CoSIA: an R Bioconductor package for CrOss Species Investigation and Analysis

**Authors:**

<table>
<thead>
<tr>
<th>Author Name</th>
<th>Affiliation</th>
<th>CRedIT Roles*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anisha Haldar*</td>
<td>1. CDIB</td>
<td>Methodology, Software, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review &amp; Editing, Visualization</td>
</tr>
<tr>
<td>Vishal H. Oza*</td>
<td>1. CDIB</td>
<td>Conceptualization, Methodology, Software, Formal Analysis, Supervision, Project administration, Writing - Original Draft, Writing - Review &amp; Editing</td>
</tr>
<tr>
<td>Nathaniel S. DeVoss</td>
<td>1. CDIB</td>
<td>Software, Data Curation, Writing - Review &amp; Editing, Visualization</td>
</tr>
<tr>
<td>Amanda D. Clark</td>
<td>1. CDIB</td>
<td>Validation, Writing - Review &amp; Editing, Supervision, Data Curation, Software</td>
</tr>
<tr>
<td>Brittany N. Lasseigne</td>
<td>1. CDIB</td>
<td>Conceptualization, Methodology, Resources, Writing - Review &amp; Editing, Supervision, Project administration, Funding acquisition</td>
</tr>
</tbody>
</table>

* Co-first authors
Abstract:

Summary:

High throughput sequencing technologies have enabled cross-species comparative transcriptomic studies; however, there are numerous challenges for these studies due to biological and technical factors. We developed CoSIA (Cross-Species Investigation and Analysis), an Bioconductor R package and Shiny app that provides an alternative framework for cross-species transcriptomic comparison of non-diseased wild-type RNA sequencing gene expression data across tissues and species through visualization of variability, diversity, and specificity metrics.

Availability and Implementation: https://github.com/lasseignelab/CoSIA

Contact: Brittany Lasseigne (bnp0001@uab.edu)

Supplementary information: See Supplementary Files

Figure 1.
(A) The CoSIA R package workflow. (B) The coefficient of variation cross-species comparison analysis depicts a lollipop plot of example data used to show median-based CV gene comparison across species. This example highlights genes with high variability (gene5), low variability (gene4, gene3, and gene1), and species-specific variability (gene2) in expression across tissues. (C) The diversity and specificity cross-species comparison analysis depicts scatter plots of example data (Supplemental file: example_data.csv) showing diversity and specificity metric for tissues and genes. This example highlights the differences in diversity and specificity across tissues (tissue1 and tissue3 in species2), and the similarities in diversity and specificity across species (gene1, gene2, gene3, gene4 in species2). Figure was created with BioRender.com.
1. Introduction

With the advent of high-throughput sequencing technologies (Goodwin, McPherson and McCombie 2016), there has been an explosion in the generation of gene expression data across multiple species (Mutz et al. 2013). Providing an excellent opportunity to leverage this available data to better study human and model organism gene expression patterns in a biomedical context, cross-species comparative studies have elucidated disease mechanisms, evolutionary patterns, and developmental differences (LoVerso and Cui 2015). However, cross-species gene expression comparison is challenging due to biological and technical factors affecting the measurements (Conesa et al. 2016; Chung et al. 2021). Previous studies have implemented a variety of comparison methods (Zhu et al. 2014; Sudmant, Alexis and Burge 2015; Söllner et al. 2017; Panahi et al. 2019; Wang et al. 2020; Bastian et al. 2021; García de la Torre et al. 2021; Liu et al. 2023) that involve either taking into account the evolutionary relationships between the species or rigorous statistical assumptions to account for species-level differences in gene expression (Fisher and Others 1948; Stuart et al. 2003; Hu, Greenwood and Beyene 2006; Lu, Rosenfeld and Bar-Joseph 2006; Lu et al. 2007, 2010; Campain and Yang 2010; Le, Oltvai and Bar-Joseph 2010; Tseng, Ghosh and Feingold 2012; Kristiansson et al. 2013). We developed CoSIA (Cross-Species Investigation and Analysis), an R package and associated Shiny app, which provides an alternative framework for cross-species RNA expression visualization and comparison across tissues and species using variability metrics. CoSIA allows users to calculate variability, diversity, and specificity metrics across *Homo sapiens* and five species commonly used in biomedical research *Mus musculus, Rattus norvegicus, Danio rerio, Drosophila melanogaster*, and *Caenorhabditis elegans.*
2. Implementation

CoSIA (Haldar et al. 2023), is a Bioconductor package accessible as part of the Bioconductor 3.17 release. CoSIA allows for relative cross-species comparison of non-diseased wild-type RNA sequencing gene expression data across tissues and species. Specifically, CoSIA implements methods for mapping gene identifiers and orthologs, visualizing gene expression by tissue and species, and comparing cross-species gene expression metrics, as shown in Fig. 1A.

2.1 Mapping between Gene Identifiers and Orthologs (getConversion)

CoSIA provides the user with a streamlined method for converting between different gene identifiers and cross-species ortholog mapping. The current gene identifiers supported are Ensembl IDs, Entrez IDs, and Gene Symbols. These conversions are performed with the BiomaRt (Durinck et al. 2005) and AnnotationDBI (Pagès et al. 2022) packages, while ortholog gene mapping is performed using the NCBI Homologene Database (Sayers et al. 2022) and the NCBI Eukaryotic Genome Annotation Pipeline Database (Thibaud-Nissen et al. 2016).

2.2 Visualizing Gene Expression Data (getGEx)

CoSIA uses the Bioconductor Data Package CoSIAdata (EH7858, EH7859, EH7860, EH7861, EH7862, EH7863) which is hosted on ExperimentHub, which contains variance stabilized and transformed gene expression values using the Variance Stabilizing Transformation (VST) method implemented in the DESeq2 (Love, Huber and Anders 2014) package of non-diseased wild-type RNA-seq read counts we retrieved from the BgeeDB (Komljenovic et al. 2018). These values are parsed (depending on the user’s choice of genes, tissues, and species) into a data frame that can be visualized as an interactive violin plot through the plotSpeciesGEx and plotTissueGEx plotting functions.
2.3 Calculating and Visualizing Gene Expression Metric Data (\textit{getGExMetrics})

Comparing transcriptional profiles across species has been challenging because of differences in gene expression patterns and batch effects. Previous attempts at directly comparing expression between species using various normalization techniques (Dunn \textit{et al.} 2018) have been shown to be affected by annotation depth and quality (Oziolor, Arat and Martin 2021). Other studies (Breschi \textit{et al.} 2016) have shown that each gene has a specific pattern, with some genes showing higher variation between organs within the same species compared to variation between species and vice versa. Direct comparison methods do not account for these aspects. To overcome these challenges, we implemented variability (Coefficient of Variation), diversity, and specificity (Shannon entropy-based) metrics that allow for relative comparison of gene expression patterns between species.

\textbf{Coefficient of Variation (CV):}

The CV in CoSIA is calculated as the standard deviation over the median using VST values. CoSIA provides two approaches for calculating the CV. ‘CV Tissue’ calculates the CV of a set of user-supplied genes across the specified tissues, while ‘CV Species’ calculates the CV of user-supplied genes across the specified species. The calculated CVs are returned as a data frame and visualized as lollipop plots using the \textit{plotCVGEx} plotting function as shown in Fig. 1B.

\textbf{Diversity and Specificity:}

To calculate the diversity and specificity metrics, median VST is determined for each gene in a specific tissue in a given species. These median VST values are rescaled using min-max scaling, which preserves the distribution of values but rescales the values between 0 and 1. These values are used to calculate the diversity and specificity metrics as described in (Martínez and Reyes-Valdés 2008). Briefly, diversity and specificity refer to the relative frequency of the
expression of a gene compared to other genes across tissues (Jones et al. 2023) reviewed in (Jones et al. 2023). Thus, a highly diverse gene will have similar expression across tissues but low specificity. A highly specific gene will have higher expression in one tissue compared to other tissues. CoSIA allows for four different calculations of diversity and specificity in a species: i) ‘DS Gene’ compares user-specified genes across user-specified tissues, ii) ‘DS Gene all’ compares user-specified genes across all tissues in a species, iii) ‘DS Tissue’ compares user-specified tissues across user-specified genes, and iv) ‘DS Tissue all’ compares user-specified tissues across all genes in a species. The metrics can either be exported as a data frame or visualized using diversity/specificity scatter plots using the plotDSGEx plotting function as shown in Fig. 1C.

2.4 Shiny App

We have also implemented the CoSIA package as a Shiny app (DeVoss et al. 2023) hosted at (https://lasseignelab.shinyapps.io/CoSIA/) which provides a graphical user interface (GUI) with similar functionalities as the package.

3. Conclusion:

Direct comparison of gene expression between species is complex as it is confounded by differing levels of expression and function of orthologous genes across species. Here, we provide the CoSIA package and Shiny app to facilitate the relative comparison of gene expression by summary metrics (i.e., coefficient of variation, diversity, and specificity) in six species. By leveraging the Bgee database of species-specific RNA-Seq expression data, we provide tools for the robust comparison of gene expression values across both species and tissues. We believe CoSIA will also aid biomedical researchers in selecting optimal model organisms for a given gene in a tissue of interest.
Funding:

This work was supported in part by the UAB Lasseigne Lab Start-Up funds (BNL, AH, ND, ADC and VHO), the UAB Pilot Center for Precision Animal Modeling (C-PAM) (1U54OD030167) (BNL and VHO), UAB Pilot Center for Precision Animal Modeling (C-PAM) - Diversity Supplement (3U54OD030167-03S1) (ADC), and Mentored Experiences in Research, Instruction, and Teaching (MERIT) Program (K12 GM088010) (ADC).

Acknowledgements:

The authors would like to thank the members of the Lassigne lab for their support and feedback, in particular, Elizabeth J. Wilk, Jordan Whitlock, and Timothy C. Howton.

Supplemental Files:

CoSIA_Supp.zip contains the following three files: (1) README.md - Note with description of files in CoSIA_Supp.zip (2) CoSIA_AppNote_Supp_Table_1.pdf - Table of Package Dependencies for CoSIA, (3) example_data.csv - Median Variance Stabilized Transformation of RNA-seq Read Counts used to render Figure 1b and 1c
References:


Fisher RA, Others. 224A: Answer to Question 14 on Combining independent tests of significance. 1948.


Haldar A, Oza VH, Clark AD et al. CoSIA. 2023, DOI: 10.18129/B9.BIOC.COSIA.


