TissueXplorer, a user-friendly bioinformatic tool for transcriptomic data analysis

Ana Luísa de Sousa Rocha
Master in Bioinformatics and Computational Biology

Department of Computer Science
Faculty of Sciences of the University of Porto
2025

TissueXplorer, a user-friendly bioinformatic tool for transcriptomic data analysis

Ana Luísa de Sousa Rocha

Dissertation carried out as part of the Master in Bioinformatics and Computational Biology

Department of Computer Science
2025

Supervisor
Mónica Lopes Marques, Researcher, CIIMAR

Co-supervisor
João Carneiro, Researcher, CBMA/IB-S

Dedicated to Avô Amadeu

Acknowledgements

First and foremost, I would like to thank my supervisor, **Mónica Marques**, for her incredible support and commitment throughout this journey. She was truly dedicated to helping me achieve my goal of submitting my thesis in July, knowing how important it was to me. From the very beginning, she welcomed me into the CIIMAR group, introducing me to everyone and ensuring I felt integrated. She generously dedicated her time to sit with me for countless hours, patiently explaining biological concepts and thoughtfully guiding the direction of my work. We celebrated the small victories, and she stood by me through every challenge. For all of this, I am deeply grateful to her. Her encouragement, expertise, and kindness were essential in making this experience the best it could possibly be.

I would also like to thank **João Carneiro**, my co-supervisor who, even from a distance, met with us weekly and shared valuable suggestions and advice. I am also very grateful for the way he fostered a sense of community among his students organizing team-building lunches and creating a welcoming environment where we could learn from each other's experiences. His dedication, support and encouraging words were a great source of motivation for me and made a lasting difference in my academic journey.

I would also like to thank my colleagues at CIIMAR for being so welcoming and for fostering such a positive environment. A special thanks to **Inês**, **Diogo**, **André**, and **Nádia** for their wonderful company, which made each day more enjoyable and helped make my experience at CIIMAR truly memorable. I am also grateful to **André Machado** and **Raquel Ruivo** for their feedback, particularly in the early stages, which played an important role in shaping the direction of my thesis. Finally, I would like to thank **Filipe Castro** for taking me in his group at CIIMAR and for his guidance during the process of selecting a topic for my thesis.

To the friends I made during my Master's degree, **Diana**, **Cecília**, **Tomás**, **Marta**, and the other members of the "Meetings" group, thank you for relating and experiencing with me all the emotions that arise during this period of our lives. Moving forward is exciting, but it's also difficult to leave behind the routine of seeing you every week. The memories we've created over the past two years are truly precious to me. I hope this was just the beginning of our friendship and that we have many more "Meetings" ahead.

To my **parents**, **grandmother**, **brother**, and other **family** members: thank you for all the times you showed genuine interest, concern, and a willingness to help, even without being from the same professional field. Your support was essential in helping me complete this journey.

Lastly, to **Gustavo**, the one who has followed my journey more closely than anyone, right from the very beginning. On the day we met, I told you how much I wanted to get into this Master's Degree. Little did I know that not only would I get in and complete it, but that you would become such an important part of that journey. Thank you for always being there for me, for the advice you gave me, for the reassurance on the hard days, for always believing in me when I struggled to believe in myself, and for the many reminders like *"Did you push to GitHub*?". You are the person who motivates me the most to keep learning and to never give up on my goals and I am deeply grateful to have you in my life.

Resumo

O rápido avanço das tecnologias de sequenciação de RNA levou a uma explosão de dados transcriptómicos de múltiplas espécies. No entanto, as ferramentas atualmente disponíveis para análise comparativa destes dados permanecem, em grande parte, inacessíveis a investigadores sem conhecimentos de programação. Para ultrapassar esta lacuna, desenvolveu-se TissueXplorer, uma ferramenta que permite comparar dados de expressão genética entre diferentes espécies de mamíferos por tecido.

O TissueXplorer utiliza dados transcriptómicos públicos de vários mamíferos e dispõe de diagramas interativos, como diagramas de venn e mapas de calor, que facilitam a identificação de padrões de expressão genética partilhados ou específicos de cada espécie. A plataforma permite o carregamento de conjuntos de dados próprios do utilizador, permite pesquisar por genes específicos e inclui um módulo de análise de enriquecimento funcional que auxilia na interpretação biológica das listas de genes obtidas. Desenvolvido com recurso às linguagens Python, R, e ao sistema de base de dados SQLite, TissueXplorer é acessível através de qualquer navegador web e não requer que o utilizador possua conhecimentos de programação, tornando a análise comparativa de dados transcriptómicos mais intuitiva e acessível para a comunidade científica.

Palavras-chave: Transcriptoma, Expressão Genética, Genes Específicos de Tecido, Análise comparativa entre espécies, Bioinformática

Abstract

The rapid advancement of RNA sequencing technologies has led to an explosion of transcriptomic data across multiple species. However, the tools available for comparative analysis of such data remain largely inaccessible to researchers without programming expertise. To address this gap, TissueXplorer was developed, a web-based application that enables researchers to compare gene expression across multiple mammalian species within specific tissues.

TissueXplorer integrates publicly available transcriptomic data from multiple mammalian species and offers interactive visualizations such as Venn diagrams and heatmaps, to highlight shared and species-specific gene expression patterns. The platform supports user-uploaded datasets, enables gene-specific queries, and includes a built-in enrichment analysis module to aid in the biological interpretation of retrieved gene lists. Developed using Python, R, Shiny, and SQLite, the tool is accessible via a web browser and requires no programming knowledge, allowing researchers to explore, compare, and interpret gene expression data across species through an intuitive and interactive interface.

Keywords: Transcriptome, Gene Expression, Tissue-Specific Genes, Comparative Analysis, Bioinformatics

Table of Contents

List of Tables

List of Charts

List of Figures

List of Abbreviations

API	APPLICATION PROGRAMMING INTERFACES
CDNA	COMPLEMENTARY DEOXYRIBONUCLEIC ACID
CSV	COMMA SEPARATED VALUES
CSS	CASCADING STYLE SHEETS
DE	DIFFERENTIAL EXPRESSION
DNA	DEOXYRIBONUCLEIC ACID
FCUP	FACULTY OF SCIENCES OF THE UNIVERSITY OF PORTO
FPKM	FRAGMENTS PER KILOBASE PER MILLION
HTML	HYPER TEXT MARKUP LANGUAGE
HTS	HIGH-THROUGHPUT SCREENING
KEGG	KYOTO ENCYCLOPEDIA OF GENES AND GENOMES
MRNA	RIBONUCLEIC ACID MESSENGER
NCRNA	NON-CODING RNA
NGS	NEXT-GENERATION SEQUENCING
PCR	POLYMERASE CHAIN REACTION
RNA	RIBONUCLEIC ACID
RPKM	READS PER KILOBASE PER MILLION
RRNA	RIBOSOMAL RIBONUCLEIC ACID
RSEM	RNA-SEQ BY EXPECTATION-MAXIMIZATION
RT-PCR	REVERSE TRANSCRIPTION POLYMERASE CHAIN
SCRNA-SEQ	SINGLE-CELL RNA SEQUENCING
SQL	STRUCTURED QUERY LANGUAGE
TPM	TRANSCRIPTS PER MILLION
TRNA	TRANSFER RIBONUCLEIC ACID
TSV	TAB SEPARATED VALUES
UI	USER INTERFACE
UP	UNIVERSITY OF PORTO

Chapter 1 - Introduction

1.1. Transcriptomics

Transcriptomics is the branch of molecular biology that studies the transcriptome, which refers to the entire collection of RNA molecules that are actively being expressed in a particular cell type or tissue at a given stage of development or under specific physiological conditions. This includes messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and various other non-coding RNAs (1,2). While the genome represents the complete, static set of DNA in an organism, the transcriptome dynamically changes based on factors like the cell's environment and developmental stage.

Therefore, studying the transcriptome is fundamental because by quantifying and analyzing RNA transcripts, researchers can identify which genes are turned on or off, at what levels, and under which circumstances (2). Transcriptome analysis can then help identify patterns in gene expression and provide insights into its regulatory networks, ultimately contributing to advancements in disease diagnosis and clinical treatments (2-4). Moreover, advances in sequencing technologies have made transcriptomic data more accessible and comprehensive, allowing scientists to investigate not only the presence or absence of transcripts but also explore alternative splicing, non-coding RNAs, and gene regulation (5).

1.2 The Evolution of RNA Measurement Techniques

Since transcription levels differ across tissues, physiological conditions, and environmental stimuli, one of the first objectives of transcriptome analysis was to measure the varying expression levels of genes under different conditions (6).

Among the first techniques developed to measure RNA levels was the Northern blot. This method involves the separation of RNA molecules by gel electrophoresis, their transfer to a membrane, and subsequent hybridization with a labeled complementary DNA or RNA probe specific to the gene of interest (7,8). The detection of the hybridized probe is what reveals the presence and abundance of the specific RNA (8-10). However, the technique is labor-intensive, has low sensitivity, and is limited to the detection of one or a few transcripts at a time (11). As a result, its utility is constrained when analyzing complex transcriptomes.

Reverse transcription followed by polymerase chain reaction (RT-PCR) later emerged as a more sensitive approach to detect and quantify RNA. In this method, RNA is first

reverse transcribed into complementary DNA (cDNA), which is then amplified using PCR (12). This increased both the sensitivity and specificity of RNA detection compared to Northern blotting. RT-PCR became widely used and remains a standard tool for validating transcriptomic results. Nevertheless, RT-PCR is a targeted method, requiring prior knowledge of the gene of interest, and thus does not allow for global transcriptome analysis (13).

The development of DNA microarrays represented a breakthrough in understanding gene regulation across diverse biological contexts. Microarrays consist of a solid surface onto which thousands of gene sequences (that can be either oligonucleotides or cDNA fragments) are immobilized in a grid-like pattern. These DNA probes can hybridize with labelled cDNA or cRNA from the sample which will be detected by a scanner. The level of hybridization to each gene fragment is proportional to the quantity of mRNA present in the original sample, thus enabling the simultaneous measurement of expression levels for thousands of genes (14-18). However, microarray technology also has inherent limitations. It depends on the availability of prior sequence information to design the probes, it cannot reliably detect novel or low-abundant transcripts, has difficulty distinguishing between closely related sequences and can be expensive (19-21).

Despite these limitations, these early RNA measurement techniques paved the way for modern transcriptomics, with each method representing a step forward towards next-generation sequencing technologies.

1.3 Next-Generation Sequencing (NGS) or High Throughput Sequencing (HTS)

In recent decades, advances in Next-generation sequencing (NGS) technologies have revolutionized the way scientists understand molecular biology. First-generation sequencing methods, such as Sanger sequencing, produced high-quality sequence data through a chain-termination technique (22,23). However, despite its high accuracy, this method was limited to processing a single DNA fragment at a time, making it labor-intensive, time-consuming, and costly for large-scale projects.

Since then, significant technological advancements have been made, that have set the scene for the development of Next-Generation Sequencing methods used today. NGS refers to a collection of high-throughput technologies that comprise a wide range of methods distinguished by their unique protocols, that enables the rapid sequencing of entire genomes, exomes, or transcriptomes. Some of the most widely used NGS

platforms include Illumina, Ion Torrent, and PacBio, each differing in chemistry, read length, and error profile. Moreover, NGS platforms can generate vast amounts of data at a very low cost (24) since they allow for hundreds of millions of DNA molecules to be sequenced at the same time (25), resulting in the exponential growth of genome information. Despite their differences, all NGS platforms share key advantages: high speed, scalability, and the ability to generate vast amounts of sequence data in a single run. With the development of NGS technologies researchers could now explore questions that were once limited by the constraints of older technologies, such as detecting rare mutations, identifying novel transcripts, or quantifying gene expression with high resolution (26-28).

1.4 RNA-Seq

Among the most transformative applications of next-generation sequencing (NGS) is RNA sequencing (RNA-Seq). In transcriptomics specifically, NGS gave rise to RNA-Seq, which replaced older hybridization-based approaches like microarrays (29). RNA-Seq offers several advantages: covers a very broad dynamic range of the genome (meaning, it can detect and quantify both highly expressed and lowly expressed genes within the same experiment), provides an accurate measure of gene expression and can be applied to any species, even if there is no reference sequence available making it suitable for both model and non-model organisms (30).

The RNA-Seq procedure starts with the extraction and isolation of RNA from the biological sample. This RNA is then fragmented and transcribed into a library of cDNA. The resulting cDNA fragments are then ligated with specialized adaptors that will then allow them to be sequenced by high-throughput sequencing platforms (6). These adaptors can be attached to either one (single-end sequencing) or both ends (paired-end sequencing) resulting in short sequences or reads that typically range from 30 to 400 base pairs depending on the sequencing platform used.

After sequencing, millions of short reads are generated, which are then aligned to a reference genome or transcriptome using computational tools. In cases where no reference is available, *de novo* transcriptome assembly may be employed. Once aligned, read counts are quantified to determine gene expression levels, the number of reads aligned to each gene measures its level of expression (5,31,32).

In this way, the RNA-Seq framework enables a high-resolution analysis of all RNAs within a sample, with both identifying their sequences and measuring their expression levels at the same time (6,29) (Figure 1).

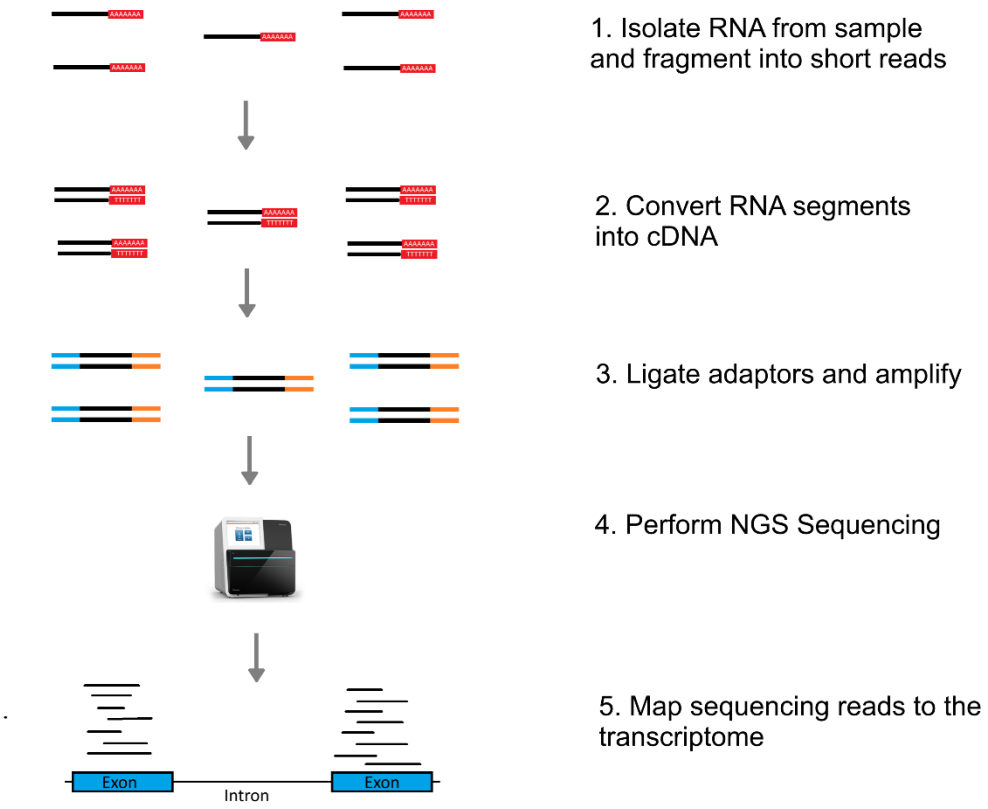

Figure 1. Schematic representation of the step-by-step process of an RNA-Seq experiment.

1.5 Types of RNA-Seq

There are different types of RNA-Seq methods, each designed for specific research goals. The main types include bulk RNA-Seq, single cell RNA-Seq (scRNA-Seq), long read RNA-Seq and non-coding RNA Analysis (33).

Bulk RNA sequencing (bulk RNA-Seq) is one of the most widely used approaches in transcriptomics for measuring gene expression across entire tissue samples or cell populations. It captures an average transcriptomic profile by extracting and sequencing RNA from a mixture of cells, making it a powerful tool for understanding the collective behavior of genes within a biological sample. Bulk RNA-Seq is also generally less expensive than other types of RNA-Seq, such as single-cell RNA-Seq (34,35), and enables differential expression analysis, detection of isoforms and alternative splicing

(35-39). However, the main limitation of bulk RNA-Seq is that it averages gene expression across all cells in the sample, meaning that transcriptomic differences between different cell types or subpopulations cannot be distinguished (35,40). Despite its limitations, bulk RNA-Seq is still a powerful, widely used and cost-effective tool for studying gene expression in large samples, offering robust and reproducible results (41).

Single-cell RNA sequencing (scRNA-Seq) is a transformative technique in transcriptomics that enables the analysis of the gene expression of individual cells. Unlike bulk RNA-Seq, scRNA-Seq extracts and sequences RNA from individual cells and allows researchers to identify distinct cell types and understand gene regulatory dynamics with an impressive level of detail (42).The scRNA-Seq process typically involves cell isolation using techniques such as microfluidics, or droplet-based methods followed by lysis, reverse transcription, amplification, and sequencing. Each cell's transcriptome is tagged with a unique molecular identifier or barcode during library preparation to ensure that sequencing reads can be traced back to their cell of origin, enabling simultaneous profiling of thousands of cells in a single experiment (43-45).The key advantage of scRNA-Seq is its ability to reveal transcriptional heterogeneity that is often not shown in bulk analyses. For example, in complex tissues like tumors, brains, or immune environments, distinct subpopulations of cells may have vastly different gene expression profiles despite appearing similar at the tissue level (46). scRNA-Seq makes it possible to classify these subtypes and even discover new or rare cell populations. Additionally, it facilitates the study of dynamic biological processes such as cell differentiation, activation, and response to stimuli (47-51).

Long-read RNA sequencing is an emerging and increasingly impactful technology in the field of transcriptomics. Unlike short-read RNA-Seq, which typically generates fragments of 50-300 base pairs, long-read RNA-Seq is designed to capture entire transcripts, using platforms like Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) to directly sequence long stretches of RNA or cDNA (often spanning entire transcripts in a single read, up to 13800 base pairs, depending on the technology used) (52). This allows for the direct observation of their isoforms, exon-intron structures, and alternative splicing events, without the need for complex computational assembly (53) but comes with higher costs and technical challenges. Long-read sequencing is currently more expensive than short-read methods, generates fewer reads per run, has higher base-calling error rates compared to short-read sequencing and requires high-quality RNA input and specialized bioinformatics pipelines for data analysis (52,54-57). To mitigate

some of these issues, hybrid approaches are commonly used, where long reads provide the transcript models, and short reads provide higher depth quantification (58,59).

Lastly, non-coding RNA (ncRNA) analysis refers to the identification, quantification, and functional characterization of RNA molecules that do not code for proteins, such as microRNAs, long non-coding RNAs (lncRNAs), and circular RNAs (60,61). Its main advantages are its ability to uncover new regulatory molecules involved in key biological processes and diseases and its high sensitivity (62,63). However, it also presents some challenges such as the complexity and diversity of ncRNAs, limited functional annotation, difficulties in distinguishing between similar sequences, and the need for advanced computational tools and expertise (64).

1.6 Measuring gene expression: The importance of normalization in transcriptomics

Normalization is a crucial step in RNA-Seq data analysis because raw read counts alone are not directly comparable across genes or samples. Several biological and technical factors can introduce biases, such as Gene Length, Sequencing Depth and GC-Content (65).

Extracting reliable information from RNA-Seq data involves two crucial steps: thoroughly validating gene identities and choosing the most appropriate way to measure gene expression levels, especially when comparing between different samples and different conditions (for example, comparing healthy vs. diseased samples).

Several normalization methods have been developed to account for differences in sequencing depth and gene length, with the most commonly used metrics being Reads Per Kilobase of transcript per Million mapped reads (RPKM), Fragments Per Kilobase of transcript per Million mapped reads (FPKM), and Transcripts Per Million (TPM) (66-68). However, the scientific community has not yet agreed on the best method for measuring RNA-seq data when comparing different samples (69). The underlying uncertainty stems from a lack of sufficient experimental data obtained from diverse types of replicates (repeated experiments conducted under varying conditions) which makes it difficult to validate any specific approach. Despite ongoing debates about the best method for RNA-Seq quantification, many recent scientific studies, publicly available databases, and online resources continue to use TPM and RPKM/FPKM for pooled data analyses (combining data from multiple samples or experiments to derive insights), cross-sample comparisons and differential expression (DE) analysis (identifying genes that are

expressed differently between experimental conditions), suggesting that, despite the recognition of potential limitations in these methods, they remain widely used in the field (69).

The goal of RPKM is to estimate the relative RNA molar concentration (rmc), which represents the proportion of a specific transcript (gene) within a sample's total RNA content (31). However, for a true measure of RNA abundance, the average expression level across all genes in a sample should remain constant, regardless of sequencing depth (70). According to Zhao, S. (71), RPKM and FPKM do not reliably represent the relative quantification of RNA molar concentration (rmc) because RPKM values fluctuate depending on the total number of transcripts mapped, making it an unreliable measure of RNA molar concentration. To address this issue, TPM was introduced as an alternative because, unlike RPKM, TPM remains constant and directly reflects the relative RNA molar concentration, providing more accurate comparisons between samples. As a result, TPM has been widely adopted in modern computational tools for transcript quantification, including RSEM (RNA-Seq by Expectation-Maximization) (68,72), Salmon (73) and Kallisto (74). TPM can be calculated using Equation 1.

$$TPM_i = \frac{\frac{q_i}{l_i}}{\sum_j (\frac{q_i}{l_i})} \times 10^6$$

where q_i refers to the reads mapped to transcript, l_i is the length of the transcript, and $\frac{q_i}{l_i}$ refers to the sum of j mapped reads to transcript normalized by transcript length (69).

1.7 Challenges and Opportunities in Transcriptomic Data Analysis

There is no denying that the amount of transcriptomic data has been increasing very rapidly over the years. In fact, a quick search in the PubMed database for the term "transcriptome" shows that in just 25 years, the number of papers published with this term has risen from just 183 in 1999 to 35 266 in 2024. This rapid growth highlights the importance and expanding role of transcriptomics in scientific research (Chart 1).

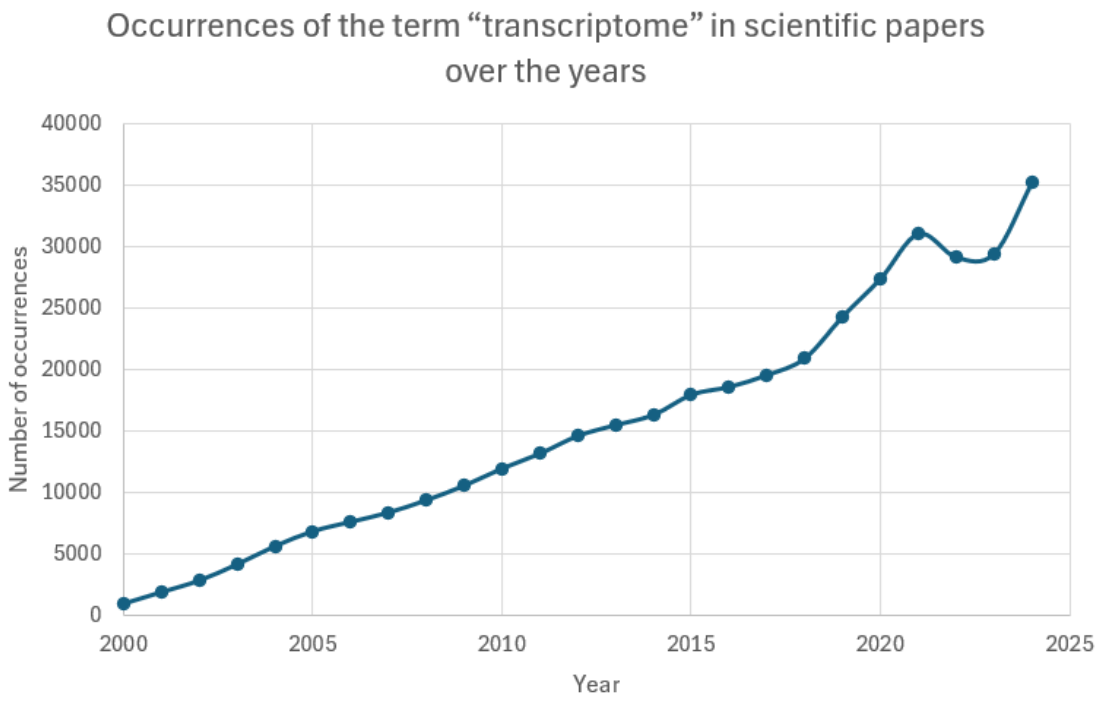

Chart 1. Count of occurrences of the term “transcriptome” in scientific papers within the PubMed database since the year 2000 (Accessed in March, 2025).

On one hand, having access to this large volume of diverse data is beneficial because it allows for more robust findings, enhances reproducibility, and enables the discovery of patterns that may not be evident in smaller and less diverse datasets. However, it also presents a significant challenge in terms of managing and extracting meaningful insights, because processing this data requires bioinformatics and programming knowledge. Therefore, it is becoming increasingly important to develop simple, intuitive, and user-friendly tools that allow researchers to access and interpret gene transcription data and draw conclusions without the need for bioinformatics skills.

Currently, researchers can access transcriptomic data through public repositories such as those maintained by NCBI, which provide extensive RNA sequencing datasets across various species and conditions. However, accessing and processing transcriptomic data often demands advanced programming skills and computational power, which many researchers simply do not have training or access to.

Additionally, some databases offer already curated and standardized data, making transcriptomic data more accessible to researchers. Some examples of these databases are Bgee (https://www.bgee.org/) (75,76), Expression Atlas (https://www.ebi.ac.uk/gxa/home) (77) and GTExPortal (https://www.gtexportal.org/home/) (78).

Bgee and Expression Atlas, are online databases that allow users to search RNA-Seq data by gene and by species with Expression Atlas also offering the option to search by tissue. While both platforms serve primarily as reference repositories for gene expression data, their main purposes differ slightly. Expression Atlas was created as a centralized repository for curated gene expression datasets. In contrast, Bgee functions not only as a data repository but also enables users to input a list of genes and compare their expression across multiple animal species, focusing exclusively on healthy, wild-type conditions. Despite their usefulness, they both lack interactive visualization features that aid in interpreting expression patterns across multiple organisms within specific tissues. Furthermore, neither of them allow the user to analyze their own datasets. As for GTExPortal, like those previously mentioned, it is also an online database. However, at the moment, it is limited to human data and does not support user-submitted data (Figure 2) (78).

In addition to reference databases, there are other bioinformatics tools specifically designed for the analysis and interpretation of RNA-Seq data. Two such tools are Evo-Devo (79) (https://apps.kaessmannlab.org/evodevoapp/) and VastDB (80) (https://vastdb.crg.eu/wiki/Main_Page), which, although similar in allowing to make queries on genes across tissues and species, were developed with distinct purposes. While VastDB focuses on the detection and exploration of alternative splicing events, Evo-Devo aims on exploring gene expression profiles across organs or tissues at different developmental stages and in multiple species. Both tools support searches by tissue, species, and gene. However, they are limited to querying one gene at a time, which makes large-scale or cross-species comparisons within the same tissue more difficult. Additionally, neither provides summary tables listing all expressed genes per tissue or species, which further complicates broader analyses of gene expression patterns.

Figure 2. Feature comparison of major databases and bioinformatics tools. A green check mark denotes that the feature is supported; a grey cross indicates it is not available.

1.8 Potential applications: Conservation of Gene Expression Across Species and Tissues

Numerous studies demonstrate that homologous tissues exhibit remarkably conserved expression patterns (81-85). Although variations may occur due to gene duplication and alternative splicing, the overall pattern of tissue-specific expression remains remarkably stable (86). For instance, studies comparing humans and chimpanzees indicate that roughly 90% of genes are similarly expressed in the brain of the two species (87,88). Another study between humans and mice shows that in the neocortex, about 79% of neural function-related genes show conserved gene expression (89). These studies demonstrate that since homologous tissues across species often share 75–90% of their genes, the distinct characteristics of each species' tissues result primarily from differences in gene regulation and expression patterns, rather than from the presence or absence of genes. These results reinforce the widely held view that non-human vertebrates like rodents remain valuable models for studying the physiology of specific human organs, even after tens of millions of years of evolutionary separation (81).

Therefore, since closely related species tend to share similar tissue-specific expression profiles, transcriptomic data from one species can serve as a useful reference for predicting gene expression in a closely related species. This method of obtaining a predicted gene expression profile would be particularly valuable for species with limited transcriptomic data, for example endangered species, allowing researchers to estimate gene expression patterns/signatures in tissues by using data from phylogenetically related species.

A tissue expression atlas, meaning, a map of genes expressed under baseline conditions for specific tissues and species can have several valuable applications. Firstly, it could be used to identify tissues in cases where its origin is uncertain or ambiguous. By comparing gene expression profiles to established tissue-specific signatures, researchers could infer the most likely source of a given RNA sample. This approach would be particularly useful in contexts such as working with incomplete datasets from non-model organisms. Additionally, a tissue expression atlas could be used to study the evolution of gene expression in a specific tissue by comparing gene expression profiles in homologous tissues across closely related species and identifying how these patterns differ among different phylogenetic groups. Lastly, since this atlas is made of transcriptomic data from baseline conditions, it could serve as a reference for comparing with data obtained under experimental or stress conditions.

1.8.1 Case study - Marine Mammals and Adipose tissue

Marine mammals have developed a series of physiological and genetic adaptations that enable them to survive and thrive in marine environments. One of the most notable adaptations is their lipid metabolism. Cetaceans, in particular, are known for accumulating substantial quantities of lipids in the form of a thick subcutaneous fat layer known as blubber (Figure 3). This specialized adipose tissue exhibits a unique lipid metabolism and serves several essential functions, such as providing thermal insulation in cold marine waters, contributing to buoyancy and acting as an energy reservoir (90). Furthermore, the cetaceans are carnivorous animals with a dietary profile, which consists primarily of high-protein, high-fat prey such as fish and squid, can also contribute to significantly modulate the expression of genes related to lipid synthesis or breakdown (91).

Therefore, studying the lipid metabolism in marine mammals is key to understanding the metabolic adaptations that support their survival in the marine environment and to identifying physiological vulnerabilities that may compromise their conservation in the face of environmental change. However, a major challenge relies in the scarcity of fresh, high-quality adipose tissue samples suitable for RNA-Seq (92). Because cetaceans are protected species and are logistically difficult to study *in vivo* (for example blubber biopsies from live animals in the wild require a highly trained team and strict permitting due to ethical and regulatory considerations), most available samples are derived from stranded individuals or museum specimens, which may be degraded or unsuitable for high-resolution transcriptomic analyses. As a result, researchers have to rely on indirect

inference methods to estimate which genes are likely expressed in cetacean adipose tissue based on data from phylogenetically related species in homologous tissues (93). Only recently has high-quality transcriptomic data become available for a limited number of cetacean species, with most datasets excluding endangered species and lacking representation across different tissues. A search in NCBI reveals that while there are now a few high-quality transcriptomes available for cetacean adipose tissue, data for other tissues remain scarce (94,95).

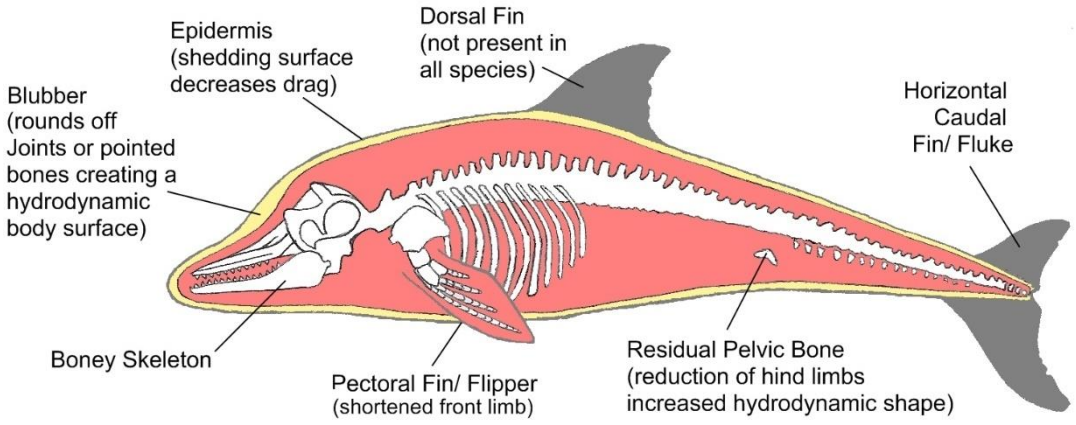

Figure 3. Anatomical adaptations of a cetacean for aquatic life. This diagram illustrates key structural features of a cetacean that enhance its hydrodynamic efficiency and swimming capabilities. Notable adaptations include a blubber layer that streamlines the body, a shedding epidermis to reduce drag, a reduced pelvic bone as a vestige of hind limbs, and modified forelimbs (flippers) for maneuverability. The dorsal fin (present in some species) aids stability, while the horizontal caudal fluke provides propulsion (96).

1.10 Motivation and Objectives

The motivation for this thesis stems from the recognition of the gap between data availability and data usability. In recent years, the field of transcriptomics has seen rapid expansion driven by the development of high-throughput RNA-Seq technologies, which have enabled the generation of a large number of datasets across diverse tissues, species, experimental conditions, developmental stages, etc. Despite this explosion in data, the tools available for non-specialist users to explore, compare, and interpret transcriptomic information remain limited. As a result, there is a growing demand for accessible platforms that empower researchers, to derive biological insights from data without expertise in coding or bioinformatics. In particular, there is a lack of intuitive tools that allow users to explore differences and overlaps in gene expression patterns across homologous tissues in different species.

To address these challenges, this thesis presents the development of TissueXplorer, a web-based bioinformatic tool designed to make comparative transcriptomic analysis more accessible. TissueXplorer integrates processed transcriptomic datasets from various mammalian species and tissues and allows users to explore shared and species-specific gene expression signatures using simple visualizations. By offering a clean, user-friendly interface, the tool lowers the barrier for users with limited computational background to use and analyze the data.

The main objectives of this thesis are as follows:

1. To collect, preprocess, and integrate transcriptomic datasets from multiple mammalian species and homologous tissues.
2. To design and develop TissueXplorer, a user-friendly web-based application for comparative transcriptomic analysis, available through a URL so that it is easily accessible for any researcher in any part of the world

To fulfil these objectives, the TissueXplorer platform aims to meet the following requirements:

- Enable the identification of gene expression overlaps and differences across species in homologous tissues through intuitive visualizations and tables that can be downloaded
- Possess a gene set enrichment analysis feature to aid in biological interpretation

- Allow users to upload and explore their own datasets, enabling comparative analyses against the platform's existing data.
- Enable users to search for individual genes or input gene lists to assess their expression across different tissues and species.
- Have a clean, intuitive, and user-friendly interface suitable for users with limited programming experience.
- Handle moderate-sized datasets efficiently and respond to user queries with minimal delay.
- Be accessible through standard internet browsers by using a consistent URL.
- Maintain a well-documented codebase and data structure to facilitate ongoing updates, maintenance, and collaboration.
- Ensure user data privacy, by not storing or integrating uploaded datasets into the platform's central database.

Chapter 2 – Material and Methods

This chapter discusses the TissueXplorer web application, as well as the set of needs that led to the development of this application and the methodologies used for its development.

2.1 Gap Analysis

The development of the TissueXplorer tool began with a gap analysis, conducted to understand the limitations of existing web applications that allow users to explore and access public transcriptomic data.

This process involved a review of several established platforms and websites including Bgee (75,76), Expression Atlas (77), GTEx (78), Evo-Devo (79), and VastDB (80) during which their respective strengths and limitations were taken into consideration. This overview played a central role in the definition of TissueXplorer's core features, and the insights gained from this gap analysis were then used to outline the key objectives for the development of the application. The collected information was summarized in a figure (Figure 2), which is presented in the Introduction.

2.2 Technologies used

To support the development of the TissueXplorer web application, a combination of programming languages and tools were employed. This section outlines the main technologies used in the development, specifically Python and R, and their roles within the overall workflow.

2.2.1 Python version 3.10.2

Python is a high-level, interpreted programming language known for its readability, simplicity, and versatility. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming (97,98). Due to its large ecosystem of libraries and tools, Python is widely used in scientific computing, data analysis, and bioinformatics.

The version of Python used through the development of TissueXplorer was Python 3.10.2 and it was primarily used for data preprocessing, cleaning, and normalization. The

Pandas library was employed to manipulate and structure the transcriptomic datasets obtained from public databases. Additional scripts were developed to convert gene identifiers into standardized gene symbols using external Application Programming Interfaces (APIs), such as the Ensembl REST API, ensuring consistency across species and finally to prepare the data for integration into a SQLite database.

2.2.2 Programming in R and Shiny

R is an open-source programming language and software environment designed specifically for statistical computing and data analysis. It provides a wide range of statistical techniques, including linear and nonlinear modeling, classification, clustering, and graphical methods. It also offers an extensive suite of packages for manipulating, analyzing, and visualizing high-dimensional biological data, making it particularly popular in bioinformatics and data science (99).

This project leverages the R programming language in conjunction with the Shiny web application framework to build an interactive and user-friendly platform. Shiny is an R-based toolkit for building web applications without requiring expertise in traditional web development languages such as HTML, CSS, or JavaScript. By combining R and Shiny, TissueXplorer offers a robust backend for data processing and analysis, alongside a modern, web-based frontend that is highly customizable. The choice of this tool is motivated by its widespread adoption in the bioinformatics community, its compatibility with SQLite databases, and the wide range of available bioinformatics and data analysis packages.

2.2.3 Database: SQLite

SQLite is a lightweight, embedded, relational database management system that requires minimal setup and runs directly within applications. Unlike traditional client-server databases, SQLite stores the entire database as a single file on disk, making it highly portable and easy to integrate into local applications. Its simplicity, speed, and reliability make it a popular choice for small to medium-sized projects, including data-driven web and desktop applications.

SQLite was selected as the database system for this project due to its simplicity, and ease of integration. Since the entire database is stored in a single local file, SQLite allows for fast data access and easy data sharing, which suits the needs of this project. Its compatibility with both Python and R also ensured seamless integration with the web

application. Additionally, unlike MySQL or other client-server database management systems, SQLite does not require a separate server or complex configuration, making it ideal for lightweight applications and for use during the development and testing phases.

2.2.4 GitHub

GitHub is a web-based platform for version control and collaborative software development that uses Git as its underlying system. It allows developers to host and review code, track changes, manage projects, and collaborate with others in real time. GitHub is widely used in both open-source and private development environments due to its powerful features and integration with continuous development workflows.

Throughout the development of the project GitHub allowed for efficient tracking of changes, and maintenance of a clean and organized development workflow. All source code, scripts, and documentation were stored in a private GitHub repository, serving as a backup for every version of the tool.

2.3 Development of TissueXplorer

2.3.1 Data collection

The transcriptomic data used in this project was obtained from two publicly available gene expression databases: Bgee (75,76) and Expression Atlas (77). These platforms provide already curated RNA-Seq datasets (that contain the final expression values in gene ID or gene symbol format measured in TPM) from a variety of studies across multiple species and tissues.

To ensure comparability across species, only baseline studies were selected, specifically those involving healthy adult individuals without specific experimental conditions. This criterion was applied to minimize variability due to developmental stage, disease state, or experimental conditions, therefore, this standardization was essential minimize variability and allow interspecies comparisons of gene expression across homologous tissues.

Table 1 summarizes the species included and their common names, the tissues in the datasets retrieved, and the respective source database with the accession or reference of the study. TissueXplorer includes a total of 14 species covering major mammalian orders.

Table 1. Overview of species, tissues, and dataset sources used in the study.

Species name	Common name	Source Database	Tissues Included	Study Reference	Citation
Bos taurus	Cow	Expression Atlas	Adipose tissue, duodenum, hypothalamus, kidney, liver, lung, muscle tissue	E-MTAB-2596	(84)
Homo sapiens	Human		Adipose tissue, adrenal gland, bone marrow, colon, duodenum, endometrium, heart, kidney, liver, lung, lymph node, ovary, skeletal muscle tissue, spleen, stomach, testis,	E-MTAB-2836	(100-102)
Monodelphis domestica	Rat		Brain, cerebellum, heart, kidney, liver, testis	E-MTAB-3719	(103)
Rattus norvegicus	Rat		adrenal gland, brain, heart, kidney, liver, lung, spleen, uterus, testis	E-GEOD-53960	(104)
Papio Anubis	Baboon		Bone marrow, cerebellum, colon, heart, kidney, liver, lung, lymph node, skeletal muscle tissue, spleen	E-MTAB-2848	(105)
Macaca mulatta	Monkey		Brain, colon, heart, kidney, liver, lung, skeletal muscle tissue, spleen, testis	E-MTAB-2799	(84)
Canis lupus familiaris	Dog	Bgee	Adipose tissue, kidney, heart, liver, lung, skeletal muscle tissue, spleen	GSE43013	(106)
Cavia porcellus	Guinea pig		Adipose tissue, kidney, heart, liver, lung, skeletal muscle tissue, spleen	GSE43013	(106)
Felis catus	Cat		Adipose tissue, kidney, heart, liver, lung, skeletal muscle tissue, spleen	GSE43013	(106)
Macaca mulatta	Monkey		Kidney, heart, liver, lung, skeletal muscle tissue, spleen	SRP009818	(107)
Mus musculus	Rat		Adipose tissue, kidney, heart, liver, lung, skeletal muscle tissue, spleen	GSE43013	(106)
Oryctolagus cuniculus	Rabbit		Adipose tissue, kidney, heart, liver, lung, skeletal muscle tissue, spleen	GSE43013	(106)

Rattus norvegicus	Rat		Adipose tissue, kidney, heart, liver, lung, skeletal muscle tissue, spleen	GSE43013	(106)
Equus caballus	Horse		Adipose tissue, heart left ventricle[1], hypothalamus, left lung[2], left ovary[3], liver left lateral lobe[4], ovary, spleen, uterus	ERP108802	(108)
Sus scrofa	Pig		Adipose tissue, kidney, cerebellum, heart left ventricle[1], liver, lung, spleen, stomach, testis, uterus	SRP266207	(109)

[1]heart left ventricle; [2]left lung; [3]left ovary; [4]liver left lateral lobe.

2.3.2 Data Preprocessing

The datasets obtained from Bgee and Expression Atlas were downloaded as TSV (Tab-Separated Values) files. However, since they originated from different databases, their internal structures varied significantly, requiring customized processing. To address this, two separate preprocessing functions were developed using Python, along with the packages pandas, numpy and os (operating system). The implemented functions standardized the file formats by enforcing a uniform structure across all datasets. Specifically, the first column contains gene identifiers (IDs), the second column lists gene symbols, and the subsequent columns correspond to tissue names, each populated with the respective TPM values for each gene. An excerpt of the code used to perform these operations is provided in Attachment 1. Figure 4 illustrates the structure of the dataset after the application of these preprocessing functions, highlighting the improvements in consistency and data organization.

Gene ID	GeneSymbol	adipose tissue	adult mammalian kidney	heart	liver	lung	skeletal muscle tissue	spleen
ENSCAFG00000000001		0	4.94	0	0	0	0	0
ENSCAFG00000000005		0	0.72	0	0	0	0	0
ENSCAFG00000000007		0	2.28	0	0	0	0	0
ENSCAFG00000000008		0	12.05	0	0	0	0	0
ENSCAFG00000000009		0	41.89	0	0	0	0	0
ENSCAFG00000000010		0	3.80	0	0	0	0	0
ENSCAFG00000000011		0	0.96	0	0	0	0	0
ENSCAFG00000000012		0	2.66	0	0	0	0	0
ENSCAFG00000000013		0	0.29	0	0	0	0	0
ENSCAFG00000000015		0	0.63	0	0	0	0	0
ENSCAFG00000000016		0	15.70	0	0	0	0	0

Figure 4. Example of the final .csv file structure after data preprocessing. The first column contains gene IDs, the second column, which is empty, is reserved for gene symbols, and the subsequent columns represent tissue names with their respective TPM values.

2.3.3 Gene Identifier to Gene Symbol Conversion

The datasets obtained from Bgee and Expression Atlas utilized Ensembl gene identifiers, which are unique and species-specific alphanumeric codes assigned to genes within the Ensembl database. Because these Ensembl gene IDs are not standardized across different species, direct comparisons of gene IDs are not possible. In contrast, gene symbols serve as standardized identifiers that are often conserved across species, making them more suitable for comparative transcriptomics analyses. Thus, Ensembl gene IDs were converted into their corresponding gene symbols using Python and the Ensembl REST API. A custom Python function was developed to automate the conversion of Ensembl gene IDs into gene symbols by querying the Ensembl database. To address API rate limitations and enhance computational efficiency, the function was engineered to process gene IDs in batches and to execute queries in parallel using the ThreadPoolExecutor from Python's concurrent.futures module. Upon completion, the processed data are exported as a new CSV file, named according to the species and appended with the suffix "_with_gene_symbols.csv".

The following dependencies were used in this process:

```
import pandas as pd

import requests

from concurrent.futures import ThreadPoolExecutor

from tqdm import tqdm
```

An excerpt of the code used for this function is shown in Attachment 2.

2.3.4 Generating the Final Dataset for Integration in SQLite

Before integrating the dataset into the SQLite database, a final preprocessing step was necessary. Specifically, all occurrences of "Unknown" in the GeneSymbol column were replaced with the corresponding GeneID values to preserve the information. This placeholder value typically appears when the API fails to match a given identifier to a known gene, when that specific gene ID is not recognized by the reference database.

After completing the conversion of Gene IDs to gene symbols, the original Gene ID column was removed from the dataset, as all relevant gene information was now contained within the "GeneSymbol" column. Additionally, gene symbols were normalized

to uppercase to ensure consistency. The resulting cleaned dataset was then saved in CSV format, ensuring compatibility with SQLite. This file was subsequently imported into a database using the SQLite Studio application (Figure 5). An excerpt of the code implementing this process is provided in Attachment 3.

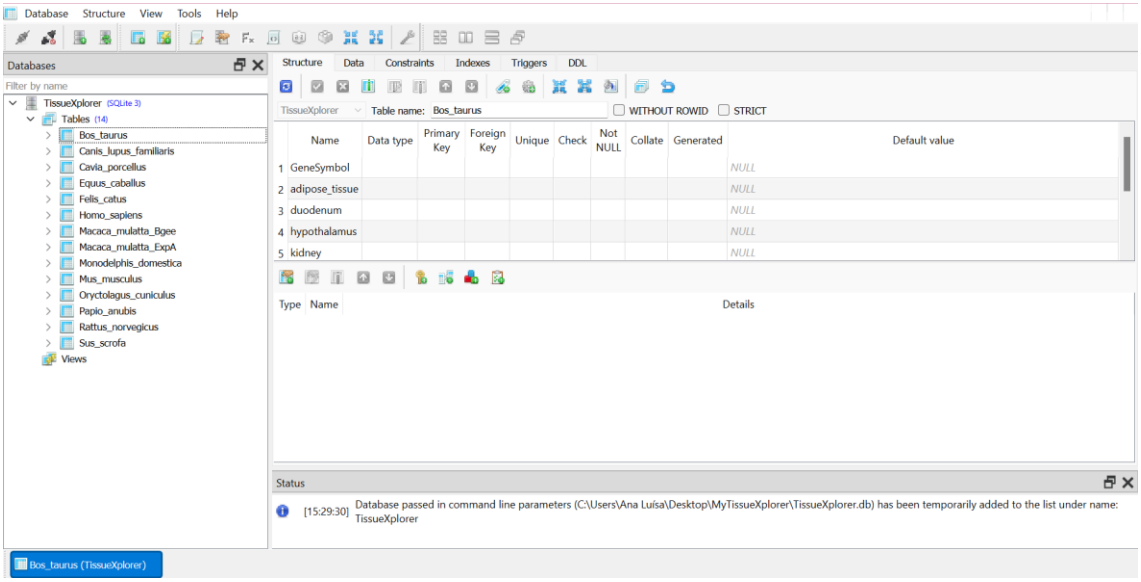

Figure 5. Structure of the TissueXplorer database visualized in SQLite Studio. The screenshot displays the organization of gene expression data for multiple species within the TissueXplorer database. Each table corresponds to a species and contains gene symbols and normalized expression values for various tissues.

2.3.5 Implementation of the TissueXplorer Web Tool Using Shiny

The TissueXplorer web application was developed using the Shiny framework in R, providing an interactive graphical interface for exploring transcriptomic data between species. Shiny facilitates the seamless integration of data processing and visualization tools with reactive components that respond to user input without requiring execution the code manually.

The application architecture follows the standard two-component structure of Shiny apps: the user interface (UI) and the server logic. The UI is defined using fluidPage() or other layout functions from Shiny, such as navbarPage, fullPage or fixedPage. The UI component is responsible for the visual layout and presentation of the application. It defines and arranges all interactive elements, including input fields, action buttons, and display panels for plots and tables. This component determines how users interact with the app and how results are visually communicated. The server component manages the backend computational logic. It includes reactive expressions (programming constructs that automatically update their output whenever the underlying input values

change) along with plotting functions. The server responds to user input from the UI by processing the relevant data and generating updated outputs, such as tables and plots, which are then rendered in the UI. For data management, the Shiny server establishes a connection to the SQLite database using the RSQLite package. When a user selects a gene or species of interest, the server queries the database for the corresponding data and dynamically generates the appropriate visualizations or tables for display in the application.

Once development and testing of the Shiny application were completed locally, the app was deployed using shinyapps.io, a cloud-based platform provided by Posit for hosting Shiny web applications. This platform enables publishing the application online without the need to own and maintain server infrastructure. The R code can be found in the following link: https://tissuexplorer.shinyapps.io/TissueXplorer/ .

The TissueXplorer application uses the following R packages:

```
library(shiny)
library(DBI)
library(RSQLite)
library(VennDiagram)
library(ggplot2)
library(reshape2)
library(pheatmap)
library(dplyr)
library(shinycssloaders)
library(enrichR)
library(DT)
library(rsconnect)
```

Chapter 3 – Results and Discussion

This chapter presents the main results arising from the development and implementation of *TissueXplorer, a* web-based platform available at https://tissuexplorer.shinyapps.io/TissueXplorer/

In addition to its online availability, TissueXplorer application can be run locally in RStudio. Users should first install R and RStudio, then execute the source code provided on GitHub (https://github.com/PTLunna/TissueXplorer).

This chapter begins with a practical walkthrough of TissueXplorer, demonstrating its core functionalities and guiding users through the process of exploring and comparing gene expression profiles across multiple species and tissues. This is followed by two case studies designed to showcase the platform's capacity to generate biological insights. Subsequently, a comparative analysis is provided, positioning TissueXplorer in the context of existing tools and databases for cross-species transcriptomics. The chapter concludes with a discussion of the platform's current limitations and potential opportunities for future development and improvement.

3.1 TissueXplorer's Interface

TissueXplorer was developed with a goal to provide an intuitive and accessible interface for researchers to explore and compare gene expression patterns in the same tissue across several mammalian species without the need for advanced computational skills. Therefore, the interface was designed with a minimal layout to ensure ease of use (Figure 6).

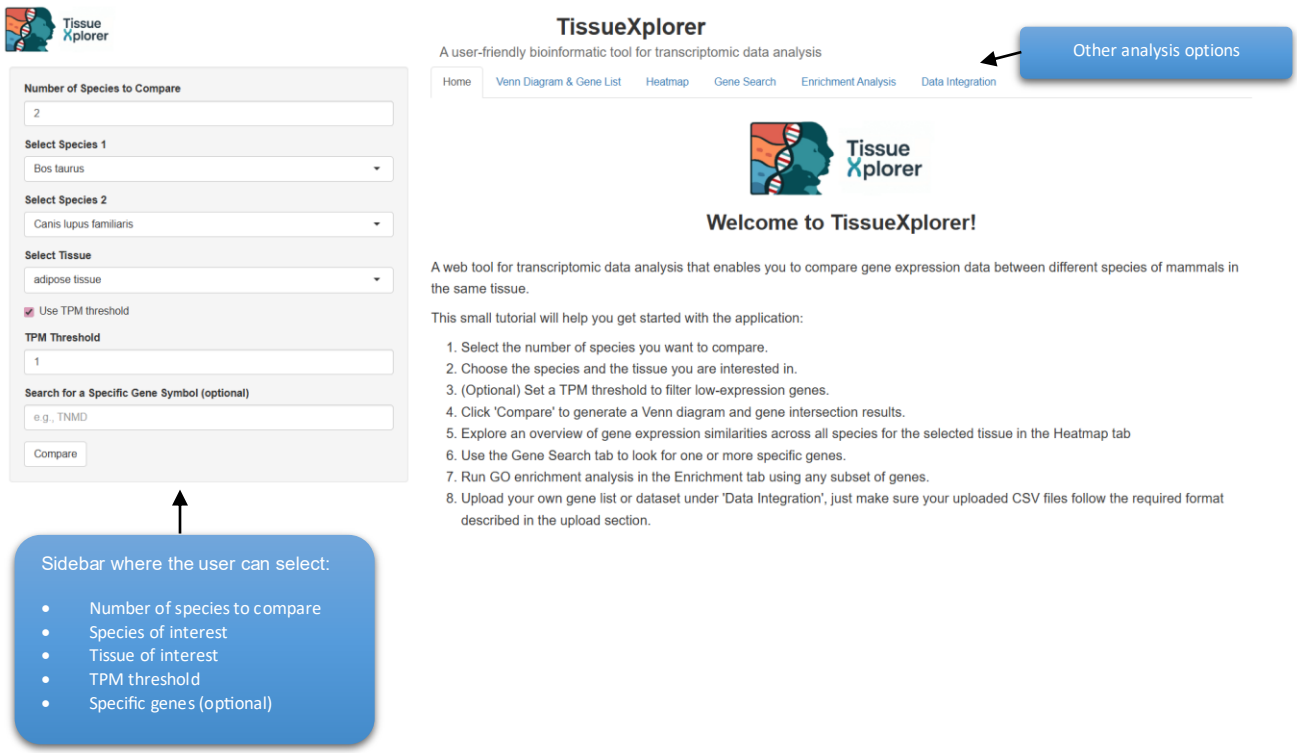

Figure 6. Home page and user interface of the TissueXplorer web platform. The figure displays the main interface of TissueXplorer, highlighting key features such as species and tissue selection, TPM threshold filtering, and gene search functionality. The central panel provides an introductory tutorial and navigation tabs for other analysis modules.

Upon accessing the application, users are greeted with a homepage that introduces the tool and its functionalities. From the main dashboard, users can navigate through the following key sections:

- Venn Diagram & Gene List;
- Heatmap;
- Gene Search;
- Enrichment Analysis;
- Data Integration;

3.2 How to use TissueXplorer and interpret the results

3.2.1 Venn Diagram & Gene List

The Venn diagram module of TissueXplorer enables comparative analysis of gene expression across multiple species and tissues. This section presents the main features and functionalities of the module as implemented in the web platform.

On the left side of the interface, users will find a sidebar panel where they can set the parameters for their analysis, such as selecting the number of species to compare, followed by choosing the species and tissue of interest. Optionally, users may also define a TPM threshold, which sets a minimum expression level for genes to be considered expressed in the selected tissue.

Once all selections are made, clicking the "Compare" button will generate a Venn diagram on the right side of the screen. This diagram visually represents the overlap of expressed genes among the selected species within the chosen tissue. The central overlap indicates genes shared across all species and the outer sections display the genes that are uniquely expressed in each individual species (Figure 7). These Venn diagrams offer users a general overview of the data, illustrating the proportion of genes shared between two species. For a more detailed examination, users can refer to the table below the diagram, which lists individual genes. Within this table, the user can select which set to visualize (shared or species-specific genes). This table can be downloaded as a .csv file by clicking the "Download Selected Genes" button. Lastly, the downloaded file contains a single column titled "GeneSymbol", which can later be uploaded in the "Data Integration" tab to perform further custom comparisons.

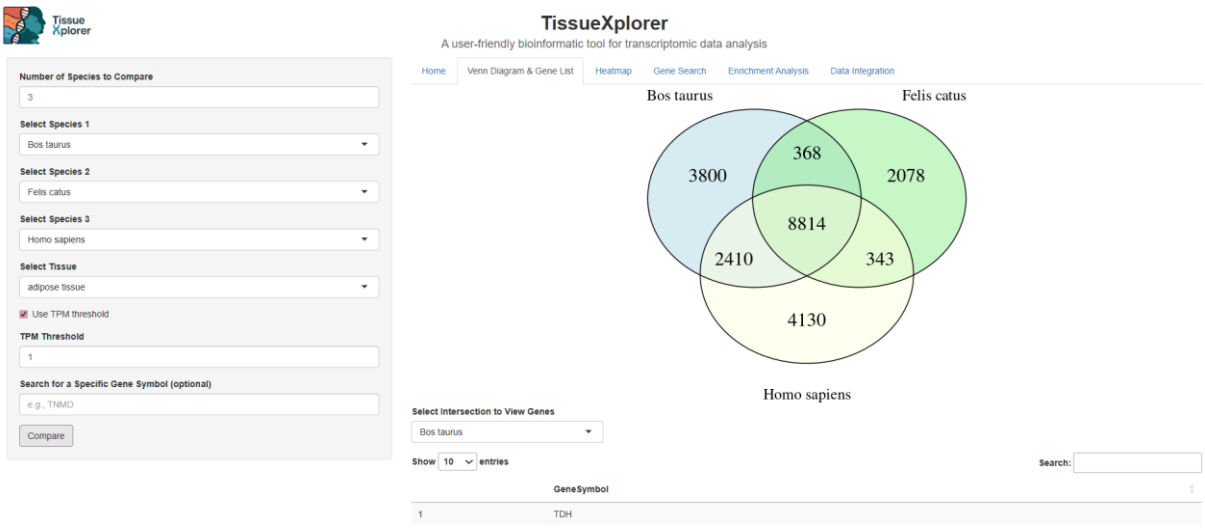

Figure 7. Venn diagram and gene list tab in the TissueXplorer platform. This figure illustrates the Venn diagram analysis feature of TissueXplorer that visually represents shared and unique gene sets among the chosen species. The corresponding gene list for each intersection can be viewed and searched in the table below.

3.2.2 Heatmap

The Heatmap module provides a comprehensive visualization of gene expression overlap across species for a selected tissue. After choosing a tissue from the sidebar, the application generates two complementary heatmaps to facilitate comparative transcriptomic analysis. The first heatmap displays the absolute number of shared genes expressed between each pair of species, with each cell representing the raw count of overlapping genes. The diagonal of the heatmap shows the total number of genes expressed in each species for the selected tissue, for example, in human adipose tissue, a total of 21150 genes are identified as expressed (Figure 8).

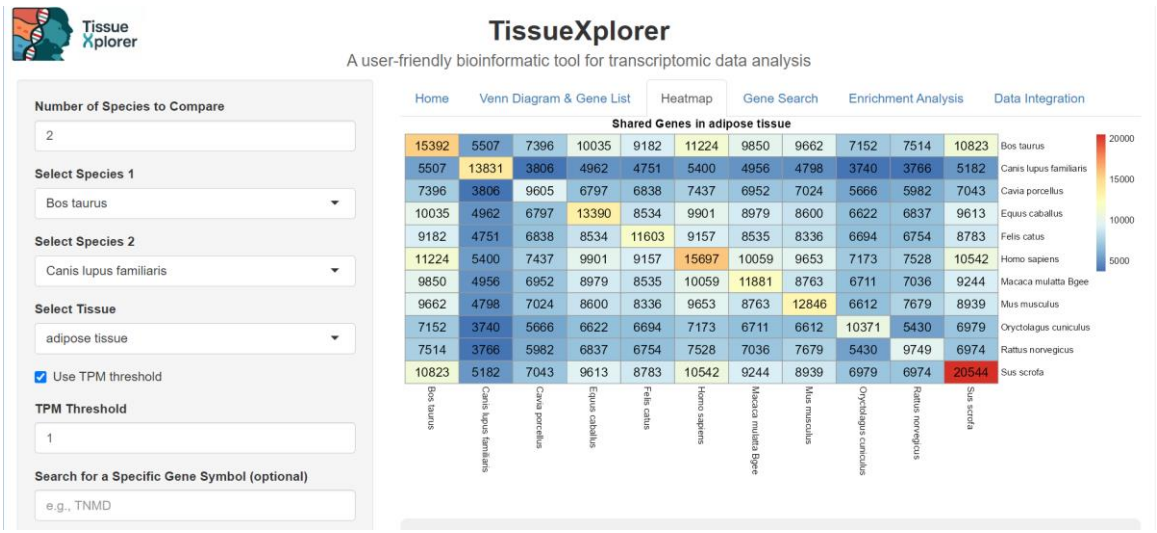

Figure 8. Heatmap of shared gene expression in adipose tissue across multiple species in TissueXplorer. This heatmap, generated from the Heatmap tab of the TissueXplorer platform, displays the number of shared expressed genes (based on a TPM threshold of 1) in adipose tissue across different species. Warmer colours (e.g., red and orange) indicate higher numbers of shared genes, while cooler colours (blue shades) indicate fewer shared genes.

The second heatmap presents the percentage of genes shared between species, offering a normalized view of expression similarity that accounts for differences in total gene counts. To interpret this heatmap, users should first locate Species 1 on the vertical axis (rows) and Species 2 on the horizontal axis (columns); the value at the intersection indicates the percentage of genes from Species 1 that are also found in Species 2 for the selected tissue and TPM threshold. For instance, if we check Figure 9, consider *Cavia porcellus* as Species 1 and *Bos taurus* as Species 2, it means that 77% of the genes expressed in *Cavia porcellus* are also expressed in *Bos taurus*. This second heatmap adds another layer of insight, helping to reveal asymmetric relationships and to identify species with unusually high or low relative overlap.

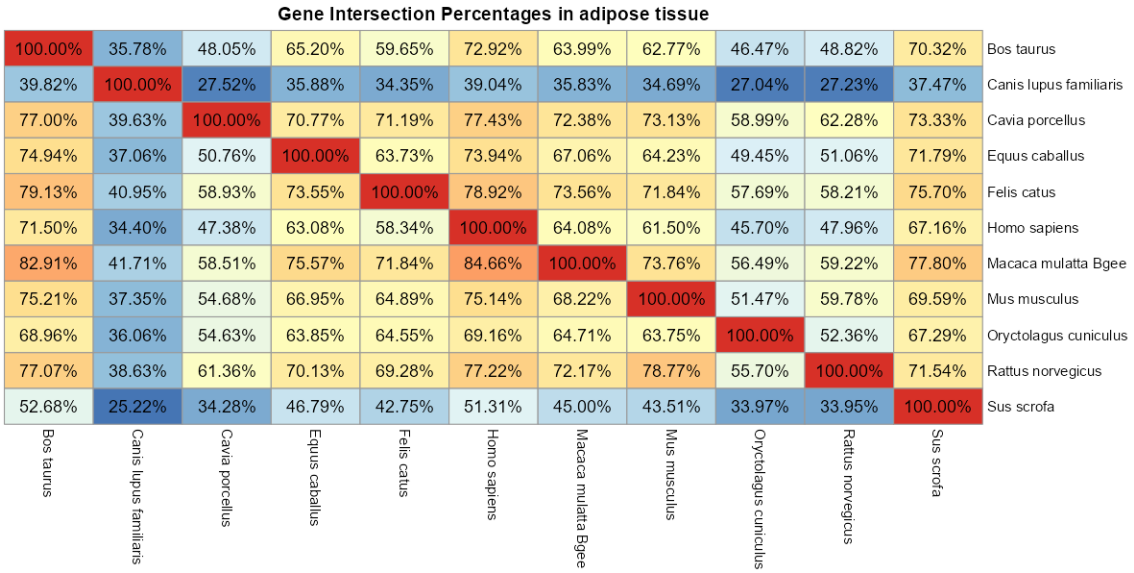

Figure 9. Heatmap illustrating pairwise gene intersection percentages in adipose tissue across multiple species. Each cell represents the percentage of genes expressed in the adipose tissue of one species (row) that are also found in the corresponding species (column). Higher percentages (red) indicate greater overlap in gene expression profiles, while lower percentages (blue) indicate less similarity. This comparative analysis highlights the degree of conservation and divergence in adipose tissue gene expression among the examined mammalian species.

3.2.3 Gene search

The Gene Search module provides users with a targeted approach to explore the expression patterns of specific genes across multiple tissues and species. To utilize this feature, users enter a gene symbol, or a list of gene symbols, separated by commas into the input field located in the sidebar (Figure 10). Upon submission, the application displays in the “Gene Search” tab the expression levels (measured in TPM values) of the selected gene(s) across all available tissues and species in the database. This feature is particularly useful for researchers who are interested in tracking the behavior of a particular gene and compare its activity across species.

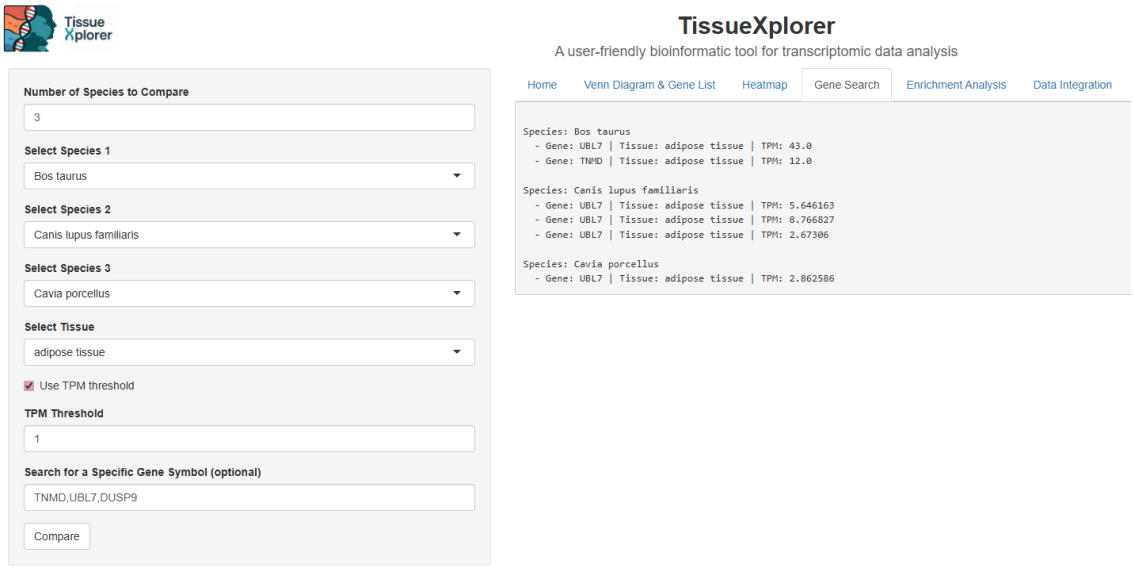

Figure 10. Gene expression query using the Gene Search tab of TissueXplorer and randomly selected gene symbols. This screenshot illustrates the use of the Gene Search feature in TissueXplorer, which allows users to explore gene expression across multiple species and tissues, filtering by TPM thresholds.

3.2.4 Enrichment Analysis

The Enrichment Analysis tab allows users to explore the biological significance of a selected gene set by identifying overrepresented functional categories. When using this feature, users can choose from six of enrichment databases, three referring to gene ontology (GO) libraries (GO Biological Process, GO Molecular Function, GO Cellular Component) and the remaining three regarding metabolic pathway libraries: KEGG (Kyoto Encyclopedia of Genes and Genomes) (110), WikiPathway (111) and Reactome (112).

To perform an enrichment analysis, users must first select a list of genes from the “Venn Diagram & Gene List” tab. Once the gene list is submitted, the tool sends the data to an external enrichment analysis API (EnrichR (113-115)), which performs statistical tests to identify overrepresented functional categories within the input gene set. The results are returned as an interactive table and a bar plot, highlighting the most significantly enriched biological processes, pathways, or molecular functions associated with the selected genes (Figure 11).

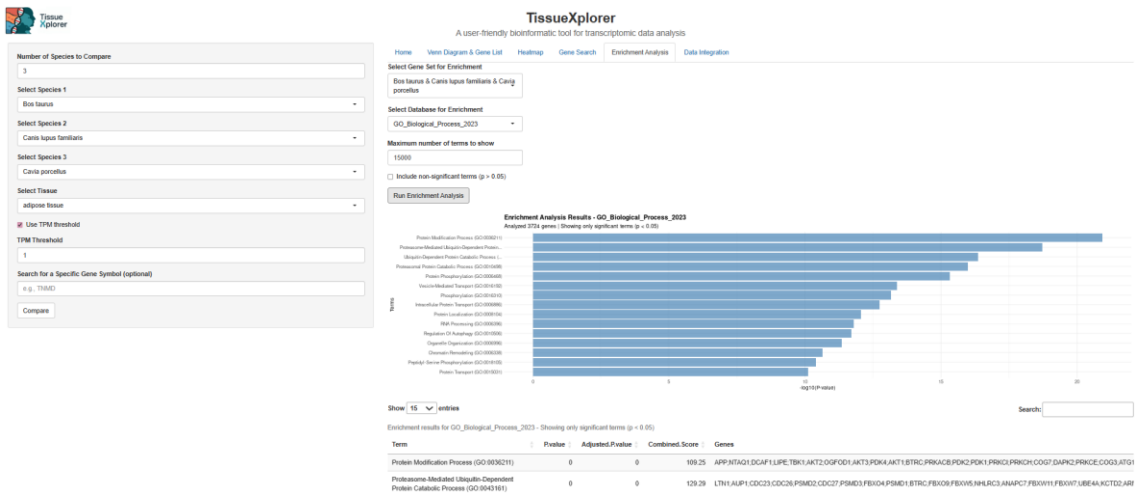

Figure 11. Functional enrichment analysis using the Enrichment Analysis tab of TissueXplorer. The bar chart displays the top enriched GO terms based on the −log10(p-value). The accompanying table includes detailed enrichment statistics such as adjusted p-values, combined scores, and a list with the associated genes.

3.2.5 Data Integration by the user

Lastly, the “Data Integration” tab allows users to incorporate their own gene expression datasets for personalized analyses into the TissueXplorer platform (Figure 12). To upload a file, simply provide a dataset name and select a .csv file. Two types of files are supported: a full dataset or a simple gene list.

A complete dataset should be a CSV file containing gene expression data from a single species across multiple tissues. The first column must be labeled 'GeneSymbol' and include the gene symbols present in all tissues. Subsequent columns should be named after the tissues, each containing the corresponding TPM values for each gene symbol. If TPM values are missing for any tissue, these should be replaced with zero and if a gene symbol is missing, the corresponding gene ID should be used as a placeholder.

A gene list is a CSV file with a single column labeled 'GeneSymbol', listing gene symbols without associated TPM values. Such lists can be generated, for example, by downloading results from the “Venn Diagram” tab. To streamline future analyses and comparisons, it is recommended to include the species and tissue information in the dataset name.

Once uploaded, the user-provided data are treated as an additional dataset within TissueXplorer. This enables users to conduct comparative analyses with existing species and tissues, visualize gene overlaps using Venn diagrams, assess gene presence across tissues, and perform enrichment analysis on their custom gene sets.

For comparisons that do not require filtering by expression level, users can disable the "Use TPM threshold" option in the sidebar. This allows all genes from the uploaded dataset to be included in the analysis, regardless of their expression values.

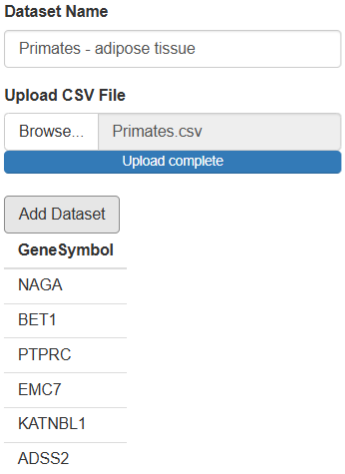

Figure 12. User data upload interface in TissueXplorer. Users can assign a dataset name and upload a CSV file containing gene expression data. Uploaded gene symbols are previewed, and datasets can be added for further analysis and visualization within the platform.

3.3 Comparative Feature Analysis of TissueXplorer and Existing Tools

In the context of transcriptomic analysis, several platforms have been developed to explore gene expression patterns across species, tissues, and developmental stages. Tools such as Bgee (75,76), Expression Atlas (77), GTEx (78), VastDB (80), and Evo-devo (79), serve as important resources for accessing gene expression data. However, these platforms can vary in their functionalities. In this section, we present a feature-based comparison of these tools with TissueXplorer, highlighting the specific strengths and innovations introduced by TissueXplorer.

Figure 13 summarises this comparison across seven key functionalities: the ability to search by gene, species, and tissue; support for user-supplied expression data; the capacity to visualise overlap between species at the tissue level; exportable tables of expressed genes; and the presence of enrichment analysis tools. Each platform is evaluated based on whether it supports these features or not.

Figure 13. Feature comparison of publicly available gene expression databases and analysis platforms. This table presents a side-by-side evaluation of six platforms Bgee, Expression Atlas, GTEx, VastDB, Evo-devo, and TissueXplorer, based on key capabilities such as search functionality (by gene, species, and tissue), support for user-supplied data, interspecies tissue overlap visualization, genes expression data, and enrichment analysis tools.

More specifically, when analyzing each feature, we find that:

- **Gene search** is a core feature present in all platforms.
- **Species search** is supported by all platforms except GTEx, which is focused exclusively on human data.
- **Tissue search** is available in all web applications except Bgee
- **Only TissueXplorer supports user data upload**, enabling researchers to analyse and visualise their own RNA-Seq data alongside public datasets.
- The ability to generate **visual overlap representations** (e.g., Venn diagrams) **between species by tissue** is exclusive to TissueXplorer, offering a unique and intuitive way to identify conserved and species-specific expression signatures.
- A **table of expressed genes** can be retrieved in Bgee, Expression Atlas, GTEx, and TissueXplorer, which is helpful for manual inspection and downstream analyses.
- **Enrichment analysis functionality** is only provided by Bgee and TissueXplorer, facilitating the interpretation of gene sets without requiring external tools.

In summary, although TissueXplorer may not match the scale or data complexity of some larger databases, it addresses key gaps by supporting customizable, multi-species comparative analyses. Its integration of features such as intuitive Venn diagram visualizations, enrichment analysis, and the ability to upload user data distinguishes it from existing tools. Additionally, TissueXplorer reduces technical barriers and offers a user-friendly interface for examining conserved and species-specific gene expression patterns. This makes it a valuable and complementary tool in evolutionary and comparative gene expression studies.

3.4 Case Study 1 - Comparative analysis of the transcriptome of 11 mammalian species in the adipose tissue.

This case study demonstrates the application of TissueXplorer in a transcriptomic analysis focused on the adipose tissue of 11 different mammalian species. Adipose tissue was selected due to its broad representation across species in available datasets, allowing for a more comprehensive interspecies comparison. Additionally, its key roles in metabolism and endocrine function make it a relevant tissue for exploring both conserved and species-specific gene expression patterns.

To replicate this case study using the TissueXplorer platform, follow the steps below:

1. Set the TPM threshold to 1 to exclude lowly expressed genes.
2. In the “Number of Species to Compare” field, select 2.
3. From the species dropdown menus, choose any two species for which adipose tissue data is available.
4. Click the “Compare” button to generate the gene intersection analysis.
5. Navigate to the Heatmap tab to visualize the overlap and divergence in gene expression patterns between the selected species.

To analyse all 11 mammalian species simultaneously, the “Heatmap” tab of TissueXplorer was used to generate two heatmaps (Figure 14 and Figure 15). Figure 14 displays the absolute number of shared genes expressed in adipose tissue between each pair of species and the diagonal cells represents the total number of genes detected in each species. Overall, the number of shared genes between species tends to be high, with most pairwise comparisons exceeding 7000 shared genes. Notably, *Sus scrofa* (pig)

presents the highest total number of expressed genes in adipose tissue (20544), followed by *Homo sapiens* (human, 15697) *Bos taurus* (cow, 15392) and *Canis lupus familiaris* (dog, 13831). The high number of genes expressed in the adipose tissue compared to the number of protein-coding genes of each species is a strong indicator that the adipose tissue metabolically active tissue, with functions that go beyond energy and fat storage.

However, a distinct dark blue pattern is visible in both the row and column corresponding to *Canis lupus familiaris*. This indicates that, despite having a high total number of expressed genes, the overlap with other species is considerably lower. This result is likely to be due to problems in annotation quality, rather than reflecting true biological divergence.

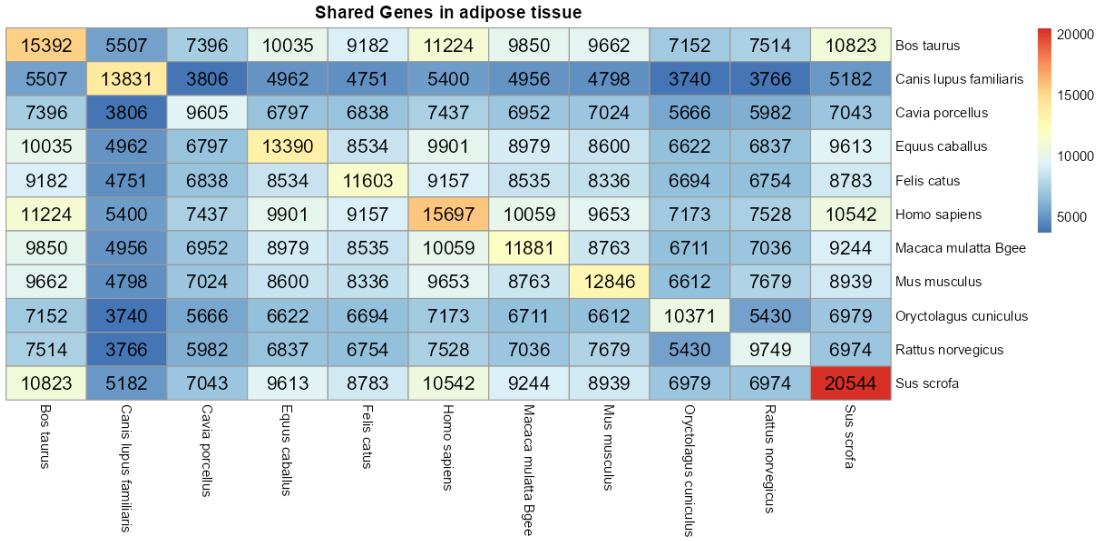

Figure 14. Heatmap showing the absolute number of genes shared between the adipose tissue transcriptomes of 11 mammalian species. Each cell represents the number of overlapping genes between a pair of species, while the diagonal indicates the total number of expressed genes in each species

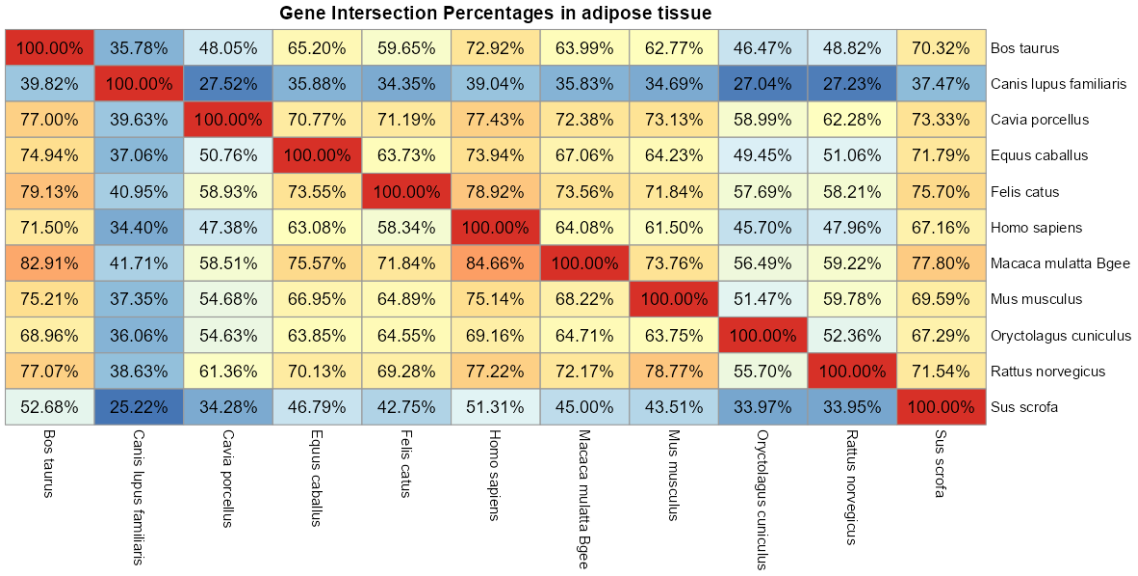

Figure 15. Heatmap illustrating the percentage of shared genes expressed in adipose tissue across 11 mammalian species. Values are normalized to account for differences in the total number of expressed genes per species. Each cell represents the proportion of genes shared between the species indicated on the corresponding row and column. Warmer colours (e.g., red) indicate higher gene intersection percentages, while cooler colours (e.g., blue) represent lower overlap.

Figure 15 shows the percentage of gene overlap between species, normalized by each species' total gene count. This allows for a clearer understanding of relative similarity, independent of dataset size. The highest percentages of gene overlap occur between phylogenetically related species, as expected. For instance, *Macaca mulatta* (Bgee) shares 84,66% of its genes with *Homo sapiens* (human), a closely related species. Similarly, *Rattus norvegicus* (rat) shares 78.77% of their adipose tissue-expressed genes with *Mus musculus* (mouse).

Once again, *Canis lupus familiaris* displays a distinct pattern. Both its row and column are dark blue, indicating that the dog shares a low percentage of genes with all other species and vice versa. This strongly suggests potential issues with annotation quality or transcriptome coverage for this species.

Sus scrofa also stands out, but displays a different pattern compared to *Canis lupus familiaris*. While Sus scrofa's row is relatively dark blue, indicating that only a small percentage of pig genes are shared with other species, its column is mostly orange, meaning that many genes from other species are also present in the pig's dataset. This asymmetry implies that *Sus scrofa* possesses a considerable number of genes that are not found in the other species. To explore the potential functions of these unique genes, a subsequent enrichment analysis was performed focusing on the genes exclusive to *Sus scrofa* in the comparison with *Bos taurus*. However, no significantly enriched terms were identified (p-value < 0.05).

In conclusion, these types of visualizations provide a useful overview of gene expression similarity across many species. Beyond highlighting general patterns of conservation and divergence, they are also particularly useful for detecting potential inconsistencies or anomalies within the datasets. One notable example is *Canis lupus familiaris*, which consistently exhibits low gene overlap with other species. This discrepancy may result from challenges during the gene conversion process from Gene IDs to standardized gene symbols, where a substantial number of canine gene IDs could not be matched to official symbols. These unmatched entries were retained as species-specific Gene IDs, which are not directly comparable across different organisms. The limited effectiveness of the normalization Python function could be attributed to the fact that the canine genome is particularly complex and exhibits high levels of structural variation, which poses challenges for accurate annotation, additionally, the *Canis lupus familiaris* genome remains partially incomplete, and many regions are still poorly characterized. (116,117). Together, these factors likely contribute to the observed divergence for *Canis lupus familiaris*, reflecting technical challenges rather than true evolutionary differences.

Nevertheless, this tool proves highly valuable, as it not only facilitates systematic comparisons across numerous species but also helps identify those with incomplete or unreliable genome annotations. By flagging such species, this tool enables researchers to account for potential data limitations and avoid drawing premature or inaccurate biological conclusions.

3.5 Case study 2 – Is There a Conserved Gene Expression Signature across different taxonomic groups in the adipose tissue?

This case study investigates whether distinct taxonomic groups such as Artiodactyla, Carnivora, and Primates share common gene expression profiles in adipose tissue or exhibit group-specific transcriptional signatures.

To obtain a representative list of genes for each taxonomic group, the “Venn Diagram & Gene List” tab was used to select two species within each taxonomic group: *Bos taurus* and *Sus scrofa* (Artiodactyla), *Felis catus* and *Canis lupus familiaris* (Carnivora), and *Homo sapiens* and *Macaca mulatta* (Primates). The results of these comparisons are shown in Figure 16.

For each pairwise comparison, the list of shared genes was extracted as a proxy for genes conserved within that taxonomic group. These shared gene sets were later used

for comparative analyses across groups. The resulting gene intersections are presented in Figure 16.

To replicate this analysis in TissueXplorer, configure the following parameters:

1. Set the TPM threshold to 1 to exclude low-expression genes.
2. Select “Number of Species to Compare”: 2.
3. Choose a pair of species belonging to the same taxonomic group.
4. In the “Venn Diagram & Gene List” tab, click “Compare” to retrieve the list of shared genes.
5. Repeat for each of the three taxonomic groups.

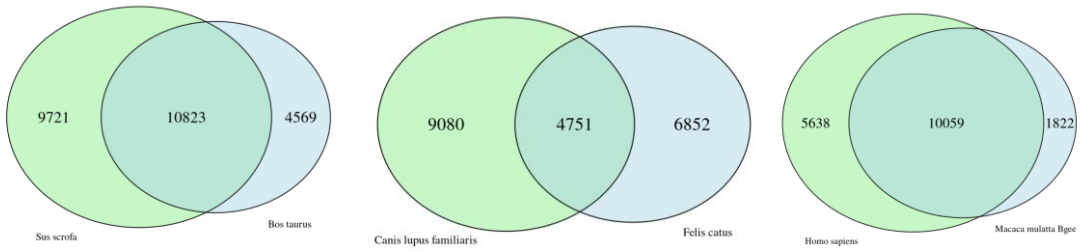

Figure 16. Venn diagrams showing shared and species-specific genes in adipose tissue across three mammalian orders. The left panel represents Artiodactyla, comparing *Sus scrofa* (green) and *Bos taurus* (blue). The center panel illustrates Carnivora, with *Canis lupus familiaris* (green) and *Felis catus* (blue). The right panel presents Primates, comparing *Homo sapiens* (green) with *Macaca mulatta* Bgce (blue). Overlapping regions indicate the number of genes commonly expressed between the two species in each panel, while non-overlapping areas represent the number of genes uniquely expressed in a given species.

After clicking the "Select Intersection to View Genes" button, we selected the shared genes between the compared species and downloaded the corresponding .csv using the “Download Selected Genes” option. The resulting .csv file was then reuploaded into the platform via the “Data Integration” tab, and a new analysis was performed using these taxonomic groups. The results are presented in Figure 17.

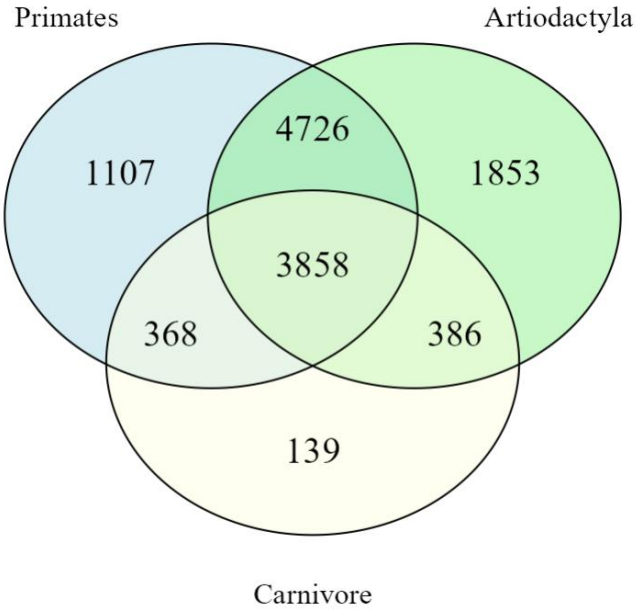

Figure 17. Venn diagram illustrating the distribution and shared expression of genes among three mammalian orders. Each circle represents one order, Primates, Artiodactyla, and Carnivora with overlapping regions indicating genes commonly expressed across two or all three groups. The diagram highlights both unique and shared gene sets, offering insight into conserved and divergent genetic profiles within adipose tissue.

The results identified 3858 commonly expressed genes across all six species in adipose tissue, indicating a highly conserved set of core biological functions. This observation is consistent with previous large-scale comparative studies that have demonstrated the existence of a conserved transcriptomic signature in adipose tissue, particularly for genes involved in lipid metabolism, and adipogenesis (118,119). Interestingly, despite this shared core, the degree of gene expression overlap between taxonomic groups varied considerably. For example, Primates and Artiodactyla exhibited a notably higher number of shared genes in adipose tissue (4726 genes) compared to their respective overlaps with Carnivores (368 genes shared between Carnivores and Primates, and 386 genes between Carnivores and Artiodactyla). These findings suggest that while there is a core gene expression signature conserved across all groups, Primates and Artiodactyla may possess additional, shared group-specific transcriptional programs in adipose tissue. A previous comparative study has reported similar findings, where closely related lineages such as Primates and Artiodactyla display greater transcriptomic similarity, likely reflecting their evolutionary proximity and shared physiological adaptations (120). Meanwhile, the lower overlap with Carnivores could indicate more divergent gene expression patterns in the adipose tissue. This divergence may be attributed to distinct metabolic adaptations related to their obligate carnivorous diet, as has been observed in studies comparing lipid metabolism and adipocyte differentiation across species (121,122). The reduced number of shared genes in Carnivores could

also reflect differences in adipose tissue function, such as altered lipid storage, mobilization, or endocrine signaling, which have been reported in both transcriptomic and physiological studies (123,124).

Next to gain insight into the main biological functions and metabolic pathways, an enrichment analysis was performed in the 3858 common genes between all groups of species, the results of this analysis are presented in Figure 18.

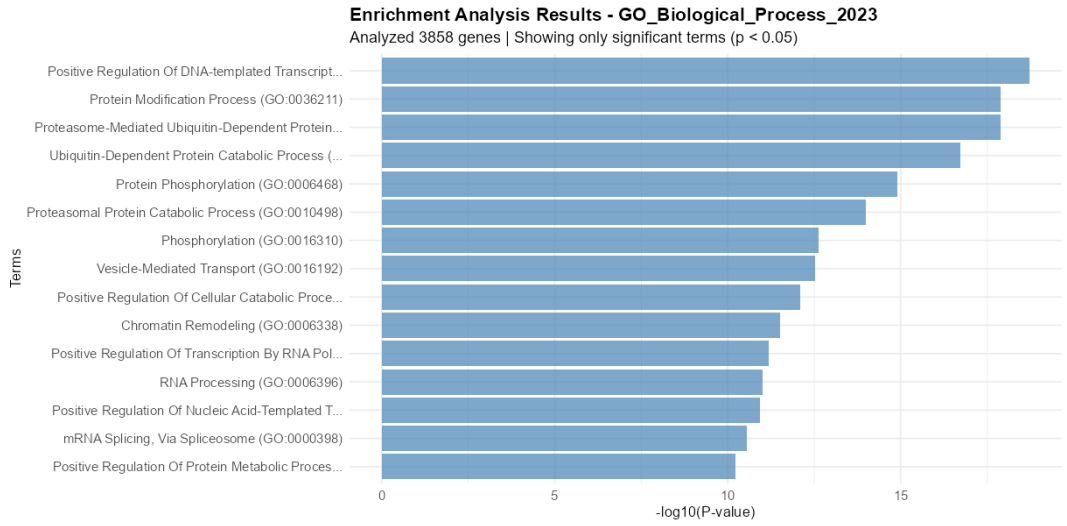

Figure 18. GO enrichment analysis of genes commonly expressed between Primates, Artiodactyla, and Carnivora in adipose tissue. The analysis was performed on 3858 shared genes, highlighting only significantly enriched biological processes ($p < 0.05$). The x-axis indicates statistical significance ($-\log_{10}$ p-value), while the y-axis lists the top GO terms associated with the shared gene set.

A total of 1179 Gene Ontology biological process terms were returned by the enrichment analysis. The most significant hits cluster into Biological functions that govern cellular regulation and homeostatic maintenance, which is not surprising since these are “core” or “housekeeping” processes, so it is expected that they dominate the most significant hits of the enrichment table. However, by taking advantage of the search bar in the upper-right corner of the interactive results table, it is possible quickly filter the 1 179 significant GO-BP terms for concepts that are more relevant to adipose physiology (Figure 19).

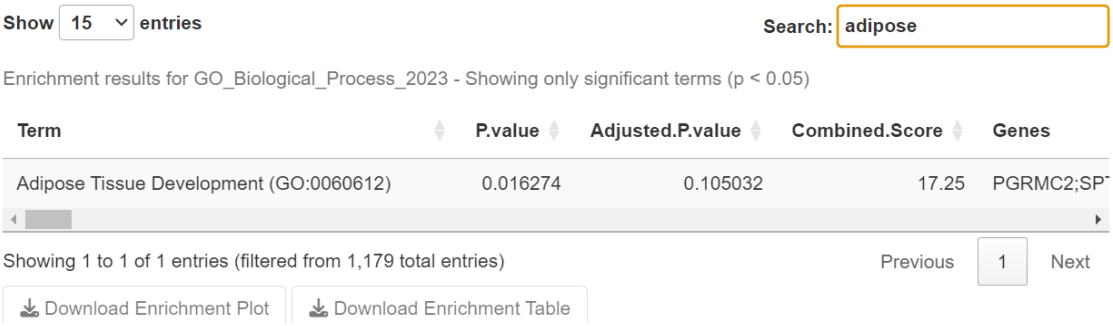

Figure 19. Example of the search functionality in the enrichment analysis output. By using the keyword "adipose" in the search bar, it was possible to filter and identify relevant biological processes, such as "Adipose Tissue Development" (GO:0060612), with a p-value of 0.016274. This feature allows for quick exploration and targeting of specific biological themes within a large list of enriched GO terms (1179 entries in total), enhancing the interpretability of results.

Thus, taking advantage of TissueXplorer's built-in quick filter, the next step was to identify significantly enriched terms specifically related to lipid metabolism and to obtain a core list of genes expressed in adipose tissue of these six species. To achieve this, a search was performed using specific keyword strings, as listed in Table 2. The resulting distribution of enriched terms and their thematic groupings is also summarised in the same table.

Table 2. Keyword-guided refinement of significant GO-BP terms.

Searched string	**Gene symbols** (only the first 10 gene symbols are shown to conserve space)	**Enriched categories**
"Adipose"	*1. PGRMC2;SPTLC2;NAMPT;SH3PXD2B; DYRK1B; VPS13B*	1.Adipose Tissue Development
"Ceramide"	*1. CERS4;CERK;SGMS1;AGK;GBA1;SGMS2; CLN8;ACER2;TM9SF2;SMPD2;*	1.Ceramide Metabolic Process
"Cholesterol"	*1. EHD1;MSR1;SCARB1;LCP1 2. EHD1;MSR1;SCARB1;NR1H2;NR1H3; LCP1 3. ARV1;HMGCS1;INSIG2;INSIG1;MSMO1; DHCR24;HMGCR;TM7SF2;FDFT1*	1.Positive Regulation Of Cholesterol Storage 2.Regulation Of Cholesterol Storage 3.Cholesterol Biosynthetic Process
"Lipid"	*1. PIGO;PIGN;WDR45B;PIGQ;ZDHHC5;*	1. Protein Lipidation

	ZDHHC6;ZDHHC7;ZDHHC21;WDR45; ZDHHC1; *2. PDGFRB;CSF1R;PDGFRA;PTAFR;OCRL; IGF1;PLCB4;RPS6KB1;PI4KA;PLCE1;* *3. HACD1;SGMS1;GBA1;SGMS2;CLN8;HACD4; SMPD2;SMPD4;SPTLC2;SMPD1;* *4. PI4K2B;VAC14;SLC27A1;MTMR2;MTMR3; SLC44A2;DGKA;MTMR14;INPPL1;MTMR4;* *5. EHD1;MSR1;SCARB1;ZC3H12A;PLIN3; FITM2;PLIN2;LCP1;OSBPL11;NFKB1* *6. RARG;TMEM161A;ALAS1;HMGB2; LY96;LIAS;FOXO1;STRN3;RXRA; IRAK1;* *7. NFKBIA;NR1H2;NR1H3;ITGAV;PTPN2* *8. RB1CC1;RAB3GAP2;GBA1; ULK1;RAB3GAP1* *9. HACD1;CERK;SGMS1;SGMS2;CLN8; HACD4;SMPD2;ARV1;SPTLC2;PPP2R1A;* *10. CERS4;HACD1;SGMS1;SPHK2;ELOVL6; SGMS2;HACD4;ACER2;SCCPDH;SPTLC2;* *11. ACADVL;SLC27A1;KLHL25;FBXW7; PRMT3;UBR4;ERLIN2;WDTC1;BMP5* *12. CD86;GSK3B;HMGB2;AKR1B1;AHR; UXT;CASP7;RUVBL2;ZC3H12A;PDK4; MED1;UPF1;MEF2C;*	2. Inositol Lipid-Mediated Signaling 3. Sphingolipid Biosynthetic Process 4. Glycerophospholipid Biosynthetic Process 5. Positive Regulation Of Lipid Storage 6. Response To Lipid 7. Negative Regulation Of Lipid Localization 8. Regulation Of Protein Lipidation 9. Sphingolipid Metabolic Process 10. Membrane Lipid Biosynthetic Process 11.Negative Regulation Of Lipid Biosynthetic Process 12. Cellular Response To Lipid
“Fat”	*1. AKT2;ACSL5;IRS2;THBS1*	1. Regulation Of Long-Chain Fatty Acid

	2. HACD1;ACSL1;ACSL5;ACSL4;ELOVL6; ACSL3;HSD17B12;ACSF3 *3. SLC25A1;HACD1;ACSL1;PPT1;ACSL5; ACSL4;ELOVL6;HSD17B12;ACSL3;ACSF3* *4. ABCD2;ACADVL;ECHS1;ETFA; HSD17B10;MCAT;HADHA;BDH2;SLC25A17; CPT2;EHHADH;* *5. HACD1;ACSL1;ACSL5;ACSL4; ELOVL6;FAR1;ACSL3;HSD17B12;ACSF3* *6. ACADVL;KLHL25;NR1H2; UBR4;PDK4;NR1H3;ERLIN2;SIRT2;WDTC1*	Import Across Plasma Membrane 2. Long-Chain fatty-acyl-CoA Biosynthetic Process 3. fatty-acyl-CoA Biosynthetic Process 4. Fatty Acid Beta-Oxidation 5. Long-Chain fatty-acyl-CoA Metabolic Process 6. Regulation Of Fatty Acid Biosynthetic Process

By applying a simple keyword-based filtering approach, the enrichment output was narrowed down to a focused subset of 20 terms, as summarised in Table 2. The search strings used: *"adipose"*, *"ceramide"*, *"cholesterol"*, *"lipid"*, and *"fat"* enabled the identification of enriched categories specifically related to lipid metabolism, fatty acid processing, and adipose tissue development. These findings not only validate the presence of a conserved transcriptional program associated with fundamental adipose tissue functions but also provide a curated gene set that can be used as a reference for future cross-species studies on metabolic regulation. Furthermore, the thematic coherence of the enriched terms many of which are central to lipid handling and adipogenesis reinforces the idea that despite phylogenetic distance, mammals share an evolutionarily conserved molecular framework underpinning adipose physiology(125).

Study Limitations and Future Work

While TissueXplorer offers a user-friendly platform for transcriptomic data analysis in tissue-specific contexts, there are a few limitations that must be acknowledged.

One limitation lies in the gene annotation process. Although one of the preprocessing steps includes a function to convert Ensembl Gene IDs into gene symbols using API queries, this conversion is not always successful. In cases where the API fails to

recognize a specific Gene ID, due to discrepancies or missing entries in the reference database, the result is labelled as "Unknown." To avoid losing information, another function replaces the missing gene symbol with the original Gene ID. However, this solution poses a limitation for comparisons between species, as Gene IDs are species-specific and therefore not directly comparable across organisms.
In addition, there are important limitations associated with the use of gene symbols in comparative analyses between species. Although gene symbols are more legible and are commonly required by downstream tools such as enrichment analysis software, they are not entirely stable or uniquely assigned across species. For example, gene symbol conventions differ across organisms: human gene symbols are typically written in uppercase (*CPS1*), while in mouse they may appear in lowercase (*Cps1*) This inconsistency presented a challenge during development and was mitigated by implementing a Python script that standardized all gene symbols to uppercase (Attachment 3).

Additionally, relying solely on gene symbols may overlook tandem duplications or species-specific duplications, especially when no clear 1 to 1 orthology relationship exists. For future development, tools such as Ensembl Compara could be integrated to provide systematic and reliable orthology mapping across species.
Despite these limitations, gene symbols were chosen in this first version of TissueXplorer for several practical reasons:

- Compatibility with functional enrichment tools: Most enrichment tools require gene symbols as input, streamlining downstream analyses.
- Cross-species comparability: Gene symbols are often conserved across species, facilitating direct comparison in multi-species studies.
- Effective for visualization: Gene symbols are useful for Venn diagrams and other comparative visualizations that illustrate gene overlap between species.
- Low risk of paralogous gene merging in mammals: In mammals, large-scale genome duplication events are relatively rare, reducing the risk of merging paralogous genes. Most exceptions are likely due to tandem duplications, which should be addressed more carefully in future updates.

The integration of more sophisticated orthology mapping algorithms in future iterations of TissueXplorer will enable accurate identification of homologous genes across a broader range of taxa, including phylogenetically distant species. This enhancement will

not only expand the platform's taxonomic coverage but also increase its utility for comparative transcriptomic analyses.

Another limitation of this approach is the reliance on transcriptomic data. Transcriptomes represent a "snapshot" of RNA transcripts expressed in an individual at a specific point in time and under specific conditions. As a result, not all genes may be actively expressed at the time of sampling, leading to the absence of certain genes in the dataset, even if those genes are in fact present in the genome of the species. To mitigate this limitation, we used samples from healthy adult individuals under baseline conditions, aiming to reduce variability and capture a representative transcriptomic profile. Nevertheless, this approach does not fully overcome the issue. For instance, when comparing transcriptomic data from *Macaca mulatta* obtained from two independent sources, Bgee and Expression Atlas in the liver, we observe a high degree of overlap, but not 100% (Figure 17). A total of 9513 genes are shared between both datasets, while 3877 genes are unique to Expression Atlas and 688 are unique to Bgee. This partial overlap illustrates the variability inherent in transcriptomic data derived from different sources, despite similar sample conditions.

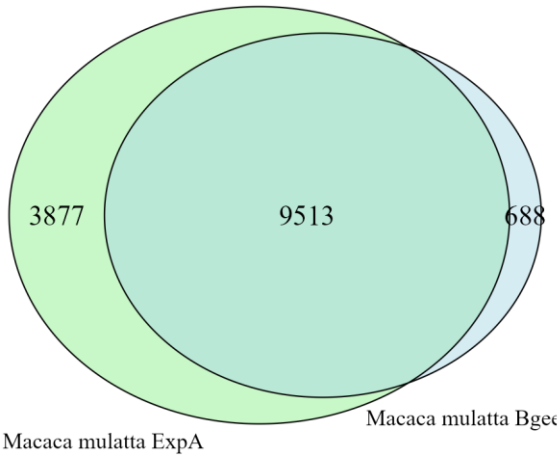

Figure 20. Venn diagram showing the overlap of expressed genes in Macaca mulatta obtained from two transcriptomic databases: Expression Atlas (ExpA) and Bgee.

These limitations underscore the inherent complexities of comparing transcriptomic data between species, tissues and studies. Nonetheless, the current framework provides a strong platform for comparative transcriptomic analysis. By focusing on usability, accessibility, and integration with tools such as enrichment analysis, TissueXplorer offers a valuable starting point for researchers interested in comparative biology and gene expression.

Conclusion

This thesis successfully achieved its primary objectives: (1) the collection, preprocessing, and integration of transcriptomic datasets from multiple mammalian species and tissues, and (2) the design and development of TissueXplorer, a user-friendly, web-based application for comparative transcriptomic analysis.

To address the first objective, a comprehensive pipeline was established to gather and standardize RNA-Seq datasets from a diverse array of mammalian species. Preprocessing steps ensured data comparability, enabling the analysis of gene expression in the same tissues of different species.

The second objective was met through the development of TissueXplorer, which incorporates a suite of features tailored to the needs of the research community:

- **Cross-species comparative analysis:** Users can explore gene expression overlaps and differences across species in homologous tissues through interactive visualizations and downloadable tables.
- **Gene set enrichment analysis:** Built-in enrichment tools help uncover biological patterns and functional associations.
- **User dataset integration:** Researchers can upload their own transcriptomic datasets and perform comparative analyses against curated reference data.
- **Flexible gene search:** Individual genes or custom gene lists can be queried to assess expression across species and tissues.
- **User-friendly interface:** Designed for accessibility, the platform is easy to use, even for those without programming expertise.
- **Efficient data handling:** Optimized to manage moderate-sized datasets with rapid query responses for a seamless user experience.
- **Web-based accessibility:** Available through standard internet browsers, ensuring broad and easy access.
- **Sustainable development:** A well-documented codebase supports long-term maintenance, updates, and collaborative improvements.
- **Data privacy:** Uploaded datasets are not stored or integrated into the central database, protecting user data and ensuring confidentiality.

By integrating these features, TissueXplorer fulfills its role as a powerful and accessible resource for comparative transcriptomic analysis. It lowers technical barriers, supports complex analyses, and enables users to derive meaningful biological insights from multi-species transcriptomic data.

TissueXplorer combines the core functionalities of existing platforms with novel features such as interactive visualizations (e.g., Venn diagrams, heatmaps), support for user-uploaded data, and integrated enrichment analysis. Its utility was demonstrated through case studies that revealed both conserved and species-specific gene expression signatures across mammalian tissues. These examples highlight its value in identifying core biological processes and detecting potential issues in data quality or annotation, underscoring its relevance for research in evolutionary and comparative biology.

Nonetheless, certain limitations must be acknowledged. The reliance on gene symbols, while practical for cross-species comparison and enrichment tools, can introduce challenges. Additionally, transcriptomic data represent only a snapshot of gene expression under specific conditions, and variability between datasets, even when collected under similar circumstances, can affect result interpretation.

Despite these constraints, TissueXplorer effectively accomplishes its primary goal: providing an intuitive, user-friendly platform for comparative transcriptomic analysis that does not require programming expertise. As the volume and diversity of RNA-Seq data continue to grow, accessible tools like TissueXplorer will be essential for enabling researchers from varied backgrounds to derive insights from increasingly complex datasets.

In conclusion, this thesis presents the successful development and validation of TissueXplorer, a versatile, web-based platform for comparative transcriptomic analysis across multiple mammalian species. By combining rigorous data processing pipelines with an accessible, feature-rich interface, TissueXplorer empowers researchers to explore gene expression patterns, conduct enrichment analyses, and draw biological conclusions with ease. Looking ahead, the platform's applicability can be extended to a broader range of species including non-mammalian vertebrates through expanded transcriptomic resources and improved orthology mapping. Ongoing development, guided by user feedback and advances in computational biology, will further enhance TissueXplorer’s analytical power and usability, ensuring its continued value to the genomics research community.

References

1. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. and Shafee, T. (2017) Transcriptomics technologies. *PLoS Comput Biol*, **13**, e1005457.
2. Dong, Z. and Chen, Y. (2013) Transcriptomics: advances and approaches. *Sci China Life Sci*, **56**, 960-967.
3. Casamassimi, A., Federico, A., Rienzo, M., Esposito, S. and Ciccodicola, A. (2017) Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *Int J Mol Sci*, **18**.
4. Robinson, E.L., Baker, A.H., Brittan, M., McCracken, I., Condorelli, G., Emanueli, C., Srivastava, P.K., Gaetano, C., Thum, T., Vanhaverbeke, M. *et al.* (2022) Dissecting the transcriptome in cardiovascular disease. *Cardiovasc Res*, **118**, 1004-1019.
5. Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12**, 87-98.
6. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57-63.
7. He, S.L. and Green, R. (2013) Northern blotting. *Methods Enzymol*, **530**, 75-87.
8. Wery, M., Foretek, D., Andjus, S., Verdys, P. and Morillon, A. (2025) Northern Blotting: Protocols for Radioactive and Nonradioactive Detection of RNA. *Methods Mol Biol*, **2863**, 13-28.
9. Green, M.R. and Sambrook, J. (2022) Analysis of RNA by Northern Blotting. *Cold Spring Harb Protoc*, **2022**.
10. Streit, S., Michalski, C.W., Erkan, M., Kleeff, J. and Friess, H. (2009) Northern blot analysis for detection and quantification of RNA in pancreatic cancer cells and tissues. *Nat Protoc*, **4**, 37-43.
11. Ewis, A.A., Zhelev, Z., Bakalova, R., Fukuoka, S., Shinohara, Y., Ishikawa, M. and Baba, Y. (2005) A history of microarrays in biomedicine. *Expert Rev Mol Diagn*, **5**, 315-328.
12. Bustin, S.A., Benes, V., Nolan, T. and Pfaffl, M.W. (2005) Quantitative real-time RT-PCR--a perspective. *J Mol Endocrinol*, **34**, 597-601.
13. Bustin, S.A. (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol*, **25**, 169-193.
14. Glaves, P.D. and Tugwood, J.D. (2011) Generation and analysis of transcriptomics data. *Methods Mol Biol*, **691**, 167-185.
15. Wang, L. and Li, P.C. (2011) Microfluidic DNA microarray analysis: a review. *Anal Chim Acta*, **687**, 12-27.
16. Stoughton, R.B. (2005) Applications of DNA microarrays in biology. *Annu Rev Biochem*, **74**, 53-82.
17. Ehrenreich, A. (2006) DNA microarray technology for the microbiologist: an overview. *Appl Microbiol Biotechnol*, **73**, 255-273.
18. Heller, M.J. (2002) DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, **4**, 129-153.
19. Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827-836.
20. Call, D.R. (2005) Challenges and opportunities for pathogen detection using DNA microarrays. *Crit Rev Microbiol*, **31**, 91-99.
21. Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*, **22**, 101-109.
22. Sanger, F. and Thompson, E.O. (1953) The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J*, **53**, 353-366.

23. Sanger, F. and Thompson, E.O. (1953) The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem J*, **53**, 366-374.
24. Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.
25. Churko, J.M., Mantalas, G.L., Snyder, M.P. and Wu, J.C. (2013) Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circ Res*, **112**, 1613-1623.
26. Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, **9**, 387-402.
27. Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**, 333-351.
28. Mardis, E.R. (2017) DNA sequencing technologies: 2006-2016. *Nat Protoc*, **12**, 213-218.
29. Finotello, F. and Di Camillo, B. (2015) Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics*, **14**, 130-142.
30. illumina. (2021) In illumina (ed.), *Advancing RNA sequencing, chromatin structure, and DNA methylation studies with the power of Illumina next-generation sequencing*.
31. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621-628.
32. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol*, **17**, 13.
33. Stark, R., Grzelak, M. and Hadfield, J. (2019) RNA sequencing: the teenage years. *Nat Rev Genet*, **20**, 631-656.
34. Thareja, G., Suryawanshi, H., Luo, X. and Muthukumar, T. (2023) Standardization and Interpretation of RNA-sequencing for Transplantation. *Transplantation*, **107**, 2155-2167.
35. Kuksin, M., Morel, D., Aglave, M., Danlos, F.X., Marabelle, A., Zinovyev, A., Gautheret, D. and Verlingue, L. (2021) Applications of single-cell and bulk RNA sequencing in onco-immunology. *Eur J Cancer*, **149**, 193-210.
36. Thind, A., Monga, I., Thakur, P., Kumari, P., Dindhoria, K., Krzak, M., Ranson, M. and Ashford, B. (2021) Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology. *Briefings in bioinformatics*.
37. Thind, A.S., Monga, I., Thakur, P.K., Kumari, P., Dindhoria, K., Krzak, M., Ranson, M. and Ashford, B. (2021) Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology. *Brief Bioinform*, **22**.
38. Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C.M., Zou, F. and Jiang, Y. (2021) SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform*, **22**, 416-427.
39. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P. and De Preter, K. (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*, **11**, 5650.
40. Li, X. and Wang, C.Y. (2021) From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci*, **13**, 36.
41. Oshlack, A., Robinson, M.D. and Young, M.D. (2010) From RNA-seq reads to differential expression results. *Genome Biol*, **11**, 220.

42. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, **6**, 377-382.
43. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*, **6**, 377-382.
44. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M. *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202-1214.
45. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, **8**, 14049.
46. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol*, **21**, 31.
47. Hedlund, E. and Deng, Q. (2018) Single-cell RNA sequencing: Technical advancements and biological applications. *Mol Aspects Med*, **59**, 36-46.
48. Zhang, Y., Wang, D., Peng, M., Tang, L., Ouyang, J., Xiong, F., Guo, C., Tang, Y., Zhou, Y., Liao, Q. *et al.* (2021) Single-cell RNA sequencing in cancer research. *J Exp Clin Cancer Res*, **40**, 81.
49. Samad, T. and Wu, S.M. (2021) Single cell RNA sequencing approaches to cardiac development and congenital heart disease. *Semin Cell Dev Biol*, **118**, 129-135.
50. Hu, Y., Zhang, Y., Liu, Y., Gao, Y., San, T., Li, X., Song, S., Yan, B. and Zhao, Z. (2022) Advances in application of single-cell RNA sequencing in cardiovascular research. *Front Cardiovasc Med*, **9**, 905151.
51. Liu, S. and Trapnell, C. (2016) Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res*, **5**.
52. Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M. and Tombácz, D. (2019) Long-Read Sequencing - A Powerful Tool in Viral Transcriptome Research. *Trends Microbiol*, **27**, 578-592.
53. Uapinyoying, P., Goecks, J., Knoblach, S., Panchapakesan, K., Bonnemann, C., Partridge, T., Jaiswal, J. and Hoffman, E. (2020) A long-read RNA-seq approach to identify novel transcripts of very large genes. *Genome Research*, **30**, 885-897.
54. Cho, H., Davis, J., Li, X., Smith, K.S., Battle, A. and Montgomery, S.B. (2014) High-resolution transcriptome analysis with long-read RNA sequencing. *PLoS One*, **9**, e108095.
55. Oikonomopoulos, S., Bayega, A., Fahiminiya, S., Djambazian, H., Berube, P. and Ragoussis, J. (2020) Methodologies for Transcript Profiling Using Long-Read Technologies. *Front Genet*, **11**, 606.
56. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, **21**, 30.
57. Tedersoo, L., Albertsen, M., Anslan, S. and Callahan, B. (2021) Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. *Appl Environ Microbiol*, **87**, e0062621.
58. Shumate, A., Wong, B., Pertea, G. and Pertea, M. (2022) Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput Biol*, **18**, e1009730.

59. Kainth, A.S., Haddad, G.A., Hall, J.M. and Ruthenburg, A.J. (2023) Merging short and stranded long reads improves transcript assembly. *PLoS Comput Biol*, **19**, e1011576.
60. He, Q., Liu, Y. and Sun, W. (2018) Statistical analysis of non-coding RNA data. *Cancer Lett*, **417**, 161-167.
61. Dindhoria, K., Monga, I. and Thind, A.S. (2022) Computational approaches and challenges for identification and annotation of non-coding RNAs using RNA-Seq. *Funct Integr Genomics*, **22**, 1105-1112.
62. Beermann, J., Piccoli, M.T., Viereck, J. and Thum, T. (2016) Non-coding RNAs in Development and Disease: Background, Mechanisms, and Therapeutic Approaches. *Physiol Rev*, **96**, 1297-1325.
63. Ye, B., Shi, J., Kang, H., Oyebamiji, O., Hill, D., Yu, H., Ness, S., Ye, F., Ping, J., He, J. *et al.* (2020) Advancing Pan-cancer Gene Expression Survial Analysis by Inclusion of Non-coding RNA. *RNA Biol*, **17**, 1666-1673.
64. Bussotti, G., Notredame, C. and Enright, A.J. (2013) Detecting and comparing non-coding RNAs in the high-throughput era. *Int J Mol Sci*, **14**, 15423-15458.
65. Evans, C., Hardin, J. and Stoebel, D.M. (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform*, **19**, 776-792.
66. Babarinde, I.A. and Hutchins, A.P. (2022) The effects of sequencing depth on the assembly of coding and noncoding transcripts in the human genome. *BMC Genomics*, **23**, 487.
67. Zhang, M.J., Ntranos, V. and Tse, D. (2020) Determining sequencing depth in a single-cell RNA-seq experiment. *Nat Commun*, **11**, 774.
68. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
69. Zhao, Y., Li, M.C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H. and McShane, L.M. (2021) TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *J Transl Med*, **19**, 269.
70. Wagner, G.P., Kin, K. and Lynch, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*, **131**, 281-285.
71. Zhao, S., Ye, Z. and Stanton, R. (2020) Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *Rna*, **26**, 903-909.
72. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
73. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, **14**, 417-419.
74. Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, **34**, 525-527.
75. Bastian, Frederic B., Cammarata, Alessandro B., Carsanaro, S., Detering, H., Huang, W.-T., Joye, S., Niknejad, A., Nyamari, M., Mendes de Farias, T., Moretti, S. *et al.* (2024) Bgee in 2024: focus on curated single-cell RNA-seq datasets, and query tools. *Nucleic Acids Research*, **53**, D878-D885.
76. Bastian, F.B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, Sara S., de Farias, T.M., Moretti, S., Parmentier, G., de Laval, V.R., Rosikiewicz, M. *et al.* (2020) The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research*, **49**, D831-D847.

77. Moreno, P., Fexova, S., George, N., Manning, J.R., Miao, Z., Mohammed, S., Muñoz-Pomer, A., Fullgrabe, A., Bi, Y., Bush, N. *et al.* (2021) Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Research*, **50**, D129-D140.
78. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, **45**, 580-585.
79. Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S. *et al.* (2019) Gene expression across mammalian organ development. *Nature*, **571**, 505-509.
80. Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y. *et al.* (2017) An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res*, **27**, 1759-1768.
81. Sudmant, P.H., Alexis, M.S. and Burge, C.B. (2015) Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol*, **16**, 287.
82. Zheng-Bradley, X., Rung, J., Parkinson, H. and Brazma, A. (2010) Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol*, **11**, R124.
83. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343-348.
84. Merkin, J., Russell, C., Chen, P. and Burge, C.B. (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593-1599.
85. Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R. *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587-1593.
86. Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2016) A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, **18**, 205-214.
87. Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigele, S., Do, H.H., Weiss, G., Enard, W. *et al.* (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*, **14**, 1462-1473.
88. Oldham, M.C., Horvath, S. and Geschwind, D.H. (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A*, **103**, 17973-17978.
89. Zeng, H., Shen, E.H., Hohmann, J.G., Oh, S.W., Bernard, A., Royall, J.J., Glattfelder, K.J., Sunkin, S.M., Morris, J.A., Guillozet-Bongaarts, A.L. *et al.* (2012) Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*, **149**, 483-496.
90. Castrillon, J. and Bengtson Nash, S. (2020) Evaluating cetacean body condition; a review of traditional approaches and new developments. *Ecol Evol*, **10**, 6144-6162.
91. Li, G., Wei, H., Bi, J., Ding, X., Li, L., Xu, S., Yang, G. and Ren, W. (2020) Insights into Dietary Switch in Cetaceans: Evidence from Molecular Evolution of Proteinases and Lipases. *J Mol Evol*, **88**, 521-535.
92. Senevirathna, J.D.M. and Asakawa, S. (2021) Multi-Omics Approaches and Radiation on Lipid Metabolism in Toothed Whales. *Life (Basel)*, **11**.
93. Tang, C.H., Lin, C.Y., Tsai, Y.L., Lee, S.H. and Wang, W.H. (2018) Lipidomics as a diagnostic tool of the metabolic and physiological state of managed whales: A correlation study of systemic metabolism. *Zoo Biol*, **37**, 440-451.

94. Bukhman, Y.V., Morin, P.A., Meyer, S., Chu, L.-F., Jacobsen, J.K., Antosiewicz-Bourget, J., Mamott, D., Gonzales, M., Argus, C., Bolin, J. *et al.* (2024) A High-Quality Blue Whale Genome, Segmental Duplications, and Historical Demography. *Molecular Biology and Evolution*, **41**.
95. Suzuki, M., Funasaka, N., Yoshimura, K., Inamori, D., Watanabe, Y., Ozaki, M., Hosono, M., Shindo, H., Kawamura, K., Tatsukawa, T. *et al.* (2024) Comprehensive expression analysis of hormone-like substances in the subcutaneous adipose tissue of the common bottlenose dolphin Tursiops truncatus. *Scientific Reports*, **14**, 12515.
96. CCaro, Cetacean Anatomy: Adaptations for Aquatic Life. https://www.ccaro.org/c-adaptations-for-aquatic-life.php
97. Foundation, T.P.S. (2001). Python.
98. Lutz, M. *Learning Python, 5th Edition*. O'Reilly Media, Inc.
99. Foundation, T.R. R Foundation.
100. Naranbhai, V., Fletcher, H.A., Tanner, R., O'Shea, M.K., McShane, H., Fairfax, B.P., Knight, J.C. and Hill, A.V. (2015) Distinct Transcriptional and Anti-Mycobacterial Profiles of Peripheral Blood Monocytes Dependent on the Ratio of Monocytes: Lymphocytes. *EBioMedicine*, **2**, 1619-1626.
101. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
102. Habuka, M., Fagerberg, L., Hallström, B.M., Pontén, F., Yamamoto, T. and Uhlen, M. (2015) The Urinary Bladder Transcriptome and Proteome Defined by Transcriptomics and Antibody-Based Profiling. *PLoS One*, **10**, e0145301.
103. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343-348.
104. Yu, Y., Fuscoe, J.C., Zhao, C., Guo, C., Jia, M., Qing, T., Bannon, D.I., Lancashire, L., Bao, W., Du, T. *et al.* (2014) A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nat Commun*, **5**, 3230.
105. Pipes, L., Li, S., Bozinoski, M., Palermo, R., Peng, X., Blood, P., Kelly, S., Weiss, J.M., Thierry-Mieg, J., Thierry-Mieg, D. *et al.* (2013) The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res*, **41**, D906-914.
106. Fushan, A.A., Turanov, A.A., Lee, S.G., Kim, E.B., Lobanov, A.V., Yim, S.H., Buffenstein, R., Lee, S.R., Chang, K.T., Rhee, H. *et al.* (2015) Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, **14**, 352-365.
107. Chen, J.Y., Peng, Z., Zhang, R., Yang, X.Z., Tan, B.C., Fang, H., Liu, C.J., Shi, M., Ye, Z.Q., Zhang, Y.E. *et al.* (2014) RNA editome in rhesus macaque shaped by purifying selection. *PLoS Genet*, **10**, e1004274.
108. Burns, E.N., Bordbari, M.H., Mienaltowski, M.J., Affolter, V.K., Barro, M.V., Gianino, F., Gianino, G., Giulotto, E., Kalbfleisch, T.S., Katzman, S.A. *et al.* (2018) Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Anim Genet*, **49**, 564-570.
109. Jin, L., Tang, Q., Hu, S., Chen, Z., Zhou, X., Zeng, B., Wang, Y., He, M., Li, Y., Gui, L. *et al.* (2021) A pig BodyMap transcriptome reveals diverse tissue physiologies and evolutionary dynamics of transcription. *Nat Commun*, **12**, 3715.
110. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27-30.
111. Agrawal, A., Balcı, H., Hanspers, K., Coort, S.L., Martens, M., Slenter, D.N., Ehrhart, F., Digles, D., Waagmeester, A., Wassink, I. *et al.* (2023)

WikiPathways 2024: next generation pathway database. *Nucleic Acids Research*, **52**, D679-D689.

112. Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B. *et al.* (2023) The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, **52**, D672-D678.
113. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R. and Ma'ayan, A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.
114. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, **44**, W90-97.
115. Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M. *et al.* (2021) Gene Set Knowledge Discovery with Enrichr. *Curr Protoc*, **1**, e90.
116. Meadows, J.R.S., Kidd, J.M., Wang, G.-D., Parker, H.G., Schall, P.Z., Bianchi, M., Christmas, M.J., Bougiouri, K., Buckley, R.M., Hitte, C. *et al.* (2023) Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture. *Genome Biology*, **24**, 187.
117. Wang, G.-D., Shao, X.-J., Bai, B., Wang, J., Wang, X., Cao, X., Liu, Y.-H., Wang, X., Yin, T.-T., Zhang, S.-J. *et al.* (2018) Structural variation during dog domestication: insights from gray wolf and dhole genomes. *National Science Review*, **6**, 110-122.
118. Villar, D., Berthelot, C., Aldridge, S., Rayner, Tim F., Lukk, M., Pignatelli, M., Park, Thomas J., Deaville, R., Erichsen, Jonathan T., Jasinska, Anna J. *et al.* (2015) Enhancer Evolution across 20 Mammalian Species. *Cell*, **160**, 554-566.
119. Perry, G.H., Melsted, P., Marioni, J.C., Wang, Y., Bainer, R., Pickrell, J.K., Michelini, K., Zehr, S., Yoder, A.D., Stephens, M. *et al.* (2012) Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res*, **22**, 602-610.
120. Yao, Y., Liu, S., Xia, C., Gao, Y., Pan, Z., Canela-Xandri, O., Khamseh, A., Rawlik, K., Wang, S., Li, B. *et al.* (2022) Comparative transcriptome in large-scale human and cattle populations. *Genome Biology*, **23**, 176.
121. Jebb, D. and Hiller, M. (2018) Recurrent loss of HMGCS2 shows that ketogenesis is not essential for the evolution of large mammalian brains. *eLife*, **7**, e38906.
122. Figueiró, H.V., Li, G., Trindade, F.J., Assis, J., Pais, F., Fernandes, G., Santos, S.H.D., Hughes, G.M., Komissarov, A., Antunes, A. *et al.* (2017) Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Science Advances*, **3**, e1700299.
123. Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H., Medrano, J.F., Van Eenennaam, A.L., Ernst, C., Ross, P. *et al.* (2018) Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics*, **19**, 684.
124. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57-63.
125. Liu, P., Li, D., Zhang, J., He, M., Li, Y., Liu, R. and Li, M. (2023) Transcriptomic and lipidomic profiling of subcutaneous and visceral adipose tissues in 15 vertebrates. *Scientific Data*, **10**, 453.

Attachment 1

Function to standardize the file formats

```
# Function to preprocess BGee dataframes
def preprocess_BGee():
    # For all the species in the Bgee folder
    for species in os.listdir("..//Datasets/Bgee"):
        species_path = os.path.join("..//Datasets/Bgee", species)
        for file in os.listdir(species_path):
            # Takes all .tsv files, which is the data from Bgee
            if file.endswith(".tsv"):
                file_path = os.path.join(species_path, file)
                # Read .tsv file and transform into a pandas dataframe
                df = pd.read_csv(file_path, sep="\t")
                # Eliminate unecessary collumns
                df = df.drop(
                    columns=[
                        "Experiment ID",
                        "Library ID",
                        "Library type",
                        "Anatomical entity ID",
                        "Stage ID",
                        "Stage name",
                        "Sex",
                        "Strain",
                        "Read count",
                        "Rank",
                        "Detection flag",
                        "pValue",
                        "State in Bgee",
                    ]
                )

                # Creates a new dataframe called df_pivot where each tissue in the collumn "Anatomical
entity name" of the df dataframe becomes a new collumn filled by the corresponding "TPM" values.
                df_pivot = df.pivot(columns="Anatomical entity name", values="TPM")
                # Joins the two dataframes in a new one called df_final
                df_final = pd.concat([df, df_pivot], axis=1)
                # Removes "Anatomical entity name" and "TPM" collumns
                df_final = df_final.drop(columns=["Anatomical entity name", "TPM"])
                # PREPROCESSING
                # Substituting missing values for 0
                df_final.fillna(0, inplace=True)
                # Removing lines where all collumns except "Gene ID" are 0
                df_final = df_final.loc[
                    ~(df_final.drop(columns=["Gene ID"], errors="ignore") == 0).all(
                        axis=1
                    )
                ]
                # Adding a collumn called "GeneSymbol"
                df_final.insert(1, "GeneSymbol", None)
                # Saving df_final in a .csv file
                output_file = os.path.join(
                    species_path, f"PP_Bgee_{file.replace('.tsv', '.csv')}"
                )
                df_final.to_csv(output_file, index=False)
                print(f"Complete for {species}")
```

```
# Function to preprocess Expression Atlas dataframes
def preprocess():
    # For all the species in the Expression Atlas folder
    for species in os.listdir("..//Datasets/Expression Atlas"):
        species_path = os.path.join("..//Datasets/Expression Atlas", species)
        for file in os.listdir(species_path):
            # Takes all .tsv files, which is the data from Expression Atlas
            if file.endswith(".tsv"):
                file_path = os.path.join(species_path, file)
                # Read .tsv file and transform into a pandas dataframe
                df = pd.read_csv(file_path, sep="\t", header=4)
                # Substite missing values with 0
                df.fillna(0, inplace=True)
                # Change the name of the collumn Gene Name to "GeneSymbol"
                df.rename(columns={"Gene Name": "GeneSymbol"}, inplace=True)
                # Delete geneIDs in the Gene Symbol collumn
                df.loc[
                    df["GeneSymbol"].astype(str).str.contains("00000"),
"GeneSymbol"    ] = np.nan
                # Save in a csv file
                output_file = os.path.join(
                    species_path,
                    f"PP_ExpAtlas_{species}_{file.replace('.tsv', '.csv')}",
                )
                df.to_csv(output_file, index=False)
                print(f"Complete for {species}")
```

Attachment 2

Functions to transform gene IDS into gene symbols

```
# API to transform gene IDs into gene symbols
BATCH_SIZE = 100
MAX_WORKERS = 10  # Number of simultaneous calls to the API


def get_gene_symbol(ensembl_id):
    "Calls API Ensembl to obtain the Gene Symbol of a single Gene ID."
    url = (
        f"https://rest.ensembl.org/lookup/id/{ensembl_id}?content-
type=application/json"
    response = requests.get(url, timeout=120)
    if response.status_code == 200:
        data = response.json()
        return ensembl_id, data.get("display_name", "Unknown")
    else:
        return ensembl_id, "Not Found"


def get_gene_symbol_batch(ensembl_ids):
    "Executes múltiple API calls in paralell using ThreadPoolExecutor."
    gene_symbols = {}
    with ThreadPoolExecutor(max_workers=MAX_WORKERS) as executor:
        results = executor.map(get_gene_symbol, ensembl_ids)
    for gene_id, symbol in results:
        gene_symbols[gene_id] = symbol
    return gene_symbols


def add_gene_names(file_path, species):
    df = pd.read_csv(file_path)
    tqdm.pandas()
    # Filters lines where GeneSymbol is empty or Na
    missing_mask = df["GeneSymbol"].isna() | (
        df["GeneSymbol"].astype(str).str.strip() == ""
    )
    missing_symbols = df[missing_mask]

    # Removes duplicates before calling the API
    unique_gene_ids = missing_symbols["Gene ID"].drop_duplicates().tolist()
    gene_symbols = {}  # Dictionary to store unique results
    # Processing in batches
    for i in tqdm(
        range(0, len(unique_gene_ids), BATCH_SIZE), desc="Processing in batches"
    ):
        batch_ids = unique_gene_ids[i : i + BATCH_SIZE]
        batch_results = get_gene_symbol_batch(batch_ids)
        gene_symbols.update(batch_results)  # Save in cache

    df["GeneSymbol"] = df["GeneSymbol"].astype("object")

    # Substituting only the gene symbols that we did not know
    df.loc[missing_mask, "GeneSymbol"] = df.loc[missing_mask, "Gene ID"].map(
        gene_symbols
    )
    df.to_csv(f"{species}_with_gene_symbols.csv", index=False)
```

Attachment 3

Last preprocessing function to transform “Unknown” in Gene IDs and set the gene symbols to uppercase

```
# Transforming "Unknown" in GeneSymbol collumn in Gene IDs and delete the collumn Gene ID
def back_geneid():
    # Obtains the list of files in the current folder that end with "_with_gene_symbols.csv"
    files = [f for f in os.listdir() if f.endswith("_with_gene_symbols.csv")]

    for file in files:
        df = pd.read_csv(file)

        if "GeneSymbol" in df.columns and "Gene ID" in df.columns:
            # Replace "Unknown" or "Not Found" with Gene ID
            mask = df["GeneSymbol"].isin(["Unknown", "Not Found"])
            df.loc[mask, "GeneSymbol"] = df.loc[mask, "Gene ID"]

            # Drop the 'Gene ID' column
            df.drop(columns=["Gene ID"], inplace=True)
            # Set GeneSymbols to uppercase
            df["GeneSymbol"] = df["GeneSymbol"].str.upper()
            # Naming and saving final .csv file
            new_filename = f"final_{file}"
            df.to_csv(new_filename, index=False)
            print(f"File saved: {new_filename}")
        else:
            print(f"Error: {file} Does not contain the necessary collumns")
```