–58.

[11] Chapman DG. Some two sample tests. Ann Math Stat 1950;21:601–6.

[12] Lee JC, Lin SH. Generalized confidence intervals for the ratio of means of two normal populations. J Stat Plan Inference 2004; 123:49–60.

[13] Armstrong P, Garrido R, de Dios Ortúzar J. Confidence intervals to bound the value of time. Transp Res Part E: Logist Transp Rev 2001; 37:143–61.

[14] Deng X, Xu J, Hui J, Wang C. Probability fold change: a robust computational approach for identifying differentially expressed gene lists. Comput Methods Programs Biomed 2009;93:124–39.

[15] Broberg P. Statistical methods for ranking differentially expressed genes. Genome Biol 2003;4:R41.

[16] Zennaro MC, Jeunemaitre X. Mutations in KCNJ5 gene cause hyperaldosteronism. Circ Res 2011;108:1417–8.

[17] Giordano TJ, Kuick R, Else T, Gauger PG, Vinco M, Bauersfeld J, et al. Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. Clin Cancer Res 2009;15:668–76.

[18] Giordano TJ, Thomas DG, Kuick R, Lizyness M, Misek DE, Smith AL, et al. Distinct transcriptional profiles of adrenocortical tumors uncovered by DNA microarray analysis. Am J Pathol 2003; 162:521–31.

[19] Boulkroun S, Samson-Couterie B, Dzib JF, Lefebvre H, Louiset E, Amar L, et al. Adrenal cortex remodeling and functional zona glomerulosa hyperplasia in primary aldosteronism. Hypertension 2010;56:885–92.

[20] Noth S, Brysbaert G, Pellay FX, Benecke A. High-sensitivity transcriptome data structure and implications for analysis and biologic interpretation comparison of AB1700 and Affymetrix. Genomics Proteomics Bioinformatics 2006;4:212–29.

[21] Benjamini Y, Hochberg Y. Controlling false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300.

[22] Storey JD. The positive false discovery rate: a Bayesian interpretation and the *q*-value. Ann Stat 2003;31:2013–35.

[23] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res 2011;39(Database issue):D1005–10.

[24] Brysbaert G, Pellay FX, Noth S, Benecke A. Quality assessment of transcriptome data using intrinsic statistical properties. Genomics Proteomics Bioinformatics 2010;8:57–71.

[25] Bécavin C, Tchitchek N, Mintsa-Eya C, Lesne A, Benecke A. Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. Bioinformatics 2011;27:1413–21.

[26] Noth S, Brysbaert G, Benecke A. Normalization using weighted negative second order exponential error functions (NeONORM) provides robustness against asymmetries in comparative transcriptome profiles and avoids false calls. Genomics Proteomics Bioinformatics 2006;4:90–109.