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

Authors:

Anisha Haldar[1]*, Vishal H. Oza[1]*, Nathaniel S. DeVoss[1], Amanda D. Clark[1], Brittany N. Lasseigne[1]

*****Co-first

1. The Department of Cell, Developmental and Integrative Biology, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, USA

Corresponding Author: Brittany N. Lasseigne (bnp0001@uab.edu)

Author Credit:

Anisha Haldar: Methodology, Software, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

Vishal H. Oza: Conceptualization, Methodology, Software, Formal Analysis, Supervision, Project administration, Writing - Original Draft, Writing - Review & Editing

Nathaniel S. DeVoss: Software, Data Curation, Writing - Review & Editing, Visualization

Amanda D. Clark: Validation, Writing - Review & Editing, Supervision, Data Curation, Software

Brittany N. Lasseigne: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Abstract:

Summary:

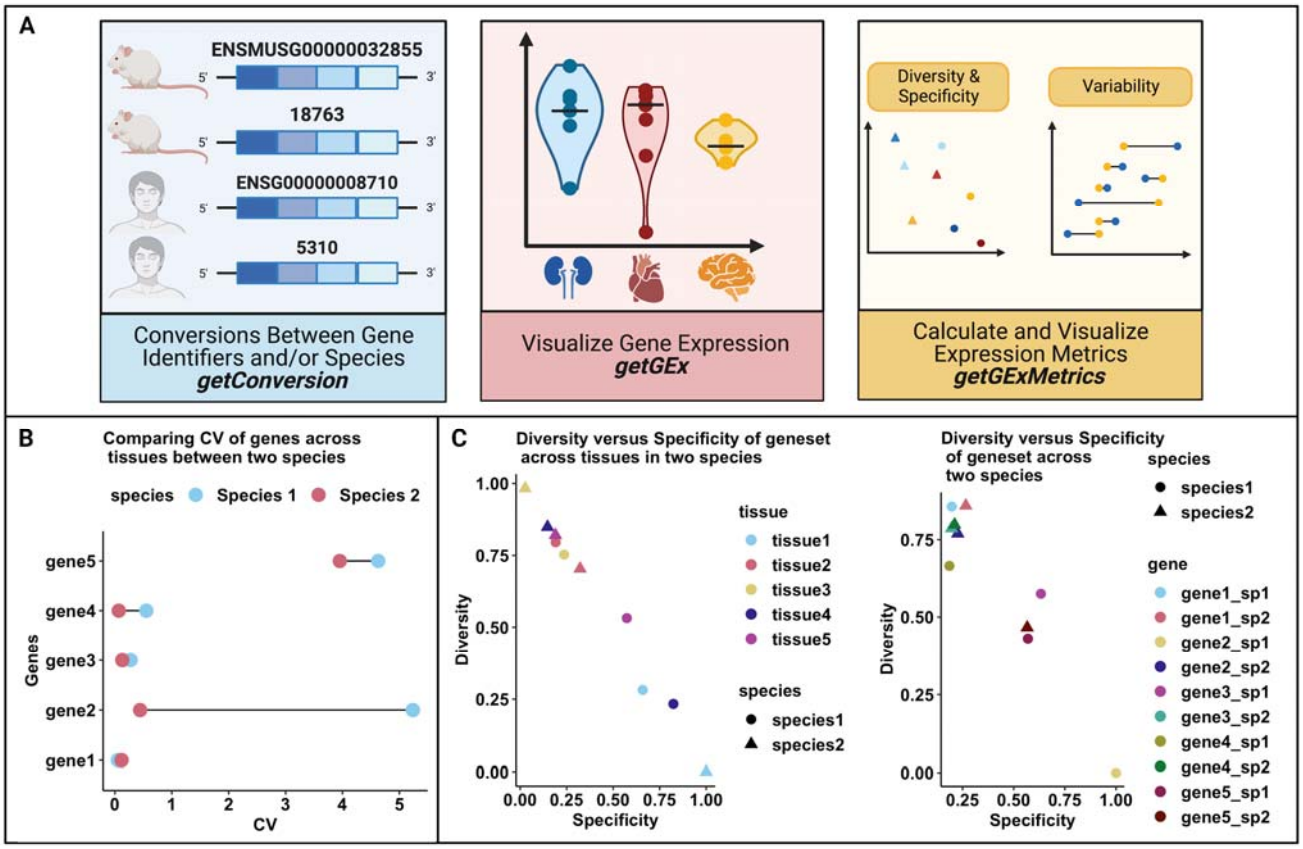
High throughput sequencing technologies have enabled cross-species comparative transcriptomic studies; however, there are numerous challenges for these studies due to biological and technical factors. We developed CoSIA (Cross-Species Investigation and Analysis), an Bioconductor R package and Shiny app that provides an alternative framework for cross-species transcriptomic comparison of non-diseased wild-type RNA sequencing gene expression data from Bgee across tissues and species (human, mouse, rat, zebrafish, fly, and nematode) 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: Table 2) 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) hosted on ExperimentHub, which contains variance stabilized Transcript per million (TPM) gene expression values of non-diseased wild-type RNA-seq read counts we retrieved using BgeeDB [package v2.26.0; database v15.0] (Komljenovic *et al.* 2018). The variance stabilization is done using the Variance Stabilizing Transformation (VST) method implemented in the DESeq2 (Love, Huber and Anders 2014) package. 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 (*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 *et al.* 2018) have been shown to be affected by annotation depth and quality (Oziolor, Arat and Martin 2021). Other studies (Breschi *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. To understand how these metrics work, we have simulated expression data of five genes and its orthologs across two species and five tissues (Supp. Table 2). The metrics calculated on these genes are plotted in Fig. 1B and C.

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 *plotCVGEx* plotting function. As shown in Fig. 1B, Gene 2 has high variation in expression across tissues in species 1 but not in species 2. However, Gene 1 and 3 have no variation in expression across both species, thus their expression doesn't change much across tissues in both species.

Diversity and Specificity:

To calculate the diversity and specificity metrics, we first calculate the median of the variance stabilized TPM values for each gene in a specific tissue in a given species. These median 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 in the context of gene expression across tissues is quantified using Shanon entropy. Diversity refers to the degree of heterogeneity or variability in the expression patterns of a gene across different tissues. It captures the distribution and relative frequency of expression across tissues, with higher diversity indicating a more evenly distributed expression pattern. Specificity, on the other hand, measures the level of concentration or selectivity of gene expression within a particular tissue. It assesses the extent to which a gene's expression is confined to a specific tissue, with higher specificity indicating a more restricted or specialized expression pattern within that tissue. The range of diversity and specificity is between 0 and 1, with 0 being low and 1 being high. Another important thing to note is that they are inversely related (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. The first half of Fig. 1C, shows that the geneset expression is very specific in tissue 1 in species 2

compared to species 1. On the contrary, the geneset expression is very diverse in tissue 3 in species 2 compared to species 1. In the second half of Fig 1C, we look at individual genes, where gene 2 has very specific expression in species 1 however, it is very diverse in species 2.

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_AppNote_Supplementary_Data.docx contains the following: (1) *Table 1* - Table of Package Dependencies for CoSIA, (2) *Table 2*- Median Variance Stabilized Transformation of RNA-seq Read Counts used to render Figure 1b and 1c (3) Figure S1 - CoSIA Usage Example: cross-species & -tissue investigation of genes associated with monogenic kidney diseases.

References:

Bastian FB, Roux J, Niknejad A *et al.* The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res* 2021;**49**:D831–47.

Breschi A, Djebali S, Gillis J *et al.* Gene-specific patterns of expression variation across organs and species. *Genome Biol* 2016;**17**:151.

Campain A, Yang YH. Comparison study of microarray meta-analysis methods. *BMC Bioinformatics* 2010;**11**:408.

Chung M, Bruno VM, Rasko DA *et al.* Best practices on the differential expression analysis of multi-species RNA-seq. *Genome Biol* 2021;**22**:121.

Conesa A, Madrigal P, Tarazona S *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;**17**:13.

Dunn CW, Zapata F, Munro C *et al.* Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc Natl Acad Sci U S A* 2018;**115**:E409–17.

Durinck S, Moreau Y, Kasprzyk A *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005;**21**:3439–40.

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

García de la Torre VS, Majorel-Loulergue C, Rigaill GJ *et al.* Wide cross-species RNA-Seq comparison reveals convergent molecular mechanisms involved in nickel hyperaccumulation across dicotyledons. *New Phytol* 2021;**229**:994–1006.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.

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

Hu P, Greenwood CMT, Beyene J. Statistical Methods for Meta-Analysis of Microarray Data: A Comparative Study. *Inf Syst Front* 2006;**8**:9–20.

Jones E, Haldar A, Oza V *et al.* Quantifying variability in gene expression profiles using transcriptome and isoform diversity. *Preprints* 2023, DOI: 10.20944/preprints202301.0257.v1.

Komljenovic A, Roux J, Wollbrett J *et al.* BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. F1000Res., 5. *DOI* 2018;**10**:f1000research.

Kristiansson E, Österlund T, Gunnarsson L *et al.* A novel method for cross-species gene

expression analysis. *BMC Bioinformatics* 2013;**14**:70.

Le H-S, Oltvai ZN, Bar-Joseph Z. Cross-species queries of large gene expression databases. *Bioinformatics* 2010;**26**:2416–23.

Liu J, Zhang Y, Zheng Y *et al.* PlantExp: a platform for exploration of gene expression and alternative splicing based on public plant RNA-seq samples. *Nucleic Acids Res* 2023;**51**:D1483–91.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

LoVerso PR, Cui F. A Computational Pipeline for Cross-Species Analysis of RNA-seq Data Using R and Bioconductor. *Bioinform Biol Insights* 2015;**9**:165–74.

Lu Y, Mahony S, Benos PV *et al.* Combined analysis reveals a core set of cycling genes. *Genome Biol* 2007;**8**:R146.

Lu Y, Rosenfeld R, Bar-Joseph Z. Identifying cycling genes by combining sequence homology and expression data. *Bioinformatics* 2006;**22**:e314–22.

Lu Y, Rosenfeld R, Nau GJ *et al.* Cross Species Expression Analysis of Innate Immune Response. *Journal of Computational Biology* 2010;**17**:253–68.

Martínez O, Reyes-Valdés MH. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proc Natl Acad Sci U S A* 2008;**105**:9709–14.

Mutz K-O, Heilkenbrinker A, Lönne M *et al.* Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol* 2013;**24**:22–30.

Oziolor E, Arat S, Martin M. Annotation depth confounds direct comparison of gene expression across species. *BMC Bioinformatics* 2021;**22**:499.

Pagès H, Carlson M, Falcon S *et al.* AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.58. 0 https://bioconductor. org/packages. 2022.

Panahi B, Frahadian M, Dums JT *et al.* Integration of Cross Species RNA-seq Meta-Analysis and Machine-Learning Models Identifies the Most Important Salt Stress-Responsive Pathways in Microalga Dunaliella. *Front Genet* 2019;**10**:752.

Sayers EW, Bolton EE, Brister JR *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2022;**50**:D20–6.

Söllner JF, Leparc G, Hildebrandt T *et al.* An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Sci Data* 2017;**4**:170185.

Stuart JM, Segal E, Koller D *et al.* A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003;**302**:249–55.

Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol* 2015;**16**:287.

Thibaud-Nissen F, DiCuccio M, Hlavina W *et al.* P8008 the NCBI eukaryotic genome annotation pipeline. *J Anim Sci* 2016;**94**:184–184.

Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* 2012;**40**:3785–99.

Wang N, Niger C, Li N *et al.* Cross-Species RNA-Seq Study Comparing Transcriptomes of Enriched Osteocyte Populations in the Tibia and Skull. *Front Endocrinol* 2020;**11**:581002.

Zhu Y, Li M, Sousa AMM *et al.* XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons. *BMC Genomics* 2014;**15**:343.