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

∗To whom correspondence should be addressed.

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 are of paramount importance for precision medicine and drug discovery. It will allow us to identify potential side effects and test the effectiveness of therapeutic approaches early in their design phase and will inform the development of accurate 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 and 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 in different scenarios of practical importance using several publicly available RNA-seq datasets.

Contact: dimitris.manatakis@emulatebio.com

Supplementary information: Supplementary data are available at bioRxiv.

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 disease states (Arbar et al. (2016); Crow et al. (2019); Mohammed et al. (2019); McDonough 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.) (Uhlen et al. (2015, 2017); Yu et al. (2015); Keen et al. (2015); Mele et al. (2015); Suntsova et al. (2019); Sollner et al. (2017)) using RNA-seq, 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. In particular, the HPA provides for every tissue type a set of genes exhibiting significantly elevated expressions compared to the other organ tissue types. It is widely accepted that these gene sets form a "transcriptomic signature" of the specific organ, and their expression patterns characterize the tissue’s underlying biological processes.

Recent advancements in bioengineering and biotechnology have enabled the development of cell-cultured based organ models that recapitulate critical functions of human organs (e.g., liver, intestine, brain, etc.) (Jang et al. (2019); Kasendra et al. (2019)). The emerge of such ex-vivo human organ models generated in turn the need for new mathematical tools for determining their "similarity" to the actual human organ they represent. Such tools will not only help us understand the model capabilities and limitations but also reveal aspects we can improve in their design to optimize their physiological relevance and increase their value for precision medicine. Next-generation sequencing data (e.g., RNA-seq) is already utilized to determine the distance between organ tissue samples (Mele et al. (2015); Suntsova et al. (2019); Sollner et al. (2017); Sudmant et al. (2015)) in conjunction with classical mathematical tools such as the Euclidean distance, or dimensionality reduction techniques (e.g., Principal Component Analysis, Linear Discriminant Analysis, Uniform Manifold Approximation and Projection, etc.). However, all these methods exhibit significant limitations due to sensitivity to noisy measurements...
and outliers, inability to capture non-linear relations, etc. (Pereira et al. 2009; Li et al. 2016)). In this work, we introduce a new distance, called Transcriptomic Signature Distance (TSD), that was inspired from the field of information retrieval, addresses the problem of tissue sample comparisons in the well-established framework of information theory, and circumvents the above-mentioned weaknesses of more classical approaches.

Text similarity is a well-studied problem in information retrieval (Pradhan et al. 2015; Nagwani 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 is applied to construct words (i.e., the genes). Based on this analogy, the transcriptome 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 genes that are considered transcriptomic signatures of the compared organ tissues (Yu et al. 2015; Lonsdale et al. 2013)). These are sizeable gene sets that can adequately characterize the “identity” of organ tissues. Using them as the basis in comparing tissue samples allows us to significantly reduce the effects of sequencing “background noise” and donor-to-donor variability in the analysis. Moreover, making use of information theory and advanced statistical methods, TSD can capture the distance/similarity between any two arbitrary organ tissue samples, or between two tissue samples that are known to belong to two different classes (sample groups, e.g., different organs, organ models etc.). In the latter case, TSD considers, in a principled manner, the intra-class variabilities and incorporates them into the distance estimation. The proposed distance space is determined using the expressions probability distributions and is related to Shannon’s entropy (see equation (2)), the divergence (distance) between the probability distributions of the reference tissue (i.e., read the stories) of tissues. These transcriptome (i.e., read the stories) of tissues. The proposed TSD 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), used to estimate the distance between two samples knowing that they belong to two different classes (tissue sets) and taking into account the intra class variabilities.

2.2 The Transcriptomic Signature Distance

The proposed TSD measures the transcriptomic distance between a tissue sample from a “reference” tissue sample 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 the two distances and their limitations when used in isolation that justifies their combined use in assessing tissue distances.

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 et al. 2004)). JSD calculates the divergence (distance) between the probability distributions \( p^A \) and \( p^B \) using the following equation:

\[
JSD(p^A, p^B) = H\left(\frac{1}{2} p^A + \frac{1}{2} p^B\right) - \frac{1}{2} H(p^A) - \frac{1}{2} H(p^B)
\]

where \( H(p^A) = \sum_{q=1}^{M} p^A_q \log_2 p^A_q \) is the Shannon’s entropy function (Jain et al. 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 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 take into account 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 expression profiles \((N = 20 \ \text{w.l.o.g.})\) of two organ tissues \(t\) and \(l\). Let’s assume that \([g^A_1, \ldots, g^A_M]\) are the
Transcriptomic Signature Distance

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

2.2.3 Transcriptomic Signature Distance

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

\[
TSD(t, \hat{t}) = \frac{1}{2} SR-JSD(\hat{p}_{\hat{A}}, \hat{p}_{A}) + \frac{1}{2} RCC(\hat{p}_{\hat{A}}, \hat{p}_{A}),
\]

where \( RCC \) is the Rectified Linear Unit activation function defined as:

\[
ReLU(x) = x, \quad \text{when } x > 0 \quad \text{and zero otherwise.}
\]

For the calculation of the \( RCC \) we assume that if two ranking vectors have \( RCC < 0 \) (i.e. are anti-correlated) then their Rankings Correlation Distance 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 Transcriptomic Signature Distance 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 \( t \) and \( \hat{t} \) 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 different tissue sets (e.g., ex-vivo organ model samples vs. human organ samples). We propose a modified version...
of the TSD, which takes into account the gene expression variability of the samples within the corresponding tissue sets and provides statistically robust and accurate estimations of their transcriptomic signature distances.

Let’s assume that we have two sets of tissue samples $\mathcal{T} = \{t_1, t_2, \ldots, t_V\}$ and $\mathcal{F} = \{f_1, f_2, \ldots, f_T\}$. Similarly to the notation used in Section 2.1, $\mathcal{T}$ corresponds to the set of tissue samples that we want to compare to the reference set $\mathcal{F}$. For each tissue sample $t_i \in \mathcal{T}$, where $i = 1, 2, \ldots, V$, and $f_j \in \mathcal{F}$, where $j = 1, 2, \ldots, U$, we form: (i) the expression profile vectors $\hat{\mathbf{P}}_i$ and $\hat{\mathbf{P}}_j$; (ii) the Atlas signature vectors $\hat{\mathbf{A}}^\mathcal{T}$ and $\hat{\mathbf{A}}^\mathcal{F}$; (iii) the discrete probability distributions of the Atlas genes $\hat{\mathbf{A}}^\mathcal{T} = [P_{A1}^\mathcal{T}, P_{A2}^\mathcal{T}, \ldots, P_{AM}^\mathcal{T}]$ and $\hat{\mathbf{A}}^\mathcal{F} = [P_{A1}^\mathcal{F}, P_{A2}^\mathcal{F}, \ldots, P_{AM}^\mathcal{F}]$; and (iv) the matrices $[\Pi, \varPi]$ that summarize the Atlas gene probability distributions of the corresponding samples.

\[
\begin{align*}
\mathbf{n} &= \begin{bmatrix}
\hat{P}_{A1}^\mathcal{T} & \hat{P}_{A2}^\mathcal{T} & \cdots & \hat{P}_{AM}^\mathcal{T} \\
\hat{P}_{A1}^\mathcal{F} & \hat{P}_{A2}^\mathcal{F} & \cdots & \hat{P}_{AM}^\mathcal{F}
\end{bmatrix} \\
\mathbf{r} &= \begin{bmatrix}
\hat{P}_{A1}^\mathcal{T} & \hat{P}_{A2}^\mathcal{T} & \cdots & \hat{P}_{AM}^\mathcal{T} \\
\hat{P}_{A1}^\mathcal{F} & \hat{P}_{A2}^\mathcal{F} & \cdots & \hat{P}_{AM}^\mathcal{F}
\end{bmatrix}
\end{align*}
\]

2.3.1 The weighted Jensen-Shannon Divergence

In $\Pi$ and $\varPi$, we assume that the probabilities of appearance of each Atlas gene (e.g., $k^{\text{th}}$ gene) across tissue samples (i.e., rows of matrices), were generated by a normal distribution ($\mu_k$, $\sigma_k^2$) with parameters:

\[
\begin{align*}
\mu_k &= \frac{1}{V} \sum_{i=1}^{V} \hat{P}_{Ai}^\mathcal{T} \\
\sigma_k^2 &= \frac{1}{V} \sum_{i=1}^{V} (\hat{P}_{Ai}^\mathcal{T} - \mu_k)^2 \\
\end{align*}
\]

\[
\begin{align*}
\mu_k &= \frac{1}{U} \sum_{j=1}^{U} \hat{P}_{Aj}^\mathcal{F} \\
\sigma_k^2 &= \frac{1}{U} \sum_{j=1}^{U} (\hat{P}_{Aj}^\mathcal{F} - \mu_k)^2 \\
\end{align*}
\]

Using this assumption, for each tissue $\{t_i, t_j\}$, the likelihood of appearance of the $k^{\text{th}}$ Atlas gene can be calculated as:

\[
\begin{align*}
F(P_{A1}^\mathcal{T}; \hat{\mu}_k, \hat{\sigma}_k^2) &= \frac{1}{\sqrt{2\pi} \sigma_k} \exp\left(-\frac{(P_{A1}^\mathcal{T} - \hat{\mu}_k)^2}{2\sigma_k^2}\right) \\
F(P_{A1}^\mathcal{F}; \hat{\mu}_k, \hat{\sigma}_k^2) &= \frac{1}{\sqrt{2\pi} \sigma_k} \exp\left(-\frac{(P_{A1}^\mathcal{F} - \hat{\mu}_k)^2}{2\sigma_k^2}\right)
\end{align*}
\]

The larger the $F(P_{A1}^\mathcal{T}; \hat{\mu}_k, \hat{\sigma}_k^2)$ ($F(P_{A1}^\mathcal{F}; \hat{\mu}_k, \hat{\sigma}_k^2)$) the more “confident” we are about the likelihood of appearance of the $k^{\text{th}}$ Atlas gene in the set of samples $\mathcal{T}$ ($\mathcal{F}$). We quantify our “confidence” as:

\[
\begin{align*}
\phi_{A1}^\mathcal{T} &= \log_{10}(F(P_{A1}^\mathcal{T}; \hat{\mu}_k, \hat{\sigma}_k^2) + 1) \\
\phi_{A1}^\mathcal{F} &= \log_{10}(F(P_{A1}^\mathcal{F}; \hat{\mu}_k, \hat{\sigma}_k^2) + 1)
\end{align*}
\]

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 TSD we utilize a weighted version of the Shannon’s
The weighted Rankings Correlation Coefficient

The second term of the ISD (see equation (4)) is the rankings correlation distance (RCD) between the reference Atlas genes ranking vectors \( \hat{P}_A \) of the compared tissues \( t, t \). For the case where we have sets of tissue samples (i.e. \( T \) and \( T \)) we use weighted Rankings Correlation Coefficient (wRCC) which can be calculated as:

\[
\text{wRCC}(\hat{P}_A, \hat{P}_A) = \frac{\sum_{q=1}^{Q} \left[ |\beta_q| \left( \frac{r_{AQ} - r_{AQ}}{|r_{AQ} - r_{AQ}|} \right) \right]}{\sum_{q=1}^{Q} \left[ \sum_{q=1}^{Q} |\beta_q| \left( \frac{r_{AQ} - r_{AQ}}{|r_{AQ} - r_{AQ}|} \right) \right]}
\]

where \( r_{AQ} \) is the correlation coefficient between the ranked gene lists of tissues \( t \) and \( t \) for gene \( q \).

2.3.2 The weighted Rankings Correlation Coefficient

The weighted Rankings Correlation Coefficient (wRCC) is defined as:

\[
\text{wRCC}(\hat{P}_A, \hat{P}_A) = \frac{\sum_{q=1}^{Q} \left[ |\beta_q| \left( \frac{r_{AQ} - r_{AQ}}{|r_{AQ} - r_{AQ}|} \right) \right]}{\sum_{q=1}^{Q} \left[ \sum_{q=1}^{Q} |\beta_q| \left( \frac{r_{AQ} - r_{AQ}}{|r_{AQ} - r_{AQ}|} \right) \right]}
\]

where \( r_{AQ} \) is the correlation coefficient between the ranked gene lists of tissues \( t \) and \( t \) for gene \( q \).

The wRCC can be interpreted as a weighted version of the standard correlation coefficient, where the weights \( |\beta_q| \) represent the importance of each gene in the ranking.

To calculate the wRCC, we first compute the correlation coefficients \( r_{AQ} \) for each gene \( q \) and then weight these coefficients by the absolute values of the elements of \( \beta_q \). Finally, we normalize these weighted coefficients to obtain the wRCC.

The wRCC is a measure of how well the two ranking vectors \( \hat{P}_A \) and \( \hat{P}_A \) agree with each other, taking into account the importance of each gene in the ranking.
Each row of matrices $\mathbf{R}$ and $\mathbf{\dot{R}}$ 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 $\mathbf{R}$ and $\mathbf{\dot{R}}$, we assume that the rankings of each Atlas gene (e.g. $k^{th}$ gene) across tissues, were generated by a normal distribution (e.g. $N(\gamma_k, \delta_k^2)$) with parameters:

$$\gamma_k = \frac{1}{V} \sum_{j=1}^{V} \gamma_{jk}$$

$$\delta_k = \sqrt{\frac{1}{V} \sum_{j=1}^{V} \delta_{jk}^2}$$

Using this assumption, the likelihood about the rankings of the $k^{th}$ Atlas gene can be calculated as:

$$\Xi(r_{jk}; \gamma_k, \delta_k^2) = \frac{1}{\delta_k \sqrt{2\pi}} e^{-\frac{(r_{jk} - \gamma_k)^2}{2\delta_k^2}}$$

The larger the $\Xi(r_{jk}; \gamma_k, \delta_k^2)$ the more “confident” we are about the ranking $r_{jk}$ ($\delta_{jk}^2$) of the $k^{th}$ Atlas gene. We quantify our “confidence” as:

$$\tau_{jk} = \log_{10}(\Xi(r_{jk}; \gamma_k, \delta_k^2) + 1)$$

To avoid negative “confidence” values we added “1” before taking the logarithm of the likelihoods. Using the $\tau_{jk}$ and $\gamma_{jk}$ we calculate the importance weight of the $k^{th}$ Atlas gene as:

$$\beta_k = \gamma_k + \tau_{jk}.$$ 

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

### 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 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 hours after death. The Human Protein Atlas (HPA) includes information for only 13 out of the 20 organs in the database, and these were used in our analysis. As we can see in 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 13 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 estimate 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).

If we examine carefully corresponding rows in Figures 3a and 3b we see that in many cases there is no correlation between the corresponding $SR.JSD$ and $wRCC$ values, indicating that these two pieces of information capture different aspects of transcriptome dissimilarities. To better illustrate this fact, Figures 4a–4b and 4c–4d depict the distances of organs from Lung and Kidney (references) respectively in a 2D-space where $SR.JSD$ and $wRCC$ 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 ($SR.JSD$ and $wRCC$) 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 $wRCC$s but different $SR.JSD$s 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 $SR.JSD$s but different $wRCC$s coordinate values. Based on these observations it is clear that the proposed new distance, $wTSD$, which combines the $SR.JSD$s and $wRCC$s information while also taking into account 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 use publicly available RNA-seq datasets characterizing two different diseases: idiopathic pulmonary fibrosis and liver cancer.
Transcriptomic Signature Distance

Fig. 3. Heatmaps of means of pairwise distances between organ tissue samples. (a) SR-wJSD, (b) wRCD, (c) wTSD. The rows correspond to the reference organs whose Atlas genes were used in the distance calculations.

Fig. 4. The inter-organ distances 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 pairwise distances (SR-wJSD and wRCD) between that organ's tissue samples and the samples of the reference organ (Lung or Kidney respectively).

3.2.1 Using wTSD with idiopathic pulmonary fibrosis dataset

Recently published research (McDonough et al. (2019)) has studied the progression mechanisms of Idiopathic Pulmonary Fibrosis (IPF), a lethal chronic lung disease which progresses the fibrosis in lungs over time, causing serious breathing difficulties. IPF affects 13 to 20 per 100K people worldwide. According to the National Institute of Health, about 30K to 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 et al. (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 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 of samples in the corresponding IPF progression group and Control group samples (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 transcriptomic signature distance from the Control class also increases. This fact demonstrates the interpretability of the proposed distance. 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 Figure 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. 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 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 employed 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 for every sample the expression information of the 936 genes, which, according to HPA, form a transcriptomic signature of the healthy human liver (see 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 between samples in the corresponding cancer group and Control samples. 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 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 7b show, for the liver and kidney datasets respectively, the pairwise distances (SR-JSD and RCD) of organ model samples (mouse, rat, organoids) from the healthy human organ tissue samples. The boxplots in Figures 7a and 7b summarize the distributions of these pairwise distances. The results indicate that the liver tissue samples

3.3 Assessing distance of human organs from different organ models

In this section, we show that the proposed transcriptomic signature distance 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 transcriptomic signature distance 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 Table S1. For mouse and rat, we obtained RNA-seq data from the database developed by Solntsev et al. (2017). For both species, the number of available samples per organ was equal to 3. Finally, we retrieved RNA-seq data for healthy human liver and kidney derived organoids from the publicly available datasets (GSE84073 and GSE99582) presented in Broutier et al. (2017) and Pipitone et al. (2017) respectively.

The number of samples of the healthy liver and kidney organoids was 6 and 3, 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)).
of mouse and rat are transcriptomically closer to the human liver than the human-derived liver organoid samples. However, for the kidney, the transcriptomic signature distances of mouse and rat are transcriptomically closer to the human liver than the organoid samples of kidney organ models and human kidney tissue samples.

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 employ (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 (Kotlar et al. (2019); Merienne et al. (2019); Thul et al. (2017)).

4 Conclusions

We presented the development of TSD, a new distance we have 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, microphysiological systems design, single-cell types comparison, etc. To facilitate its use we plan to make its R code implementation available on Github.

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 1U01TR002188. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Alhaj S, Alhajjar M, Campos-Laborio FJ et al. “Identification of expression patterns in the progression of disease stages by integration of transcriptomic data”. BMC Bioinformatics. 2016;17(Suppl 15):432.

Boutron I., Mastrogiovanni, G., Vanstien, M. et al. “Human primary liver cancer–derived organ cultures for disease modeling and drug screening.” Nat Med 23, 1424–1435 (2017). doi: 10.1038/nm.4438.

Crow M., N Limb, S Balouza, P. Pardis, and J. Gillis “Predictability of human differential gene expression”. Proceedings of the National Academy of Sciences. (2019), 116(13): 6491-6500.

Friedman J., Hastie T., Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–441.
Fuglede B, Topsoe F. "Jensen-Shannon divergence and Hilbert space embedding". IEEE International Symposium on Information Theory, 2004:30–30.

Jang K.J., Otteno M.A., Romiti J. et al. "Reproducing human and cross-species drug toxicities using a Liver-Chip". Science Translational Medicine 06 Nov 2019. Vol. 11, Issue 517, eaax516. DOI: 10.1126/scitranslmed.aax516.

Jihanlou H. 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1):145–151.

Jiaxing C. and Kaow N.Y., Lu L., Yiqi J. et al. "On triangular Inequalities of correlation-based distances for gene expression profiles". bioRxiv, Cold Spring Harbor Laboratory, 2019. 10.10158/2106.

Kasendra M., Raymond L., Jianyi Y. et al. "Duodenum Intestine-Chip for pre-clinical drug assessment in a human relevant system". eLife 2020:e50135. DOI: 10.7554/eLife.50135.

Kern J.C., Moore H.M. "The Genotype-Tissue Expression (GTEX) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine". J Pers Med, 2015, 5: 22-29.

Kerr D.S., Collard H.R., King T.E. "Classification and natural history of the idiopathic interstitial pneumonias". Proc Am Thorac Soc, 2006:3(4):285–292.

Kotliar, D. et al. "Idiopathic pulmonary fibrosis". N Engl J Med, 2015, 5: 22-29.

Kuipers J., Leparc G., Hildebrandt T. et al., "A subcellular map of the human proteome". Science 26 May 2017, Vol. 357, Issue W1, Pages W589–W598, doi:10.1126/science.aan2507.

Li W.V., Chen Y., Li J.J. "Stability approach to regularization selection for high dimensional graphical models. Advances in Neural Information Processing Systems, 2010.

Lonsdale J., Thomas J., Salvatore M. et al. "The Genotype-Tissue Expression (GTEX) project". Nat. Genet. 2013 45 S8058.

McDonough J.E., Ahangari F., Li Q. et al. "Transcriptional regulatory model of fibrosis progression in the human lung". PLoS ONE 2019;14(2):e0213597. doi:10.1177/2151161X18334290.

Mehr, M. et al. "Human genomics. The human transcriptome across tissues and individuals". Science 348, 666-663 (2015).

Metzner, N. et al. Cell-Type-Specific Gene Expression Profiling in Adult Mouse Brain Reveals Normal and Disease-State Signatures. Cell Rep. 26, 2477-2493.e9 (2019).

Mohammed, A., Cui, Y., Mao, V.R. et al. “Differential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients”. Sci Rep 9, 11270 (2019). doi:10.1038/s41598-019-47703-6.

Nagwani N.K., “A Comment on “A Similarity Measure for Text Classification and Clustering””, in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 9, pp. 2589-2590, 1 Sept. 2015. doi: 10.1109/TKDE.2015.2454616.

Nagwani N.K., “A Comment on “A Similarity Measure for Text Classification and Clustering””, in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 9, pp. 2589-2590, 1 Sept. 2015. doi: 10.1109/TKDE.2015.2454616.

Nagwani N.K., “A Comment on “A Similarity Measure for Text Classification and Clustering””, in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 9, pp. 2589-2590, 1 Sept. 2015. doi: 10.1109/TKDE.2015.2454616.

Pradhan N., Gyanchandani M., Wadhvani R., “A Problem With the Correlation Coefficient as a Measure of Gene Expression Divergence” GENETICS December 1, 2009 vol. 183 no. 4 1597-1600. https://doi.org/10.1534/genetics.109.101247.

Phipson B., Efron B. "Evaluation of variability in human kidney organoids." Nat Methods 11, Issue 517, eaax5516 DOI: 10.1126/scitranslmed.aax5516.

Rui H., Roeder K., and Wasserman L. "Evaluation of variability in human kidney organoids." Nat Methods 11, Issue 517, eaax5516 DOI: 10.1126/scitranslmed.aax5516.

Smedley D., Haider S., Durinck S., "The BioMart community portal: an innovative alternative to large, centralized data repositories." Nucleic Acids Research, Volume 43, Issue W1, 1 July 2015, Pages W366-W369, https://doi.org/10.1093/nar/gkv350.

Sollner J., Leparc G., Hildebrandt T. et al. “An RNA-Seq atlas of gene expression in mouse and rat normal tissues.” Scientific Data 4, 170385 (2017) doi:10.1038/sdata.2017.185.

Sudmant P.H., Alexis M.S., Burge, C.B. “Meta-analysis of RNA-seq expression data across species, tissues and studies.” Genome Biol 16, 287 (2015) doi:10.1186/s13059-015-0853-4.

Sunitova, M., Gafullin, N., Allina, D. et al. “Atlas of RNA sequencing profiles for normal human tissues.” Scientific Data 6, 36 (2019), doi:10.1038/s41597-019-0043-4.

Thul P.J., Akerson L., Wiking M., et al., “A subcellular map of the human proteome.” Science 26 May 2017.Vol. 357, Issue 356, Issue 6340, eaal3321 DOI: 10.1126/science.aal3321.

Thul P.J., Akerson L., Wiking M., et al., “A subcellular map of the human proteome.” Science 26 May 2017.Vol. 357, Issue 356, Issue 6340, eaal3321 DOI: 10.1126/science.aal3321.

Uhlen M. et al. “Tissue-based map of the human proteome.” Science 347, Issue 6220, DOI: 10.1126/science.1260419.

Uhlen M. et al. “A pathology atlas of the human cancer transcriptome.” Science 18 Aug 2017, Vol. 357, Issue 6352, DOI: 10.1126/science.aan2507.

Yu N.Y., Hallstrom B.M., Fagerberg L. et al. "A pathology atlas of the human cancer transcriptome". Nat Commun 7, 10986 (2016) doi:10.1038/srep10986.