Bioinformatics, 36(21), 2020, 5194–5204
doi: 10.1093/bioinformatics/btaa654
Advance Access Publication Date: 19 July 2020
Original Paper

Gene expression

An information-theoretic approach for measuring the distance of organ tissue samples using their transcriptomic signatures

Dimitris V. Manatakis [1,*], Aaron VanDevender[2] and Elias S. Manolakos[3,4]

[1]Emulate Inc., Boston, MA 02210, USA, [2]Founders Fund, San Francisco, CA 94129, USA, [3]Department of Informatics and Telecommunications, University of Athens, Athens 15784, Greece and [4]Bouve College of Health Sciences, Northeastern University, Boston, MA 02115, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on January 30, 2020; revised on May 26, 2020; editorial decision on July 10, 2020; accepted on July 14, 2020

Abstract

Motivation: Recapitulating aspects of human organ functions using *in vitro* (e.g. plates, transwells, etc.), *in vivo* (e.g. mouse, rat, etc.), or *ex vivo* (e.g. organ chips, 3D systems, etc.) organ models is of paramount importance for drug discovery and precision medicine. It will allow us to identify potential side effects and test the effectiveness of new therapeutic approaches early in their design phase, and will inform the development of better disease models. Developing mathematical methods to reliably compare the 'distance/similarity' of organ models from/to the real human organ they represent is an understudied problem with important applications in biomedicine and tissue engineering.

Results: We introduce the Transcriptomic Signature Distance (*TSD*), an information-theoretic distance for assessing the transcriptomic similarity of two tissue samples, or two groups of tissue samples. In developing *TSD*, we are leveraging next-generation sequencing data as well as information retrieved from well-curated databases providing signature gene sets characteristic for human organs. We present the justification and mathematical development of the new distance and demonstrate its effectiveness and advantages in different scenarios of practical importance using several publicly available RNA-seq datasets.

Availability and Implementation: The computation of both *TSD* versions (simple and weighted) has been implemented in R and can be downloaded from https://github.com/Cod3B3nd3R/Transcriptomic-Signature-Distance.

Contact: dimitris.manatakis@emulatebio.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Assessing the transcriptomic distance of biological samples (e.g. organ tissues, cells of different types, etc.) is essential for understanding their functional differences and recognizing different disease states (Aibar *et al.*, 2016; Crow *et al.*, 2019; McDonough *et al.*, 2019; Mohammed *et al.*, 2019). Recently, significant efforts have been invested towards characterizing organ tissues (e.g. liver, kidney, intestine, etc.) of different species (e.g. human, mouse, rat, etc.) at various states (e.g. healthy, diseased, etc.) (Keen and Moore, 2015; Mele *et al.*, 2015; Sollner *et al.*, 2017; Suntsova *et al.*, 2019; Uhlen *et al.*, 2015, 2017; Yu *et al.*, 2015) using RNA-sequencing, a mature technology for quantifying gene transcripts in biological samples. A notable effort is the Human Protein Atlas (HPA) project (Uhlen *et al.*, 2015), a Swedish-based program providing, among others, gene expression signatures for 37 healthy human organ tissues. Importantly, for each tissue type, the HPA provides gene sets exhibiting significantly elevated expressions compared to the other organ tissue types. It is widely accepted that these gene sets can be used to form a 'transcriptomic signature' of the specific organ, and their expression patterns characterize the tissue's underlying biological processes (Uhlen *et al.*, 2015).

Recent advancements in bioengineering and biotechnology have enabled the development of cell-culture-based organ models recapitulating critical functions of human organs (e.g. liver, intestine, brain, etc.) (Jang *et al.*, 2019; Kasendra *et al.*, 2020). The emergence of such *ex vivo* organ models has generated, in turn, the need for new mathematical tools for assessing their 'similarity' to the actual human organ they represent. Such tools will not only help us understand the model strengths and limitations but also reveal aspects we can improve in their design to optimize their physiological relevance and increase their value for precision medicine. Transcriptomic data (e.g. RNA-seq) is extensively utilized to determine the distance/similarity between biological samples (Chen *et al.*, 2015; Gentleman *et al.*, 2005; Jaskowiak *et al.*, 2013; Mele *et al.*, 2015; Skinnider *et al.*, 2019; Sollner *et al.*, 2017; Souto *et al.*, 2008; Sudmant *et al.*, 2015; Suntsova *et al.*, 2019) in conjunction with classical mathematical tools such as the Euclidean distance, Pearson's correlation, dimensionality reduction techniques (e.g. Principal Component Analysis, Linear Discriminant Analysis, Uniform Manifold Approximation and Projection) and so on. However, these methods exhibit significant limitations due to their sensitivity to noisy measurements, outliers, inability to capture non-linear relations, etc. (Li *et al.*, 2016; Pereira *et al.*, 2009). In this work, we introduce a new distance, called *Transcriptomic Signature Distance (TSD)*, that was inspired from the field of information retrieval, which addresses the problem of tissue sample comparisons in the framework of information theory, and circumvents the above-mentioned weaknesses of classical approaches.

Text similarity is a well-studied problem in information retrieval (Nagwani, 2015; Pradhan *et al.*, 2015). Over the years, many techniques have been proposed to measure the distance/similarity of documents based on features such as word frequencies, word patterns in sentences, etc. They process vector representations of documents and assume that documents with similar content exhibit similar feature patterns. RNA sequencing, on the other hand, allows us to read the transcriptome (i.e. read the stories) of tissues. These transcriptome 'stories' are written using a four nucleotide bases alphabet, which are assembled to construct words (i.e. the genes). Based on this analogy, gene expression patterns of homologous tissues should 'tell' similar stories, and therefore the set of words (i.e. genes), their relative frequencies of appearance and rankings in the stories are expected to be similar as well.

Our method exploits well-curated databases (e.g. HPA, GTExPortal, etc.) to retrieve gene sets that are considered as transcriptomic signatures of the different organ tissues (Lonsdale *et al.*, 2013; Yu *et al.*, 2015) and can adequately characterize them. Using such sets as a basis to assess the distance of tissue samples from a reference tissue allows us to significantly reduce the effects of sequencing 'background noise' and donor-to-donor variability in the data analysis. Moreover, using information theory and advanced statistical methods, *TSD* can capture the distance of a tissue sample from a reference organ tissue sample, while also exploiting any available knowledge on the different groups (e.g. different organs, organ models etc.) the two samples may belong to. If sample group (class) information is available, *TSD* exploits, in a principled manner, the intra-class variabilities and incorporates them into the sample distance calculation.

The proposed distance space is determined using the probability distribution of the expression of the signature genes as well as their rankings profile in the corresponding transcriptomes of the two tissues. We explain the advantages of the proposed metric and experimentally validate its ability to resolve distances between organ tissues in many practical situations of interest where more classical methods may fail.

The rest of the article is organized as follows: In Section 2, we present the development of *TSD* and justify why it can better capture the actual transcriptomic distance between two tissues. In Section 3, we present and discuss extensive experimental validation and comparison results to other methods in different scenarios of practical importance. Finally, we summarize our findings and point to future work in Section 4.

2 Materials and methods

In this section, we present *TSD*, a new method to measure the transcriptomic distance between organ tissue samples. *TSD* requires as input the gene expression levels (e.g. hit-counts, CPM, TPM, FPKM, etc.) of two tissue samples (or two sets of tissue samples) where one tissue sample is assumed to be the 'reference' (i.e. gold standard) from which we want to measure the distance of the other tissue sample.

2.1 Preliminaries

The HPA (Uhlen *et al.*, 2015) provides for each human organ three characteristic gene sets (called 'tissue enriched', 'group enriched' and 'tissue enhanced', in order of tissue specificity) that specify genes exhibiting significantly higher expression levels relatively to other organ tissues in the Atlas (see Supplementary Section S1 in Supplementary Material for the exact classification definitions). These gene sets, provide important information for understanding human organs' biology and functions (Uhlen *et al.*, 2015). In our approach, we consider the signature genes of a tissue to be the union of these three HPA gene sets. Utilizing an HPA-based gene signature, allows us to base distance calculations on the most informative genes for each specific organ tissue while removing from the analysis in an unbiased manner many 'noisy' lowly expressed genes that may severely mask the gene expression signal.

In the article, we will use lowercase letters to denote scalars, bold lowercase (uppercase) letters for vectors (matrices) and bold uppercase calligraphic letters for sets. Let $\dot{\boldsymbol{s}} = [\dot{g}_1, \dot{g}_2, \ldots, \dot{g}_N]$ and $\ddot{\boldsymbol{s}} = [\ddot{g}_1, \ddot{g}_2, \ldots, \ddot{g}_N]$ be two vectors storing the expression levels of N genes after applying RNA-sequencing on tissue samples $\dot{t}$ and $\ddot{t}$, respectively. For presentation purposes, we assume that $\dot{t}$ is our *'reference'* tissue sample and $\ddot{t}$ is the sample of tissue that we want to measure its distance from $\dot{t}$. From the HPA database, we retrieve the $M \leq N$ genes that characterize the reference organ as explained above, where tissue $\dot{t}$ was sampled from and are a subset of genes $\{g_1, g_2, \ldots, g_N\}$. Then, using $\dot{\boldsymbol{s}}$ and $\ddot{\boldsymbol{s}}$, we form the corresponding Atlas signature vectors $\dot{\boldsymbol{s}}^A = [\dot{g}_1^A, \dot{g}_2^A, \ldots, \dot{g}_M^A]$ and $\ddot{\boldsymbol{s}}^A = [\ddot{g}_1^A, \ddot{g}_2^A, \ldots, \ddot{g}_M^A]$.

For each Atlas signature vector, we estimate the corresponding discrete probability distribution $\dot{\boldsymbol{p}}^A = [\dot{P}_1^A, \dot{P}_2^A, \ldots, \dot{P}_M^A]$ and $\ddot{\boldsymbol{p}}^A = [\ddot{P}_1^A, \ddot{P}_2^A, \ldots, \ddot{P}_M^A]$. The probabilities of each Atlas gene are calculated using:

$$\dot{P}_k^A = \frac{\dot{g}_k^A}{\sum_{q=1}^{M} \dot{g}_q^A}, \quad \ddot{P}_k^A = \frac{\ddot{g}_k^A}{\sum_{q=1}^{M} \ddot{g}_q^A}, \quad \text{where} \quad k = \{1, 2, \ldots, M\}. \tag{1}$$

In addition, we form the vectors $\dot{\boldsymbol{\rho}}^A = [\dot{r}_1^A, \dot{r}_2^A, \ldots, \dot{r}_M^A]$ and $\ddot{\boldsymbol{\rho}}^A = [\ddot{r}_1^A, \ddot{r}_2^A, \ldots, \ddot{r}_M^A]$ containing the expression level rankings of the Atlas genes of the reference tissue ($\dot{t}$) in the full transcriptome gene expression vectors $\dot{\boldsymbol{s}}$ and $\ddot{\boldsymbol{s}}$. Note that $1 \leq r_i^A \leq N$.

In Section 2.2, we present the 'simple' version of the *TSD* that measures the transcriptomic distance between any two organ tissue samples. In Section 2.3, we present the development of the more general version, the so called *weighted-TSD (wTSD)*, which is used to estimate the distance between two samples knowing that they belong to two different classes (tissue sets) and considering the intra-class variabilities.

2.2 The Transcriptomic Signature Distance

TSD measures the transcriptomic distance of a tissue sample $\ddot{t}$ from a *'reference'* tissue sample $\dot{t}$ as the average of the *Jensen–Shannon Divergence (JSD)* (Section 2.2.1) and the *Rankings Correlation Distance (RCD)* (Section 2.2.2). We present below these two distances and their limitations when used in isolation to justify their combined use in *TSD*.

2.2.1 The Jensen–Shannon Divergence

JSD is popular for measuring the similarity between two probability distributions and is related to Shannon's entropy, *Kullback–Leibler Divergence (KLD)* and mutual information (Fuglede and Topsoe, 2004). *JSD* calculates the divergence (distance) between the probability distributions $\dot{\boldsymbol{p}}^A$ and $\ddot{\boldsymbol{p}}^A$ using the following equation:

$$JSD(\dot{\boldsymbol{p}}^A, \ddot{\boldsymbol{p}}^A) = H(\frac{1}{2} \cdot \dot{\boldsymbol{p}}^A + \frac{1}{2} \cdot \ddot{\boldsymbol{p}}^A) - \frac{1}{2} \cdot H(\dot{\boldsymbol{p}}^A) - \frac{1}{2} \cdot H(\ddot{\boldsymbol{p}}^A) \text{ where}$$
$$H(\boldsymbol{p}^A) = -\sum_{q=1}^{M} P_q^A \cdot \log_2 P_q^A \tag{2}$$

is the Shannon's entropy function (Jianhua, 1991). Unlike *KLD*, the square root of the *JSD* (*SR-JSD*) satisfies all the basic properties of a 'true' metric, such as symmetry, non-negativity, triangle inequality and identity of indiscernibles. Given that, we use the base-2

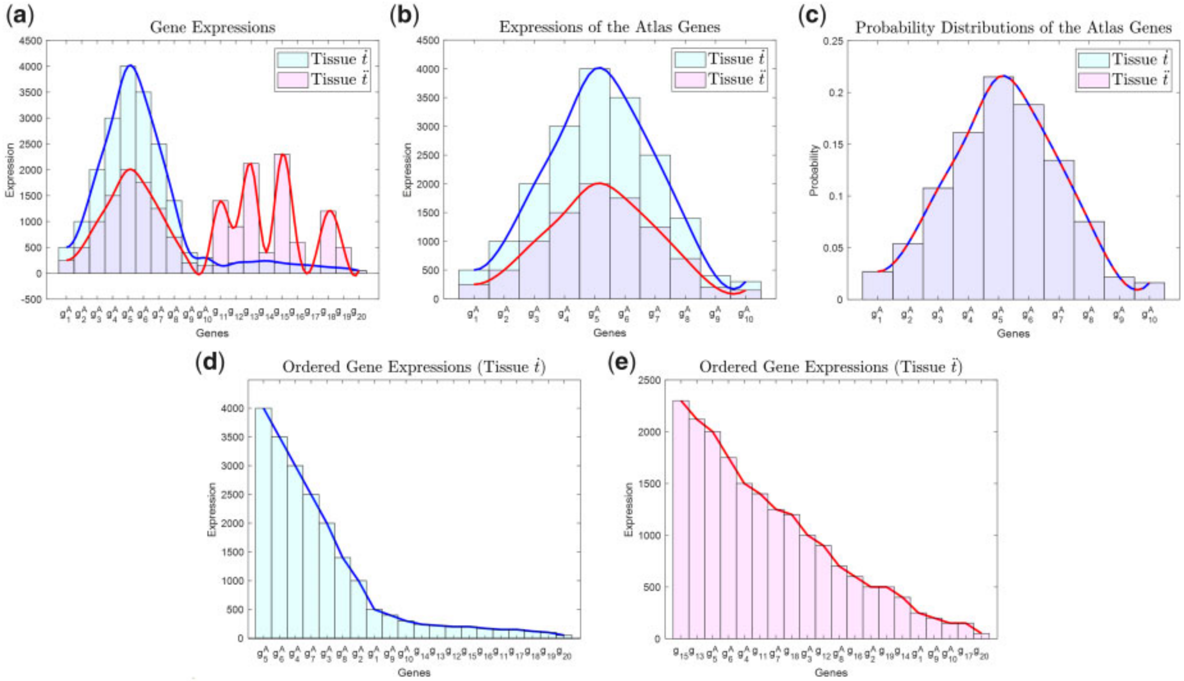

Fig. 1. (**a**) Gene expression vectors of the two tissues $\dot{t}$ (reference) and $\ddot{t}$. (**b**) Expression histograms of the Atlas genes $\{g_1^A, g_2^A, \ldots, g_{10}^A\}$ of the tissues $\dot{t}$ (reference) and $\ddot{t}$. (**c**) The discrete probability distributions $\{\dot{\boldsymbol{p}}^A, \ddot{\boldsymbol{p}}^A\}$ of the Atlas genes are identical due to the proportional gene expression levels (see b) and therefore SR-JSD$(\dot{\boldsymbol{p}}^A, \ddot{\boldsymbol{p}}^A) = 0$. (**d**, **e**) The gene expression rankings of the tissues $\{\dot{t}, \ddot{t}\}$. Note that the rankings of the Atlas genes in the whole transcriptome $\dot{\boldsymbol{\rho}}^A = [8, 7, 5, 3, 1, 2, 4, 6, 9, 10]$ and $\ddot{\boldsymbol{\rho}}^A = [16, 13, 9, 5, 3, 4, 7, 11, 17, 18]$, respectively, differ significantly which captures the tissue differences in this case

logarithm for calculating the Shannon's entropy (see Equation 2), the *SR-JSD* is bounded in the interval $[0, 1]$ where $0(1)$ corresponds to minimum (maximum) distance.

SR-JSD uses the probability distributions of the reference Atlas genes to measure the distance between the two samples and therefore it does not considers the expression levels of these genes in the whole transcriptome which is very important for the development of an accurate tissue distance metric. The following example demonstrates this limitation.

Figure 1a shows the whole gene expression profiles ($N = 20$ w.l.o.g) of two organ tissues $\dot{t}$ and $\ddot{t}$. Let us assume w.l.o.g. that $\{g_1^A, \ldots, g_{10}^A\}$ are the reference Atlas genes ($M = 10$) that characterize the tissue $\dot{t}$. Using their expressions in both tissues (see Fig. 1b), we form vectors $\dot{\boldsymbol{s}}^A$ and $\ddot{\boldsymbol{s}}^A$ and calculate the corresponding probability distributions $\dot{\boldsymbol{p}}^A$ and $\ddot{\boldsymbol{p}}^A$ (Fig. 1c). Since both tissues in this example $\{\dot{t}, \ddot{t}\}$ exhibit proportional gene expression for the Atlas genes, this results to identical probability distributions (Fig. 1c), and therefore, $SR\text{-}JSD(\dot{\boldsymbol{p}}^A, \ddot{\boldsymbol{p}}^A) = 0$, suggesting that $\dot{t}$ and $\ddot{t}$ are transcriptomically very close. However, as shown in Figure 1a, this is apparently not the case, which shows that using *SR-JSD* alone may fail to capture the tissues transcriptomic distance. To deal with this limitation, we also use in TSD, the correlation coefficient of the rankings of the Atlas genes in the whole transcriptome that can correct this situation (see Fig. 1d and e).

2.2.2 The Rankings Correlation Coefficient

The Rankings Correlation Coefficient (RCC) is defined as the Pearson's correlation of the Atlas gene ranking vectors, namely $\dot{\boldsymbol{\rho}}^A$ and $\ddot{\boldsymbol{\rho}}^A$, in the whole transcriptome. It is calculated based on the formula:

$$RCC(\dot{\boldsymbol{\rho}}^A, \ddot{\boldsymbol{\rho}}^A) = \frac{\sum_{q=1}^{M} (\dot{r}_q^A - \bar{\dot{r}}^A)(\ddot{r}_q^A - \bar{\ddot{r}}^A)}{\sqrt{\sum_{q=1}^{M} (\dot{r}_q^A - \bar{\dot{r}}^A)^2 \sum_{q=1}^{M} (\ddot{r}_q^A - \bar{\ddot{r}}^A)^2}}, \tag{3}$$

where $\bar{\dot{r}}^A$ and $\bar{\ddot{r}}^A$ are the corresponding mean rankings. The range of RCC is [-1,1] where 1(-1) implies perfect linear relation between the compared ranking vectors.

Applying Equation (3) to the Atlas ranking vectors provides us information about the linear relation of the reference Atlas genes in the whole transcriptome of the two tissues based on their expressions. Figure 1d and e demonstrates the ranking profile difference of the Atlas genes in the transcriptomes of $\dot{t}$ and $\ddot{t}$, respectively. This difference can successfully capture the tissues dissimilarity when *SR-JSD* may fail to do so as in the example of Figure 1.

2.2.3 Transcriptomic Signature Distance

The example presented in Figure 1 demonstrates the limitation of the *SR-JSD* to represent with accuracy the transcriptomic distance between two tissues. Figure 2 provides a similar example that demonstrates the same limitation when *RCC* is used alone. In this example, tissues $\dot{t}$ and $\ddot{t}$ have identical reference Atlas gene rankings in the corresponding transcriptomes (Fig. 2d and e) but different probability distributions (Fig. 2c). In this case, using *RCC* alone would suggest that the transcriptomic signatures of the tissues are identical which evidently is not the case. To address the limitations introduced when using either *SR-JSD* or *RCC* independently, we introduce the *TSD* which combines them:

$$TSD(\dot{t}, \ddot{t}) = \frac{1}{2} \cdot SR\text{-}JSD(\dot{\boldsymbol{p}}^A, \ddot{\boldsymbol{p}}^A) + \frac{1}{2} \cdot RCD(\dot{\boldsymbol{\rho}}^A, \ddot{\boldsymbol{\rho}}^A), \tag{4}$$

where $RCD = \sqrt{1 - ReLU(RCC)}$ is the *RCD*. *ReLU* is the *Rectified Linear Unit* activation function defined as: $ReLU(x) = x$, when $x > 0$ and zero otherwise. It can be shown that *RCD* is a 'true' metric (Jiaxing *et al.*, 2019). For the calculation of the *RCD*, we assume that if two ranking vectors have $RCC < 0$ (i.e. are anti-correlated) then their *RCD* is maximal (i.e. 1). Note that, *TSD* is bounded in the interval $[0, 1]$ where $0(1)$ corresponds to minimum (maximum) distance.

2.3 *TSD* of samples belonging to two different tissue sets

In Section 2.2, we presented *TSD* that can measure the distance between any two tissue samples $\dot{t}$ and $\ddot{t}$ without considering their classification. Here, we study the case where we want to measure the distance between two samples knowing that they belong to two

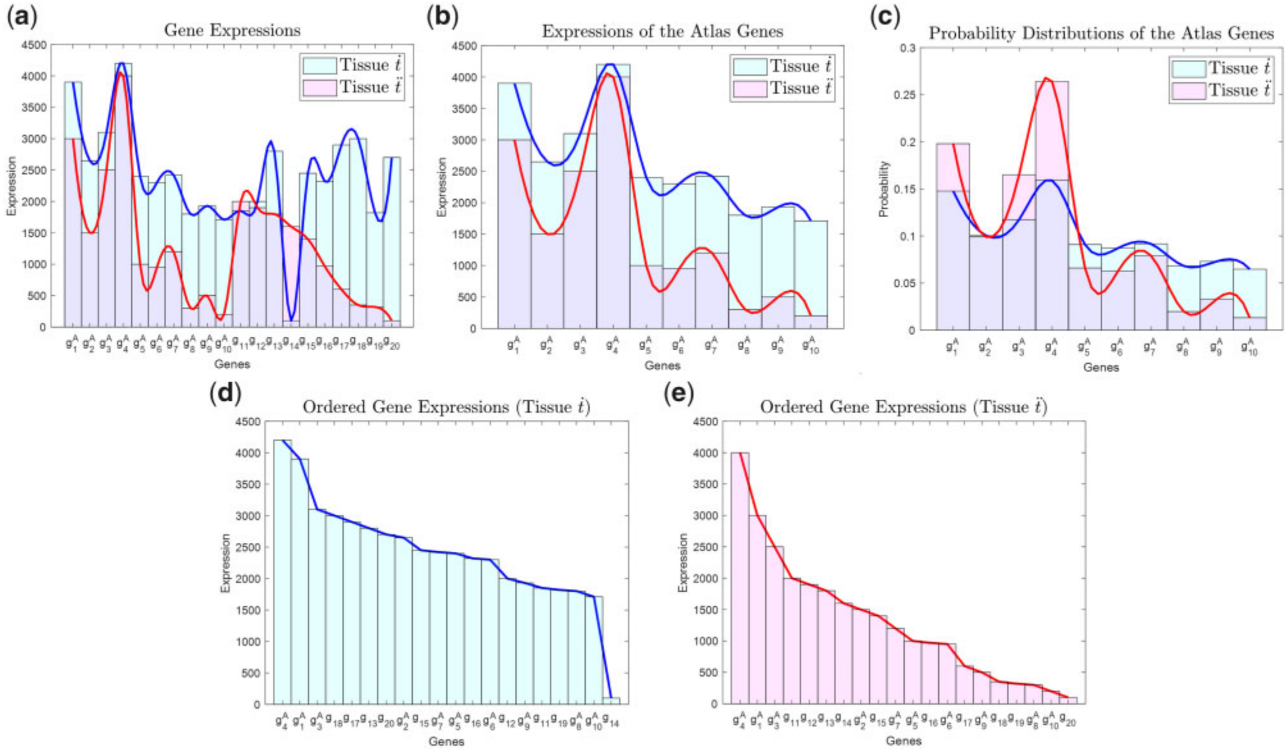

Fig. 2. (**a**) Gene expression vectors of two tissues $\dot{t}$ (reference) and $\ddot{t}$. (**b**) Expression histograms of the Atlas genes $\{g_1^A, g_2^A, \ldots, g_{10}^A\}$ of the tissues $\dot{t}$ (reference) and $\ddot{t}$. (**c**) The discrete probability distributions $\{\dot{\boldsymbol{p}}^A, \ddot{\boldsymbol{p}}^A\}$ of the reference Atlas genes can capture the gene expression differences of the tissues. (**d, e**) The sorted gene expression profiles of the two tissues. Note that the rankings $\dot{\boldsymbol{\rho}}^A = \ddot{\boldsymbol{\rho}}^A = [2, 8, 3, 1, 11, 13, 10, 18, 15, 19]$ of the Atlas genes in the transcriptome are identical for both tissues, which results to $RCC = 1$. In this example, using the RCC of the Atlas genes alone fails to capture the transcriptomic differences of the two tissues

different tissue sets (e.g. *ex vivo* organ model samples versus human organ samples). We introduce a modified version of the *TSD*, which by considering the gene expression variability of the samples within the corresponding tissue sets provides statistically robust and accurate estimations of their TSDs.

Let us assume that we have two sets of tissue samples $\dot{\mathcal{T}} = \{\dot{t}_1, \dot{t}_2, \ldots, \dot{t}_V\}$ and $\ddot{\mathcal{T}} = \{\ddot{t}_1, \ddot{t}_2, \ldots, \ddot{t}_U\}$. Similarly to the notation used in Section 2.1, $\ddot{\mathcal{T}}$ corresponds to the set of tissue samples that we want to compare to the reference set $\dot{\mathcal{T}}$. For each tissue sample $\dot{t}_i \in \dot{\mathcal{T}}$, where $i = \{1, 2, \ldots, V\}$, and $\ddot{t}_j \in \ddot{\mathcal{T}}$, where $j = \{1, 2, \ldots, U\}$, we form the: (i) gene expression profile vectors $\dot{\boldsymbol{s}}_i$ and $\ddot{\boldsymbol{s}}_j$; (ii) the Atlas signature vectors $\dot{\boldsymbol{s}}_i^A$ and $\ddot{\boldsymbol{s}}_j^A$; (iii) the discrete probability distributions of the Atlas genes $\dot{\boldsymbol{p}}_i^A = [\dot{P}_{j1}^A, \dot{P}_{i2}^A, \ldots, \dot{P}_{iM}^A]$ and $\ddot{\boldsymbol{p}}_j^A = [\ddot{P}_{j1}^A, \ddot{P}_{j2}^A, \ldots, \ddot{P}_{jM}^A]$ and (iv) the matrices $\{\dot{\boldsymbol{\Pi}}, \ddot{\boldsymbol{\Pi}}\}$ that summarize the Atlas gene probability distributions of the corresponding samples.

$$\dot{\boldsymbol{\Pi}} = \begin{bmatrix} \dot{\boldsymbol{p}}_1^A \\ \dot{\boldsymbol{p}}_2^A \\ \vdots \\ \dot{\boldsymbol{p}}_V^A \end{bmatrix} = \begin{bmatrix} \dot{P}_{11}^A & \dot{P}_{12}^A & \cdots & \dot{P}_{1M}^A \\ \dot{P}_{21}^A & \dot{P}_{22}^A & \cdots & \dot{P}_{2M}^A \\ \vdots & \vdots & \ddots & \vdots \\ \dot{P}_{V1}^A & \dot{P}_{V2}^A & \cdots & \dot{P}_{VM}^A \end{bmatrix}$$
$$\ddot{\boldsymbol{\Pi}} = \begin{bmatrix} \ddot{\boldsymbol{p}}_1^A \\ \ddot{\boldsymbol{p}}_2^A \\ \vdots \\ \ddot{\boldsymbol{p}}_U^A \end{bmatrix} = \begin{bmatrix} \ddot{P}_{11}^A & \ddot{P}_{12}^A & \cdots & \ddot{P}_{1M}^A \\ \ddot{P}_{21}^A & \ddot{P}_{22}^A & \cdots & \ddot{P}_{2M}^A \\ \vdots & \vdots & \ddots & \vdots \\ \ddot{P}_{U1}^A & \ddot{P}_{U2}^A & \cdots & \ddot{P}_{UM}^A \end{bmatrix}. \tag{5}$$

2.3.1 The weighted Jensen–Shannon Divergence

In $\dot{\boldsymbol{\Pi}}$ and $\ddot{\boldsymbol{\Pi}}$, we assume that the probabilities of appearance of each Atlas gene (e.g. kth gene) across tissue samples (i.e. rows of matrices), were generated by a normal distribution (e.g. $\mathcal{N}(\dot{\mu}_k, \dot{\sigma}_k^2)$ and $\mathcal{N}(\ddot{\mu}_k, \ddot{\sigma}_k^2)$) with parameters:

$$\begin{aligned} \dot{\mu}_k &= \frac{1}{V}\sum_{i=1}^{V} \dot{P}_{ik}^A, \quad \dot{\sigma}_k^2 = \frac{1}{V}\sum_{i=1}^{V} (\dot{P}_{ik}^A - \dot{\mu}_k)^2 \\ \ddot{\mu}_k &= \frac{1}{U}\sum_{j=1}^{U} \ddot{P}_{jk}^A, \quad \ddot{\sigma}_k^2 = \frac{1}{U}\sum_{j=1}^{U} (\ddot{P}_{jk}^A - \ddot{\mu}_k)^2 \end{aligned} \tag{6}$$

Using this assumption, for each tissue $\{\dot{t}_i, \ddot{t}_j\}$, the likelihood of appearance of the kth Atlas gene can be calculated as:

$$\begin{aligned} F(\dot{P}_{ik}^A; \dot{\mu}_k, \dot{\sigma}_k^2) &= \frac{1}{\dot{\sigma}_k \sqrt{2\pi}} e^{-\frac{(\dot{P}_{ik}^A - \dot{\mu}_k)^2}{2\dot{\sigma}_k^2}} \\ F(\ddot{P}_{jk}^A; \ddot{\mu}_k, \ddot{\sigma}_k^2) &= \frac{1}{\ddot{\sigma}_k \sqrt{2\pi}} e^{-\frac{(\ddot{P}_{jk}^A - \ddot{\mu}_k)^2}{2\ddot{\sigma}_k^2}} \end{aligned} \tag{7}$$

The larger the $F(\dot{P}_{ik}^A; \dot{\mu}_k, \dot{\sigma}_k^2)$ $(F(\ddot{P}_{jk}^A; \ddot{\mu}_k, \ddot{\sigma}_k^2))$ the more ‘confident’ we are about the likelihood of appearance of the kth Atlas gene in the set of samples $\dot{\mathcal{T}}(\ddot{\mathcal{T}})$. We quantify our ‘confidence’ as:

$$\begin{aligned} \dot{\phi}_{ik}^A &= log_{10}(F(\dot{P}_{ik}^A; \dot{\mu}_k, \dot{\sigma}_k^2) + 1) \\ \ddot{\phi}_{jk}^A &= log_{10}(F(\ddot{P}_{jk}^A; \ddot{\mu}_k, \ddot{\sigma}_k^2) + 1) \end{aligned} \tag{8}$$

where to avoid negative ‘confidence’ values we added ‘1’ before taking the logarithm of the likelihoods.

To incorporate our ‘confidence’ about the likelihood of appearance of the Atlas genes in the *JSD*, we utilize a weighted version of the Shannon’s entropy H (Jianhua, 1991):

$$wJSD(\dot{\tilde{\boldsymbol{p}}}_i^A, \ddot{\tilde{\boldsymbol{p}}}_j^A) = H(\dot{w}_i \cdot \dot{\tilde{\boldsymbol{p}}}_i^A + \ddot{w}_j \cdot \ddot{\tilde{\boldsymbol{p}}}_j^A) - \dot{w}_i \cdot H(\dot{\tilde{\boldsymbol{p}}}_i^A) - \ddot{w}_j \cdot H(\ddot{\tilde{\boldsymbol{p}}}_j^A) \tag{9}$$

where $H(\tilde{\boldsymbol{p}}_h^A) = -\sum_{q=1}^{M} \tilde{P}_{hq}^A \cdot log_2 \tilde{P}_{hq}^A$.

$\dot{\tilde{\boldsymbol{p}}}_i^A = [\dot{\tilde{P}}_1^A, \dot{\tilde{P}}_2^A, \ldots, \dot{\tilde{P}}_M^A]$, and $\ddot{\tilde{\boldsymbol{p}}}_j^A = [\ddot{\tilde{P}}_1^A, \ddot{\tilde{P}}_2^A, \ldots, \ddot{\tilde{P}}_M^A]$ are the corresponding *weighted* discrete probability distributions that describe the probabilities of the reference Atlas genes to appear in the transcriptome of the corresponding tissues. The *weighted* probabilities are calculated as:

$$\tilde{\dot{P}}_{ik}^{A} = \frac{\dot{\phi}_{ik}^{A}\dot{P}_{ik}^{A}}{\sum\limits_{q=1}^{M} \dot{\phi}_{iq}^{A}\dot{P}_{iq}^{A}}, \quad \tilde{\ddot{P}}_{jk}^{A} = \frac{\ddot{\phi}_{jk}^{A}\ddot{P}_{jk}^{A}}{\sum\limits_{q=1}^{M} \ddot{\phi}_{jq}^{A}\ddot{P}_{jq}^{A}}, \quad \text{where} \quad k = \{1, 2, \ldots, M\}. \tag{10}$$

To calculate the *wJSD* (see Equation 9), we need also to determine the weights $\dot{w}_i$ and $\ddot{w}_j$ of the corresponding probability distributions $\tilde{\dot{\boldsymbol{p}}}_i^A$ and $\tilde{\ddot{\boldsymbol{p}}}_j^A$. Next, we present a novel method that quantifies our 'confidence' on how well the tissue samples $\{\dot{t}_i, \ddot{t}_j\}$ 'represent' their corresponding tissue sets $\{\dot{\mathcal{T}}, \ddot{\mathcal{T}}\}$, and appropriately adjust the weight values $\{\dot{w}_i, \ddot{w}_j\}$.

For each tissue set $\{\dot{\mathcal{T}}, \ddot{\mathcal{T}}\}$, we form the matrices $\dot{\boldsymbol{\Psi}}$ and $\ddot{\boldsymbol{\Psi}}$ where each of their rows (i.e. $\dot{\boldsymbol{\psi}}_i^A$ and $\ddot{\boldsymbol{\psi}}_j^A$ correspond to the standardized versions (e.g. *z*-scores) of the Atlas signature vectors ($\dot{\boldsymbol{s}}_i^A$, $\ddot{\boldsymbol{s}}_j^A$) of the corresponding tissue samples.

$$\dot{\boldsymbol{\Psi}} = \begin{bmatrix} \dot{\boldsymbol{\psi}}_1^A \\ \dot{\boldsymbol{\psi}}_2^A \\ \vdots \\ \dot{\boldsymbol{\psi}}_V^A \end{bmatrix} = \begin{bmatrix} \dot{\Psi}_{11}^A & \dot{\Psi}_{12}^A & \cdots & \dot{\Psi}_{1M}^A \\ \dot{\Psi}_{21}^A & \dot{\Psi}_{22}^A & \cdots & \dot{\Psi}_{2M}^A \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\Psi}_{V1}^A & \dot{\Psi}_{V2}^A & \cdots & \dot{\Psi}_{VM}^A \end{bmatrix} \tag{11}$$

$$\ddot{\boldsymbol{\Psi}} = \begin{bmatrix} \ddot{\boldsymbol{\psi}}_1^A \\ \ddot{\boldsymbol{\psi}}_2^A \\ \vdots \\ \ddot{\boldsymbol{\psi}}_U^A \end{bmatrix} = \begin{bmatrix} \ddot{\Psi}_{11}^A & \ddot{\Psi}_{12}^A & \cdots & \ddot{\Psi}_{1M}^A \\ \ddot{\Psi}_{21}^A & \ddot{\Psi}_{22}^A & \cdots & \ddot{\Psi}_{2M}^A \\ \vdots & \vdots & \ddots & \vdots \\ \ddot{\Psi}_{U1}^A & \ddot{\Psi}_{U2}^A & \cdots & \ddot{\Psi}_{UM}^A \end{bmatrix}$$

We assume that each standardized reference Atlas signature vector (e.g. $\dot{\boldsymbol{\psi}}_i^A$ and $\ddot{\boldsymbol{\psi}}_j^A$) is a random realization of a multivariate normal distribution [e.g. $\mathcal{N}(0, \dot{\boldsymbol{\Sigma}})$ and $\mathcal{N}(0, \ddot{\boldsymbol{\Sigma}})$]. To estimate the covariance matrices $\{\dot{\boldsymbol{\Sigma}}, \ddot{\boldsymbol{\Sigma}}\}$ of these distributions, we apply the *Graphical Lasso* (*GL*) algorithm to the corresponding data matrices $\{\dot{\boldsymbol{\Psi}}, \ddot{\boldsymbol{\Psi}}\}$. *GL* is a computationally efficient algorithm which has been extensively used to identify gene–gene interaction networks from RNA-seq datasets (Friedman *et al.*, 2008). Its main advantage, is its ability to estimate the precision matrix (i.e. $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$) even in cases where the number of samples is far less than the number of variables ($n \ll p$) which holds for transcriptomic datasets (i.e. number of samples $\ll$ number of genes). In such cases, other covariance estimation methods, such as Maximum Likelihood Estimation, fail since the sample (or empirical) covariance matrix $\boldsymbol{S}$ is rank deficient. *GL* addresses this limitation using the assumption that precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse (i.e. the graph structure of the variables' interactions is sparse). To estimate the precision matrix $\boldsymbol{\Omega}$, *GL* efficiently solves the following optimization problem:

$$\hat{\boldsymbol{\Omega}} = \text{argmin}_{\boldsymbol{\Omega} \geq 0}\{tr(S\boldsymbol{\Omega}) - \log(det(\boldsymbol{\Omega})) + \lambda||\boldsymbol{\Omega}||_1\} \tag{12}$$

where $||.||_1$ is the L_1-norm (i.e. the sum of the absolute values of the elements of $\boldsymbol{\Omega}$); $det(\boldsymbol{\Omega})$ is the determinant of $\boldsymbol{\Omega}$; λ is the sparsity parameter that controls the density (i.e. the number of edges) of the graphical model and S the $M \times M$ sample covariance matrix calculated as $\boldsymbol{S} = \boldsymbol{\Psi}^T\boldsymbol{\Psi}$. To optimally choose the value of λ, we use the *StARS* method (Liu *et al.*, 2010) which is stability-based approach for selecting the regularization parameter in high-dimensional graphical models.

Using the estimated covariance matrices of the multivariate normal distributions $\mathcal{N}(0, \dot{\boldsymbol{\Sigma}})$ and $\mathcal{N}(0, \ddot{\boldsymbol{\Sigma}})$, we calculate the likelihood of the tissue samples $\dot{t}_i$ and $\ddot{t}_j$ to belong to the corresponding distributions as:

$$Z(\dot{\boldsymbol{\psi}}_i^A; 0, \dot{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^M det(\dot{\boldsymbol{\Sigma}})}} e^{\frac{-(\dot{\boldsymbol{\psi}}_i^A)^T \dot{\boldsymbol{\Sigma}}^{-1} \dot{\boldsymbol{\psi}}_i^A}{2}}$$

$$Z(\ddot{\boldsymbol{\psi}}_j^A; 0, \ddot{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^M det(\ddot{\boldsymbol{\Sigma}})}} e^{\frac{-(\ddot{\boldsymbol{\psi}}_j^A)^T \ddot{\boldsymbol{\Sigma}}^{-1} \ddot{\boldsymbol{\psi}}_j^A}{2}} \tag{13}$$

The likelihoods $Z(\dot{\boldsymbol{\psi}}_i^A; 0, \dot{\boldsymbol{\Sigma}})$ and $Z(\ddot{\boldsymbol{\psi}}_j^A; 0, \ddot{\boldsymbol{\Sigma}})$, provide information on how 'confident' we are that the tissue samples $\dot{t}_i$ and $\ddot{t}_j$ are 'good' representatives of the corresponding tissue set $\{\dot{\mathcal{T}}, \ddot{\mathcal{T}}\}$. Using this information, we calculate the weights $\{\dot{w}_i, \ddot{w}_j\}$ as:

$$\begin{aligned} &\dot{w}_i = \frac{\dot{l}_i}{\dot{l}_i + \ddot{l}_j}, \quad \ddot{w}_j = \frac{\ddot{l}_j}{\dot{l}_i + \ddot{l}_j} \qquad \text{where} \\ &\dot{l}_i = log_{10}(V \cdot Z(\dot{\boldsymbol{\psi}}_i^A; 0, \dot{\boldsymbol{\Sigma}}) + 1) \\ &\ddot{l}_j = log_{10}(U \cdot Z(\ddot{\boldsymbol{\psi}}_j^A; 0, \ddot{\boldsymbol{\Sigma}}) + 1) \end{aligned} \tag{14}$$

where V and U are the number of samples in tissue sets $\dot{\mathcal{T}}$ and $\ddot{\mathcal{T}}$, respectively. Note that, $\dot{w}_i + \ddot{w}_j = 1$. Note that, for the calculation of the weight, we also consider the sample sizes of the corresponding tissue sets. More specifically, the larger the sample size the larger the weight, we assign to the corresponding distribution. This can be justified if we consider that larger the sample sizes provide more confidence about the accuracy of the estimated parameters (i.e. covariance matrix) of the multivariate normal distribution.

2.3.2 The weighted Rankings Correlation Coefficient

The second term of the *TSD* (see Equation 4) is the *RCD* between the reference Atlas genes ranking vectors $\{\dot{\boldsymbol{\rho}}^A, \ddot{\boldsymbol{\rho}}^A\}$ of the compared tissues $\dot{t}, \ddot{t}$. For the case where we have sets of tissue samples (i.e. $\dot{\mathcal{T}}$ and $\ddot{\mathcal{T}}$), we use weighted Rankings Correlation Coefficient (*wRCC*) which can be calculated as:

$$wRCC\left(\dot{\boldsymbol{\rho}}_i^A, \ddot{\boldsymbol{\rho}}_j^A\right) = \frac{\sum\limits_{q=1}^{M} \left[\beta_q \cdot \left(\dot{r}_{iq}^A - \bar{\dot{r}}_i^A\right)\left(\ddot{r}_{jq}^A - \bar{\ddot{r}}_j^A\right)\right]}{\sqrt{\sum\limits_{q=1}^{M} \left[\beta_q \cdot \left(\dot{r}_{iq}^A - \bar{\dot{r}}_i^A\right)^2\right] \sum\limits_{q=1}^{M} \left[\beta_q \cdot \left(\ddot{r}_{jq}^A - \bar{\ddot{r}}_j^A\right)^2\right]}}$$

$$\text{where} \qquad \bar{\dot{r}}_i^A = \frac{\sum\limits_{q=1}^{M} \beta_q \cdot \dot{r}_{iq}^A}{\sum\limits_{q=1}^{M} \beta_q}, \quad \bar{\ddot{r}}_j^A = \frac{\sum\limits_{q=1}^{M} \beta_q \cdot \ddot{r}_{jq}^A}{\sum\limits_{q=1}^{M} \beta_q} \tag{15}$$

are the corresponding weighted means of the rankings of the Atlas genes. In Equation (15), β_q, where $q = \{1, \ldots, M\}$, are the weights which assign different 'confidence' to the corresponding rankings of the M Atlas genes. To calculate the values of these weights, we propose the following method.

Using the ranking vectors $\{\dot{\boldsymbol{\rho}}_i^A, \ddot{\boldsymbol{\rho}}_j^A\}$ of the Atlas genes of the tissue samples that contained in each set (i.e. $\dot{\mathcal{T}}$ and $\ddot{\mathcal{T}}$), we form the following matrices:

$$\dot{\boldsymbol{R}} = \begin{bmatrix} \dot{\boldsymbol{\rho}}_1^A \\ \dot{\boldsymbol{\rho}}_2^A \\ \vdots \\ \dot{\boldsymbol{\rho}}_V^A \end{bmatrix} = \begin{bmatrix} \dot{r}_{11}^A & \dot{r}_{12}^A & \cdots & \dot{r}_{1M}^A \\ \dot{r}_{21}^A & \dot{r}_{22}^A & \cdots & \dot{r}_{2M}^A \\ \vdots & \vdots & \ddots & \vdots \\ \dot{r}_{V1}^A & \dot{r}_{V2}^A & \cdots & \dot{r}_{VM}^A \end{bmatrix}$$

$$\ddot{\boldsymbol{R}} = \begin{bmatrix} \ddot{\boldsymbol{\rho}}_1^A \\ \ddot{\boldsymbol{\rho}}_2^A \\ \vdots \\ \ddot{\boldsymbol{\rho}}_U^A \end{bmatrix} = \begin{bmatrix} \ddot{r}_{11}^A & \ddot{r}_{12}^A & \cdots & \ddot{r}_{1M}^A \\ \ddot{r}_{21}^A & \ddot{r}_{22}^A & \cdots & \ddot{r}_{2M}^A \\ \vdots & \vdots & \ddots & \vdots \\ \ddot{r}_{U1}^A & \ddot{r}_{U2}^A & \cdots & \ddot{r}_{UM}^A \end{bmatrix}. \tag{16}$$

Each row of matrices $\dot{\boldsymbol{R}}$ and $\ddot{\boldsymbol{R}}$ contains the rankings of the expressions of the Atlas genes in the transcriptome of the corresponding tissue sample, and each column includes the rankings of the expressions of a specific Atlas gene across the tissue samples of the corresponding set.

In matrices $\dot{\boldsymbol{R}}$ and $\ddot{\boldsymbol{R}}$, we assume that the rankings of each Atlas gene (e.g. kth gene) across tissues were generated by a normal distribution [e.g. $\mathcal{N}(\dot{\gamma}_k, \dot{\delta}_k^2)$ and $\mathcal{N}(\ddot{\gamma}_k, \ddot{\delta}_k^2)$] with parameters:

$$\begin{aligned} \dot{\gamma}_k &= \frac{1}{V}\sum_{i=1}^{V} \dot{r}_{ik}^A, \quad \dot{\delta}_k^2 = \frac{1}{V}\sum_{i=1}^{V} (\dot{r}_{ik}^A - \dot{\gamma}_k)^2 \\ \ddot{\gamma}_k &= \frac{1}{U}\sum_{j=1}^{U} \ddot{r}_{jk}^A, \quad \ddot{\delta}_k^2 = \frac{1}{U}\sum_{j=1}^{U} (\ddot{r}_{jk}^A - \ddot{\gamma}_k)^2 \end{aligned} \tag{17}$$

Using this assumption, the likelihood about the rankings of the kth Atlas gene can be calculated as:

$$\begin{aligned} \Xi(\dot{r}_{ik}^A; \dot{\gamma}_k, \dot{\delta}_k^2) &= \frac{1}{\dot{\delta}_k\sqrt{2\pi}} e^{-\frac{(\dot{r}_{ik}^A - \dot{\gamma}_k)^2}{2\dot{\delta}_k^2}} \\ \Xi(\ddot{r}_{jk}^A; \ddot{\gamma}_k, \ddot{\delta}_k^2) &= \frac{1}{\ddot{\delta}_k\sqrt{2\pi}} e^{-\frac{(\ddot{r}_{jk}^A - \ddot{\gamma}_k)^2}{2\ddot{\delta}_k^2}} \end{aligned} \tag{18}$$

The larger the $\Xi(\dot{r}_{ik}^A; \dot{\gamma}_k, \dot{\delta}_k^2)$ ($\Xi(\ddot{r}_{jk}^A; \ddot{\gamma}_k, \ddot{\delta}_k^2)$) the more 'confident' we are about the ranking $\dot{r}_{ik}^A$ ($\ddot{r}_{jk}^A$) of the kth Atlas gene, We quantify our 'confidence' as:

$$\begin{aligned} \dot{\tau}_{ik}^A &= log_{10}(\Xi(\dot{r}_{ik}^A; \dot{\gamma}_k, \dot{\delta}_k^2) + 1) \\ \ddot{\tau}_{jk}^A &= log_{10}(\Xi(\ddot{r}_{jk}^A; \ddot{\gamma}_k, \ddot{\delta}_k^2) + 1) \end{aligned} \tag{19}$$

To avoid negative 'confidence' values, we added '1' before taking the logarithm of the likelihoods. Using the $\dot{\tau}_{ik}^A$ and $\ddot{\tau}_{jk}^A$, we calculate the importance weight of the kth Atlas gene as:

$$\beta_k = \dot{\tau}_{ik} + \ddot{\tau}_{jk}. \tag{20}$$

By applying these weights to Equation (15), we can calculate the $wRCC$ which also takes values in range [-1,1].

2.3.3 The weighted *TSD*

After calculating $wJSD$ and $wRCC$, we can calculate the weighted version of the transcriptomic signature distance ($wTSD$) as:

$$wTSD(\dot{t}_i, \ddot{t}_j) = \frac{1}{2}\cdot SR\text{-}wJSD(\dot{\tilde{\boldsymbol{p}}}_i^A, \ddot{\tilde{\boldsymbol{p}}}_j^A) + \frac{1}{2}\cdot wRCD(\dot{\boldsymbol{\rho}}_i^A, \ddot{\boldsymbol{\rho}}_j^A), \tag{21}$$

where $wRCD = \sqrt{1 - ReLU(wRCC)}\}$ is the *weighted Rankings Correlation Distance (wRCD)*. It can be shown that $wRCD$ is a 'true' metric (Jiaxing *et al.*, 2019). For the calculation of the $wRCD$, we assume that if two vectors have $wRCC < 0$ (i.e. anti-correlated) then their RCD is maximum (i.e. 1). Similar to the TSD (see Equation 4), $wTSD$ takes its values in $[0, 1]$.

In Equations (4) and (21), *(w)TSD* is defined as the average of *SR-(w)JSD* and *(w)RCD*. However, this can be easily generalized to a weighted average by assigned weights (say w_1 and w_2) to *SR-(w)JSD* and *(w)RCD*, respectively (assuming that the weights are positive and sum to one). This more general version provides the flexibility to assign different importance to *SR-(w)JSD* and *(w)RCD* if the user has reasons to believe that this is justified for a particular use-case.

3 Results and discussion

We present here extensive results generated using publicly available RNA-seq datasets demonstrating the validity and value of the proposed TSD (i) for measuring the distance between different organ tissues, (ii) for assessing the distance between healthy and diseased organ tissues and (iii) for evaluating the similarity of organ models (e.g. organoids, animals, etc.) to the corresponding human organ. Finally, we compare *(w)TSD* to other traditional metrics which have been used for measuring transcriptomic distance/similarity.

3.1 Using *wTSD* with real data to assess distance of human organ tissues

In Section 2.2, we used two hypothetical scenarios to illustrate that using *SR-JSD* or RCC alone may fail to adequately capture the transcriptomic differences of tissue samples. In this section, we use real data to show this ineffectiveness and justify the advantages of the proposed $wTSD$ as a higher-resolution method. For this purpose, we used the publicly available dataset in GEO GSE120795, also presented in the study by Suntsova *et al.* (2019), a comprehensive gene expression database of normal human tissues based on uniformly screened RNA-seq data. This database includes 142 tissue samples taken from 20 organs of healthy human donors of different ages, collected no later than 36 h after death. From dataset GSE120795, we selected the organ tissues that have been characterized by the HPA project, and also discarded organ tissues with less than four samples ($n < 4$) passing the quality control. The $n \geq 4$ condition was necessary to guarantee the execution of the *StARS* method (Liu *et al.*, 2010), which estimates the sparsity parameter of the graphical lasso algorithm (see Section 2.3) using random subsampling. The HPA includes information for only 15 out of the 20 organs in the database, and these were used in our analysis. As we can see in Supplementary Table S1 in Supplementary Material, the number of samples as well as the number of signature genes identified by the HPA project for each one of the 15 organs used vary considerably. The $wTSD$ distance is designed to deal with this kind of situations in a principled manner.

The Heatmaps of Figure 3 depict the mean inter-/intra-organ distances. Each row corresponds to a specific organ being used as reference whose HPA signature genes were utilized to calculate the distances of the other organs (columns) from it. As expected, the mean intra-organ tissue sample distances (main diagonal elements) are smaller than the corresponding inter-organ distances (off-diagonal elements of the same row). Observe that these Heatmaps (matrices) are not symmetric. This happens because we are measuring the distance of different types of tissues from a reference tissue (row) and each reference tissue is represented by a different Atlas signature gene vector. However, this asymmetry should be expected as we will explain using the following example: let us assume w.l.o.g. that $\dot{s}^A$, $\ddot{s}^A$ are samples extracted from liver and intestine tissues, respectively. $TSD(\dot{s}^A, \ddot{s}^A)$ measures how similar is the tissue sample $\ddot{s}^A$ to the reference sample $\dot{s}^A$. In other words, it quantifies the 'liver-ness' of the intestine sample $\ddot{s}^A$. Similarly, the $TSD(\ddot{s}^A, \dot{s}^A)$ is trying to quantify the 'intestine-ness' of the liver sample $\dot{s}^A$. TSD uses different distance spaces to measure 'liver-ness' and the 'intestine-ness' which are determined by the tissue-specific Atlas signature genes. Of course, TSD can become symmetric and a true-metric (see Section 3.3) if we force it to use the same set of genes for all tissues, but as we have shown (see Supplementary Section S2.2 in Supplementary Material) using only the HPA characteristic genes improves our ability to distinguish between different groups of tissue samples.

If we examine carefully the corresponding rows of the *SR-wJSD* and $wRCD$ heatmaps in Figure 3a and b, we see significant element-wise differences across each row. This indicates that *SR-wJSD* and $wRCD$ capture different aspects of transcriptome dissimilarities. To better illustrate this fact, Figure 4a,b and c,d depicts the distances of other organs from Lung and Kidney (references), respectively, in a 2D-space where *SR-wJSD* and $wRCD$ are used as coordinates. In these plots, each organ's name label is centered at the mean value of the corresponding pairwise tissue sample distances (*SR-wJSD* and $wRCD$) from the reference organ. In the zoomed-in version of Figure 4b, we see that the {Liver, Brain} and {Small Intestine, Thyroid} pairs have almost equal $wRCDs$ but different *SR-wJSDs* coordinate values. On the other hand, in the zoomed-in Figure 4d, {Thyroid, Esophagus}, {Pancreas, Small Intestine} as well as {Bladder and Prostate} pairs have almost equal *SR-wJSDs* but different $wRCDs$ coordinate values. Based on these observations, it is

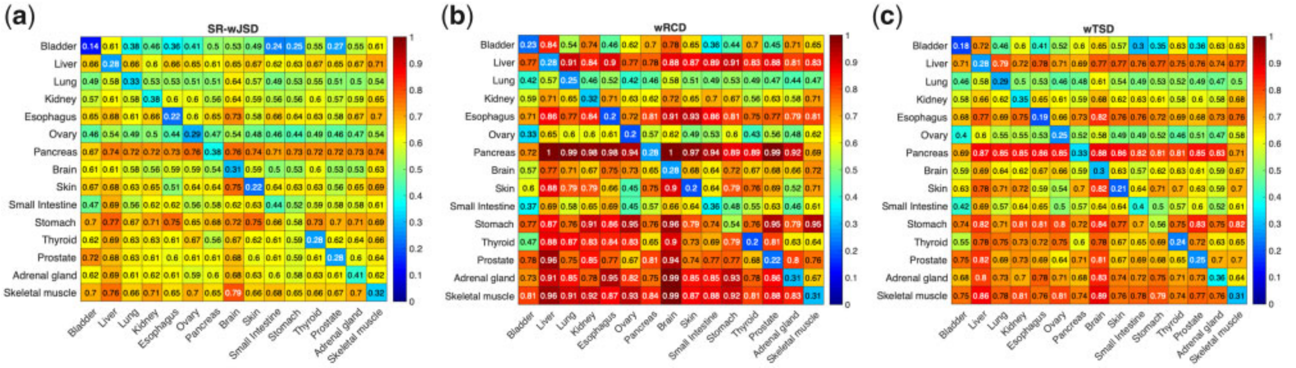

Fig. 3. Heatmaps of means of pairwise distances between sample groups of organ tissues. (**a**) *SR-wJSD*, (**b**) *wRCD* and (**c**) *wTSD*. The rows correspond to the reference organs whose Atlas genes were used in the distance calculations of the other tissues (columns) from the reference

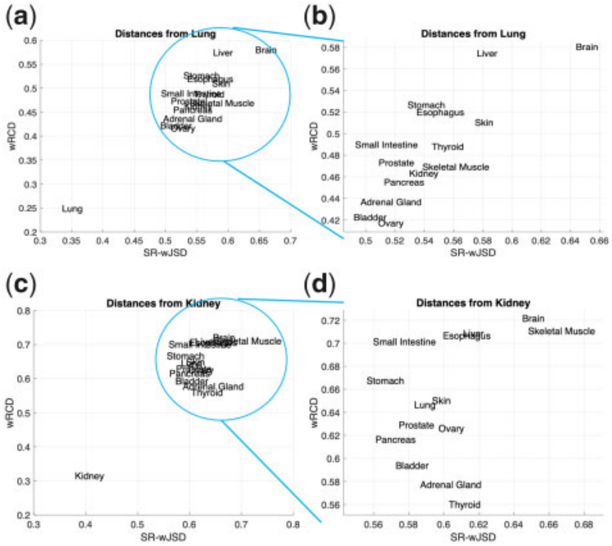

Fig. 4. The inter-organ distances of other tissues from: (**a**, **b**) Lung and (**c**, **d**) Kidney used as reference organ. The name label of each organ is centered at the mean value of the pairwise distances (*SR-wJSD* and *wRCD*) between that organ's tissue samples and samples of the reference organ (Lung or Kidney, respectively)

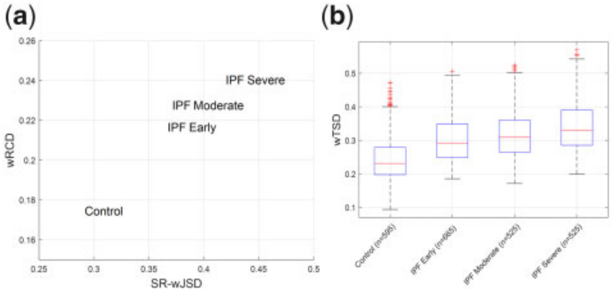

Fig. 5. The distances of the different IPF progression stages from the healthy lung. (**a**) The labels of IPF stages are centered at the coordinates determined by the means of the pairwise distances (*SR-wJSD* and *wRCD*) between tissue samples in the corresponding groups and the Controls. (**b**) Boxplots summarizing the distributions of the corresponding pairwise *wTSD* distances. Here 'n' denotes the number of pairwise distances (*wTSD*) where one sample in the pair belongs to the Controls and the other either in the Controls group or in an IPF progression stage group

clear that the proposed new distance, *wTSD*, which combines the *SR-wJSDs* and *wRCDs* information while also considering the intra-class tissue samples variability of every organ (see Section 2), provides a higher-resolution picture of the reality.

3.2 Using *wTSD* to assess tissue distance of disease subtypes and progression stages

In this section, we present results demonstrating that *TSD* can be used to resolve transcriptomic distance of normal tissues from tissues of disease subtypes as well as tissues of different disease progression stages. For this purpose, we are using publicly available RNA-seq datasets characterizing two different diseases: idiopathic pulmonary fibrosis (IPF) and liver cancer.

3.2.1 Using *wTSD* with IPF dataset

Recently published research (McDonough *et al.*, 2019) has studied the progression mechanisms of IPF, a lethal chronic lung disease which progresses the fibrosis in lungs over time, causing serious breathing difficulties. IPF affects 13–20 per 100k people worldwide. According to the National Institute of Health, about 30k–40k patients in the USA are diagnosed with IPF every year. More than 50% of IPF patients die within 3–5 years after the initial diagnosis (Kim *et al.*, 2006; Lederer and Martinez, 2018). The RNA-seq dataset of this study, available in NCBI's Gene Expression Omnibus (GEO GSE124685), consists of 84 samples classified in the following four categories: (i) Controls ($n = 35$), (ii) IPF Early ($n = 19$), (iii) IPF Moderate ($n = 15$) and (iv) IPF Severe ($n = 15$). The samples' categorization was made based on the extent of lung fibrosis, assessed using microCT quantitative imaging and tissue histology (McDonough *et al.*, 2019). Using this dataset and the information of the 239 genes which, according to HPA, can be considered as the transcriptomic signature of the healthy human lung (see Supplementary Table S1), we calculated the transcriptomic distances (*SR-wJSD*, *wRCD* and *wTSD*) of all pairs of tissue samples, one sample belonging in the Controls (reference) group and the other in the diseased groups.

Figure 5a shows the transcriptomic distances (*SR-wJSD* and *wRCD*) of the different IPF progression stages from the healthy lung tissues. The label of each IPF progression stage name is centered at the coordinates of the mean value of the pairwise distances where one sample in the pair belongs to the Controls and the other either in the Controls group or in an IPF progression stage group (all pair combinations considered). Figure 5b shows boxplots of the corresponding distributions of the pairwise *wTSD* distances. The results indicate that as the severity of IPF increases, the corresponding TSD from the Control class also increases. This fact demonstrates the interpretability of the proposed distance. Supplementary Table S2 in Supplementary Material summarizes the results of the two-sample *t*-test between the corresponding distributions of the pairwise *wTSD* distances (presented in Fig. 5b). The decision of the test is equal to 1 ($h = 1$) if the test rejects the null hypothesis (that the groups of the distances have equal means and equal but unknown variances) at the 1% significance level. Supplementary Table S2 results clearly show that *wTSD* can successfully identify the different IPF progression stages based on the HPA lung signature genes. Moreover, the *wTSD* differences are statistically significant for all comparisons

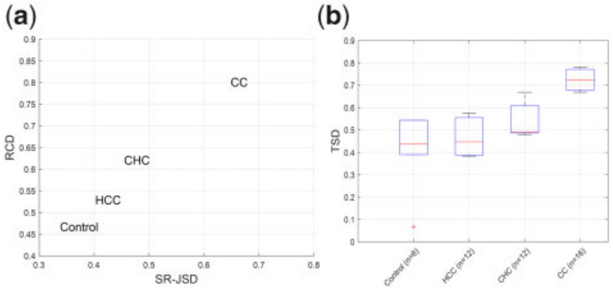

Fig. 6. Distances of different liver cancer subtypes stages from the healthy liver. (**a**) The liver cancer subtype name labels are centered at the coordinates determined by the means of the pairwise distances (*SR-wJSD* and *wRCD*) between the samples in the corresponding subtype groups and the Controls. (**b**) Boxplots summarizing the distributions of the corresponding pairwise *TSD* distances. Here 'n' denotes the number of pairwise distances (*TSD*) where one sample in the pair belongs to the Controls and the other either in the Controls group or in a Liver Cancer Subtype group

between (i) the Control and IPF stages, and (ii) different IPF stages, a fact that demonstrates the ability of *wTSD* to capture the transcriptomic differences of the corresponding categories.

3.2.2 Using *TSD* with human liver cancer dataset

In a recent study (Broutier *et al.*, 2017), human primary liver cancer-derived organoids were used to recapitulate the pathophysiology of human liver tumors. From the provided RNA-seq dataset (GEO GSE84073), we extracted samples for healthy human liver tissue (Controls) and different human liver tumor subtypes, in particular: Hepato-Cellular Carcinoma (HCC), Cholangio-Carcinoma (CC) and combined HCC/CC (CHC). The number of tissue samples in each group was relatively small: (i) Controls ($n = 4$), (ii) HCC ($n = 3$), (iii) CC ($n = 4$) and (iv) CHC ($n = 3$). We also used the expression information of the 936 genes for every sample, which, according to HPA, form a transcriptomic signature of the healthy human liver (see Supplementary Table S1). Next, using as reference organ, the healthy human liver, we computed the corresponding pairwise distances (*SR-JSD*, *RCD* and *TSD*) between its samples and samples in the different cancer subtype groups. We remark here that due to the limited number of samples in the cancer groups, we decided to use the simple version (not weighted) of the *TSD*.

Figure 6a shows the transcriptomic distances (*SR-JSD* and *RCD*) of the different cancer subtypes from the healthy liver. Each cancer subtype's name label is centered at the coordinates of the mean value of the pairwise distances where one sample in the pair belongs to the Controls and the other either in the Controls group or in a liver cancer subtype group. Figure 6b shows boxplots of the distributions of these pairwise *TSD* distances. These results demonstrate the ability of *TSD* to represent the distance difference of controls from the tumor subtypes. It is interesting to remark that the distance of the CHC group (tumor tissue, which is a combination of HCC and CC) from the Controls is in-between the corresponding distances of the HCC and CC, a fact that conforms with our human intuition. Moreover, Supplementary Table S3 in the Supplementary Material summarizes the results of the two-sample *t*-test between the corresponding distributions of the pairwise *TSD* distances (presented in Fig. 6b). In Supplementary Table S3, the decision of the test is equal to 1 ($h = 1$) if the test rejects the null hypothesis at the 5% significance level. The results show that all comparisons except one (Control versus HCC) have significantly different *TSD* distances, which indicates the ability of *TSD* to identify the transcriptomic differences of the corresponding groups.

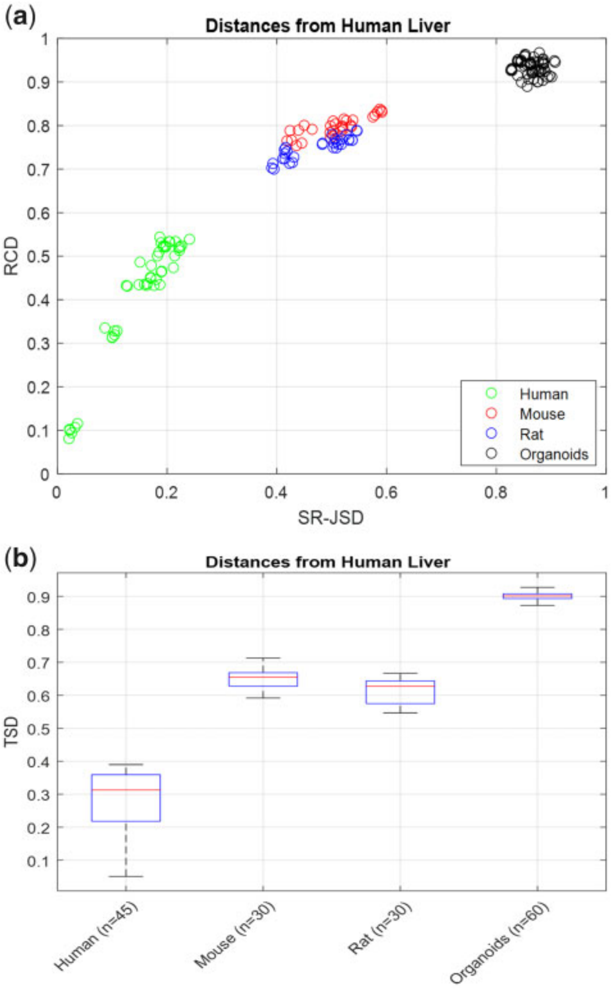

Fig. 7. The distances of the different liver models from the human liver. (**a**) The pairwise distances (*SR-JSD* and *RCD*) of organ model tissue samples [mouse (red circles), rat (blue circles), organoids (black circles)] from the corresponding healthy human organ tissue samples. (**b**) Boxplots summarizing the distributions of corresponding pairwise *TSD* distances. Here 'n' denotes the number of pairwise distances (*TSD*) where one sample in the pair belongs to the human liver group and the other either in the human liver group or in a liver organ model

3.3 Assessing distance of human organs from different organ models

In this section, we show that the proposed TSD can be used to assess the 'physiological relevance' of different organ models to a human organ based on RNA-seq data. Specifically, we computed the *TSD* of: (i) *Mus musculus* (mouse) organs, (ii) *Rattus norvegicus* (rat) organs and (iii) human-derived organoids from the human liver and kidney used as reference organs. We obtained the RNA-seq data for the healthy human organ tissues from the publicly available database developed by Suntsova *et al.* (2019) (described in Section 3.1). The number of available samples for each healthy human organ is provided in Supplementary Table S1. For mouse and rat, we obtained RNA-seq data from the database developed by Sollner *et al.* (2017). For both species, the number of available samples per organ was equal to three. Finally, we retrieved RNA-seq data for healthy human liver- and kidney-derived organoids from the publicly available datasets (GSE84073 and GSE99582) presented in the studies by Broutier *et al.* (2017) and Phipson *et al.* (2019), respectively. The number of samples of the healthy liver- and kidney-derived organoids were six and three, respectively. To compare the transcriptomic signatures between the different species, we associated the mouse and rat genes to human homologous genes using the R-package *biomart* (Smedley *et al.*, 2015). Due to the limited number of samples ($n = 3$) in some of the categories under comparison, we used the 'simple' version of *TSD* (not weighted).

Figures 7a and 8a show, for the liver and kidney datasets, respectively, the pairwise distances (*SR-JSD* and *RCD*) of organ model samples (mouse, rat and organoids) from healthy human organ

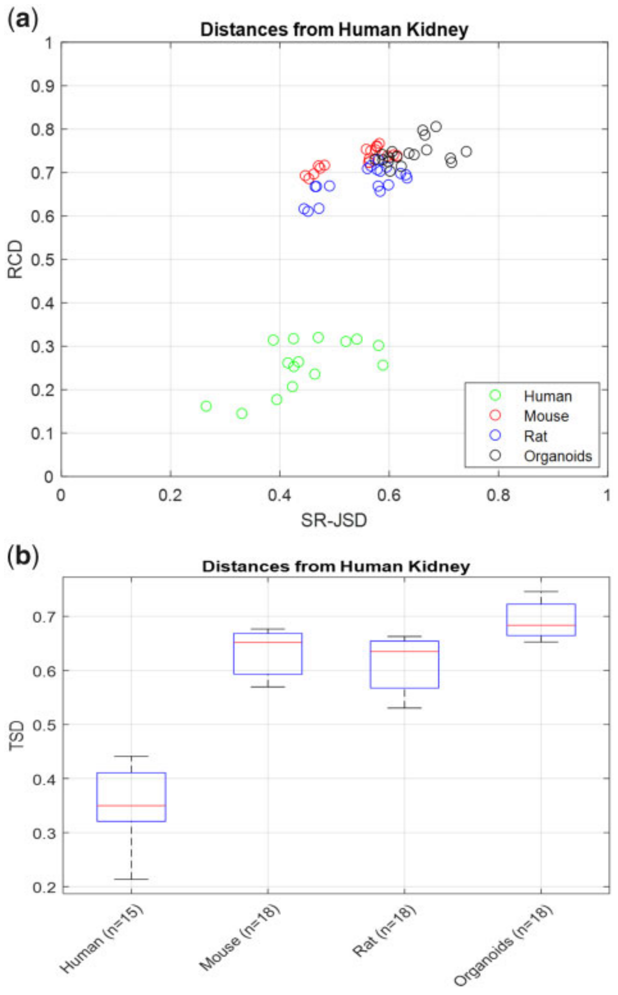

Fig. 8. The distances of different kidney models from the human kidney. (**a**) Pairwise distances (*SR-JSD* and *RCD*) of organ model samples represented as small circles; mouse (red), rat (blue), organoids (black) from the corresponding healthy human organ tissue samples. (**b**) Boxplots are summarizing the distributions of corresponding pairwise *TSD* distances. Here 'n' denotes the number of pairwise distances (*TSD*) where one sample in the pair belongs to the human kidney group and the other either in the human kidney group or in a kidney organ model

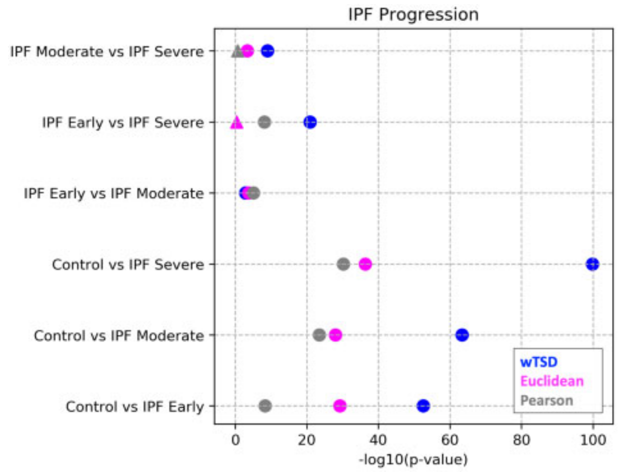

Fig. 9. IPF progression dataset. Scatter plot summarizing the results of metric comparisons when using the HPA tissue signature genes [see Supplementary Tables S6–S11 in Supplementary Section S2.2.1). Using *wTSD* (blue) we can achieve much smaller *P*-values [larger $-\log_{10}$ (*P*-values)] in almost all considered comparisons (rows). Circular (Triangular) shape glyphs are used to mark when the null hypothesis is (is not) rejected. Only *wTSD* passes the statistical significance test in all tissue group comparisons. Moreover, it achieves much smaller *P*-values in all cases but one

tissue samples. In both Figures, a circle depicts the distance of a model tissue sample (represented by color black, blue and red) to a control sample. On the other hand, green circles are used for pairwise distances between samples of the control group. The boxplots in Figures 7b and 8b summarize the distributions of these pairwise distances. The results indicate that the liver tissue samples of mouse and rat are transcriptomically closer to the human liver than the human-derived liver organoid samples. However, for the kidney, the *TSDs* of the mouse, rat and human-derived kidney organoids, from the healthy human kidney, are very similar. Another interesting observation is that for both organs, the transcriptomic distances of mouse and rat from the corresponding healthy human organs are similar. As shown in the study by Sudmant *et al.* (2015), mouse and rat have similar transcriptomic profiles between homologous organ tissues which is also confirmed by Figures 7 and 8. Supplementary Tables S4 and S5 summarize for liver and kidney, respectively, the results of the two-sample *t*-tests between the corresponding distributions (see Figs 7b and 8b) of the pairwise *TSD* distances.

At this point, we should mention a very interesting property of the *(w)TSD*. When all the compared tissues (e.g. $\{\dot{t},\ddot{t},\dddot{t}\}$) represent the same organ (i.e. are all compared based using the same set of signature genes) as the case here for liver and kidney, *(w)TSD* satisfies all the necessary conditions of a true metric. Based on *SR-(w)JSD* and *(w)RCD* characteristics (Jianhua, 1991; Jiaxing *et al.*, 2019), it is a matter of simple algebra to show that:

1. Is symmetric: $(w)TSD(\dot{t},\ddot{t}) = (w)TSD(\ddot{t},\dot{t})$
2. Is non-negative: $0 \leq (w)TSD(\dot{t},\ddot{t}) \leq 1$
3. $(w)TSD(\dot{t},\ddot{t}) = 0$ *iff*:
   SR-$(w)JSD(\dot{\boldsymbol{p}}^A,\ddot{\boldsymbol{p}}^A) = 0$ and $(w)RCD(\dot{\boldsymbol{p}}^A,\ddot{\boldsymbol{p}}^A) = 0$
4. $(w)TSD(\dot{t},\ddot{t}) \leq (w)TSD(\dot{t},\dddot{t}) + (w)TSD(\dddot{t},\dot{t})$ (triangle inequality), $\forall$ triplet of tissues $\{\dot{t},\ddot{t},\dddot{t}\}$.

3.4 Comparison to other metrics

Comparing *TSD* to other traditional distance metrics is hampered by the fact that there is no dataset that can be used as an absolute 'ground truth', providing the true transcriptomic 'distances' between organ tissues. In this section, we compare the ability of *TSD*, Euclidean distance (*ED*) and Pearson correlation (*PC*) to discriminate different groups of evidently different tissue samples using the 'IPF progression' and the 'Human Liver Cancer' datasets (see Sections 3.2.1 and 3.2.2). To compare these methods, we use the resulting *P*-values after performing two-sample *t*-tests to the calculated distances/correlations between the Controls group (reference) and the rest of the tissue sample groups. The smaller the corresponding *P*-values, the better the ability of a method to resolve the different subgroups, which implies an advantage in translating transcriptomic expression differences to a distance.

McDonough *et al.* (2019) identified significant transcriptomic differences between Control and IPF progression groups of samples. These differences were also reflected to the PCA plot provided in Figure 1E of their paper, where we observe that the distance of the various IPF progression groups to the Control group increases with disease progression. Figure 9 shows that when using the HPA tissue-characteristic gene signatures, *wTSD* can better capture this behavior compared to *PC*, while *ED* fails. More specifically, unlike *ED*, the *P*-values between the Control and IPF progression groups in *wTSD* tend to decrease significantly with the progression of the disease which shows the better discrimination of the groups (e.g. the *P*-value of Control versus IPF-Moderate is smaller than the *P*-value of the Control versus IPF-Early). For all comparison cases but one, *wTSD* achieves much smaller *P*-values than the other methods, which demonstrates its ability to discern better the existing group differences.

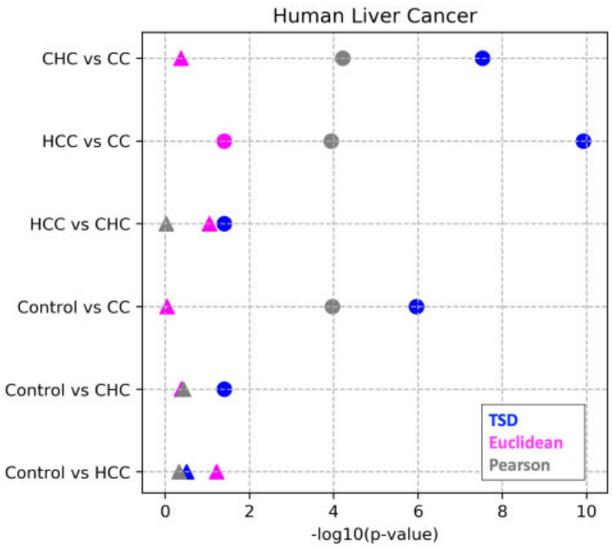

Fig. 10. Liver cancer dataset. Scatter plot summarizing the results of different comparisons when using the HPA characteristic genes as signature for each tissue (see also Supplementary Tables S12–S17 in Supplementary Section S2.2.2). Using *TSD* (blue) we achieve much smaller *P*-values [larger $-\log_{10}$ (*P*-values)] in almost all group comparisons. Circular (Triangular) shape glyphs are used to mark when the null hypothesis is (not) rejected. Moreover, using *TSD* passes the statistical significance test in almost all tissue group comparisons as opposed to using the Euclidean distance or Pearson's correlation

Broutier *et al.* (2017) studied how three of the most common human primary liver cancer (PLC)-derived organoids can recapitulate the pathophysiology of human liver tumors. The results show that *PLC-derived organoid cultures preserve the histological and genomic features of the original tumor, allowing the discrimination between different tumor tissues and subtypes*. Based on these findings, we tested the ability of *TSD*, *ED* and *PC* to discriminate the different cancer subtypes. Figure 10 depicts the corresponding *P*-values when using the liver HPA signature genes. Similar to the IPF disease progression results, *TSD* achieves in almost all cases better discrimination performance (i.e. smaller *P*-values) between the groups. Moreover, *TSD* can identify statistically significant differences in 5 out of 6 comparisons (fails only in one case where all methods fail). On the contrary, *ED* and *PC* can find significant distances in only one and three comparisons, respectively, out of the six. This demonstrates the ability of *TSD* to resolve better the different existing tissue groups even in cases where the number of available samples per group is small.

Overall, due to its better discriminative ability *TSD* can be considered as a more consistent and 'higher resolution' transcriptomic distance which is also confirmed by the boxplots provided in Supplementary Section S2.1. In addition, we have shown using both datasets that a significantly better discrimination is achieved between the sample groups when using the HPA characteristic genes only as tissue signatures, relatively to using all genes. Although *TSD* can also work with the full gene set, our results demonstrate the advantage of exploiting the HPA provided information while at the same time achieving substantial computational savings (see Supplementary Section S2.2 for details).

In this section, we presented how to use *(w)TSD* as a distance to measure the similarity between organ tissue samples in different scenarios arising in practice. We remark that we can also use *(w)TSD* in a variety of other situations for measuring distance of biological samples as long as we have access to their gene expression data and information about their signature genes. For example, *(w)TSD* can be used to assess the distances between different cell types using single-cell RNA-seq data and information about cell-type-specific signature gene sets that we can retrieve from publicly available databases or the expanding literature on the subject (Kotliar *et al.*, 2019; Merienne *et al.*, 2019; Thul *et al.*, 2017).

3.4.1 Availability of the software

The computation of both *TSD* versions (simple and weighted) has been implemented in R.

(https://github.com/Cod3B3nd3R/Transcriptomic-Signature-Distance).

4 Conclusions

We presented the development and utility of *TSD*, a new distance we introduced for quantifying the transcriptomic similarity of organ tissues. The development of *TSD* is grounded on information theory and advanced statistics. Also, *TSD* exploits the availability of 'signature' genes for human organs, provided in the well-curated publicly available HPA database, to emphasize organ tissue differences and mask the effects of measurement noise and inter-donor variability in the distance calculations. We also presented a novel method that considers the gene expression and ranking variations across homologous tissue samples and appropriately incorporates this information into the distance calculations. We justified the effectiveness and reliability of the proposed distance and evaluated its performance using many different publicly available RNA-seq datasets. We presented extensive experimental results that validate the ability of *TSD* to represent distances between different organ tissues coherently. Moreover, we have shown how *TSD* can be used to assess the distance of alternative organ model technologies (*in vivo*, *ex vivo*, etc.) to the corresponding human organ. To the best of our knowledge, *TSD* is the first distance based on information theory, which allows us to assess the similarity of model organ tissue samples to the human organ they represent based on a reference gene set. We are confident that *TSD* can be a valuable tool in many disciplines, such as tissue engineering, micro-physiological systems design, single-cell type comparison and so on. For this purpose, we make available openly the R code that computes *TSD* and *wTSD* for pairs of tissue samples, either in isolation or as members of two tissue sample groups. We also provide instructions that can be used to reproduce all the results presented in the article.

Funding

Research reported in this publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UG3TR002188. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

Aibar,S. *et al.* (2016) Identification of expression patterns in the progression of disease stages by integration of transcriptomic data. *BMC Bioinformatics*, **17**, 432.

Broutier,L. *et al.* (2017) Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. *Nat. Med.*, **23**, 1424–1435.

Chen,B. *et al.* (2015) Relating chemical structure to cellular response: an integrative analysis of gene expression, bioactivity, and structural data across 11,000 compounds. *CPT Pharmacomet. Syst. Pharmacol.*, **4**, 576–584.

Crow,M. *et al.* (2019) Predictability of human differential gene expression. *Proc. Natl. Acad. Sci. USA*, **116**, 6491–6500.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Fuglede,B. and Topsoe,F. (2004) Jensen–Shannon divergence and Hilbert space embedding. In: *IEEE International Symposium on Information Theory*, pp. 30–30. IEEE, Chicago, IL, USA.

Gentleman,R. *et al.* (2005) Distance measures in DNA microarray data analysis. In: Robert,G. (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health*. Springer, New York, NY, pp. 189–208.

Jang,K.-J. *et al.* (2019) Reproducing human and cross-species drug toxicities using a Liver-Chip. *Sci. Transl. Med.*, **11**, eaax5516.

Jaskowiak,P,A. et al. (2013) Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. IEEE/ACM Trans. Comput. Biol Bioinf, **37**, 145.

Jianhua,L. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, **37**, 145–151.

Jiaxing,C. *et al.* (2019) On triangular Inequalities of correlation-based distances for gene expression profiles. bioRxiv, Cold Spring Harbor Laboratory, doi:10.1101/582106.

Kasendra,M. *et al.* (2020) Duodenum Intestine-Chip for preclinical drug assessment in a human relevant system. *eLife*, **9**, e50135.

Keen,J.C. and Moore,H.M. (2015) The Genotype-Tissue Expression (GTEx) project: linking clinical data with molecular analysis to advance personalized medicine. *J. Pers. Med.*, **5**, 22–29.

Kim,D.S. *et al.* (2006) Classification and natural history of the idiopathic interstitial pneumonias. *Proc. Am. Thorac. Soc.*, **3**, 285–292.

Kotliar,D. *et al.* (2019) Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife*, **8**, e43803.

Lederer,D.J. and Martinez,F.J. (2018) Idiopathic pulmonary fibrosis. *N. Engl. J. Med.*, **378**, 1811–1823.

Li,W. V. *et al.*. (2017) TROM: A Testing-Based Method for Finding Transcriptomic Similarity of Biological Samples. Statistics in Biosciences, **9**, 105–136. 10.1007/s12561-016-9163-y

Liu,H. *et al.* (2010) Stability approach to regularization selection for high dimensional graphical models. *Adv. Neural Inf. Process. Syst*.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

McDonough,J.E. *et al.* (2019) Transcriptional regulatory model of fibrosis progression in the human lung. *JCI Insight*, **4**, e131597. https://doi.org/10.1172/jci.insight.131597.

Mele,M. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

Merienne,N. *et al.* (2019) Cell-type-specific gene expression profiling in adult mouse brain reveals normal and disease-state signatures. *Cell. Rep.*, **26**, 2477–2493.e9.

Mohammed,A. *et al.* (2019) Differential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients. *Sci. Rep.*, **9**, 11270.

Nagwani,N.K. (2015) *A Comment on "A Similarity Measure for Text Classification and Clustering". IEEE Trans. Knowl. Data Eng.*, **26**, 1575–1590.

Pereira,V. *et al.* (2009) A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics*, **183**, 1597–1600.

Phipson,B. *et al.* (2019) Evaluation of variability in human kidney organoids. *Nat. Methods*, **16**, 79–87.

Pradhan,N. *et al.* (2015) A review on text similarity technique used in IR and its application. *Int. J. Comput. Appl.*, **120**, 29–34.

Skinnider,M.A. *et al.* (2019) Evaluating measures of association for single-cell transcriptomics. *Nat. Methods*, **16**, 381–386.

Smedley,D. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories.. *Nucleic Acids Res.*, **43**, W589–W598., Volume Issue Pages https://doi.org/10.1093/nar/gkv350.

Sollner,J. *et al.* (2017) An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Sci. Data*, **4**, 170185.

Souto,M. *et al.* (2008) Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **9**, 497. 10.1186/1471-2105-9-497.

Sudmant,P.H. *et al.* (2015) Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.*, **16**, 287.

Suntsova,M. *et al.* (2019) Atlas of RNA sequencing profiles for normal human tissues. *Sci. Data*, **6**, 36.

Thul,P.J. *et al.* (2017) A subcellular map of the human proteome. *Science*, **356**, eaal3321.

Uhlen,M. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**. doi:10.1126/science.1260419.

Uhlen,M. *et al.* (2017) A pathology atlas of the human cancer transcriptome. *Science*, **357**, eaan2507.

Yu,N.Y. *et al.* (2015) Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic Acids Res.*, **43**, 6787–6798.