Modeling and Implementation of a Web Database for RNA-Seq of Bovine Embryonic Cells

Natalia Soriani Daleffi[1], **Marcella Pecora Milazzotto**[2],
Fernanda Nascimento Almeida[1*]

[1]Bioinformatic and Health Informatic Group – BIHG
Center for Engineering, Modeling and Applied Social Science
Federal University of ABC (UFABC) – São Bernardo do Campo, SP – Brazil

[2]Laboratory of Embryonic Metabolism and Epigenetics
Center of Natural and Human Science
Federal University of ABC (UFABC) – Santo André, SP – Brazil

*fernanda.almeida@ufabc.edu.br

Abstract. *This paper aims to develop a dedicated RNA-Seq database for bovine embryos, generated to gain insights into reproductive metabolism. The data is categorized into three groups, each obtained from distinct experiments. The primary objective is to streamline data analysis through a platform, named TranscriptomicsSeqDB, which standardizes and organizes RNA-Seq information from the Laboratory of Embryonic Metabolism and Epigenetics at UFABC, São Paulo, Brazil. Apart from data storage and management, TranscriptomicsSeqDB provides a user-friendly search interface with predefined queries to facilitate gene-specific indicator analysis.*

1. Introduction

Genetics is indispensable for understanding life. Every day, new correlations between genes and diseases, habits, and even personality traits are being discovered. They are crucial for both diagnosis and the development of new medicines and vaccines. Genetic sequencing has gained momentum in recent years, especially with next-generation sequencing (NGS) techniques. NGS allows for the rapid, automated, and high-throughput sequencing of all nitrogenous bases in the genetic material. This enables complete genetic sequencing, resulting in a vast amount of data. Therefore, there is a need to structure this data in a way that allows for practical analysis of the obtained results. One option is to organize it into databases, which are the most efficient means of accessing and analyzing biological data.

In light of this scenario, the Laboratory of Embryonic Metabolism and Epigenetics at UFABC conducted a study focused on RNA sequencing of bovine embryonic cells. The RNA-Seq method was utilized, enabling a detailed understanding of the cellular transcriptome and the metabolic processes involved in bovine reproduction from the material sequenced through NGS. The study involves RNA sequencing divided into three groups, each subjected to different tests. Each group consists of three subgroups: a control group with no manipulation, and two other groups in which the genetic material was exposed to different substances affecting metabolism. While these data were stored in spreadsheets, this type of sequencing generates a vast amount of data, making it challenging to access and analyze in its current format.

The objective of this work will be to aggregate the data generated from the genetic sequencing of bovine embryonic cells using the RNA-Seq method at the Laboratory of Embryonic Metabolism and Epigenetics of UFABC into a database. It will be necessary to standardize the available information and provide analysis tools. In addition to storing and managing the data, a web-based access platform will be developed, offering a common search interface for databases, using predefined queries.

2. RNA-Seq

One of the major focuses of genetic research is to understand gene expression, i.e., the relationship between DNA and protein production. This line of study revolves around the transcription process and RNA, which can contribute to understanding gene expression through the quantification of gene alterations during each transcription. With the aim of characterizing the transcriptome, various sequencing methods have been explored over time, such as microarray hybridization and Sanger sequencing. Both approaches sequenced complementary DNA (cDNA) but not in a comprehensive manner. Their limitations include low specificity and sensitivity for certain genes [Hrdlicková et al. 2016]. In response to the demand for faster and more complete sequencing, next-generation sequencing (NGS) emerged. NGS revolutionized the study of genetic material by providing a large amount of data in a short period.

Within NGS technology, there are several methods for genetic sequencing, including RNA Sequencing (RNA-Seq). RNA-Seq is an excellent method for mapping and quantifying transcriptomes to analyze gene expression in different tissues [Cánovas et al. 2010]. RNA-Seq has the advantage of sequencing a larger number of bases more rapidly and accurately. As a result, RNA-Seq has become widely used, being an indispensable method for studies involving gene expression, aspects of RNA biogenesis and metabolism [Hrdlicková et al. 2016].

Different types of results are analyzed using this type of sequencing. The primary result is the reading and quantity of reads, which are fragments of DNA or RNA sequences. RNA-Seq generates a large volume of reads. This volume is relevant because the reliability of the data is related to the number of independent sequences read. RPKM (Reads per kilobase per million mapped reads) is an essential metric to qualify transcribed genes. This measure reflects the molar concentration of a transcript in the sample by normalizing the RNA strand size and the total number of reads, facilitating the comparison between genes in the same sample [Mortazavi et al. 2008].

The methodology of cDNA sequencing by RNA-Seq has various biological applications, as it is an important research tool for detecting genetic elements in different types of cells and species. Due to its application in SNP (Single Nucleotide Polymorphisms) discovery and the description of transcript density, embryo sequencing is of great importance in describing the early stages of development [Chitwood et al. 2013]. Embryonic development is dynamic and influenced by various factors, including the speed of development according to published data [Desai et al. 2014]. Therefore, the morphokinetics of the early divisions of the embryonic cell can determine the overall development capacity of the embryo, but there is limited information regarding the relationship between embryonic morphokinetic characteristics and genetic physiology [Milazzotto et al. 2016].

3. Biological databases

The foundation of all applied bioinformatics studies lies in the data obtained through sequencing, which is produced in large volume in this type of work. Data are pieces of information that can be managed in various ways, one of which is storing them in a database. This approach has several advantages, such as speed, data protection, and up-to-date information. There are various types of architectures for modeling a database, with the most commonly used being the relational model. It can be applied in cases where data has some form of relationship. In a relational database, information is stored in tables and queried through Structured Query Language (SQL), which is a database query language.

A biological database shares several similarities in its construction with any other type of database [Ullah et al. 2022]. For biologists, geneticists, and professionals involved in genetic studies, biological databases are essential for disseminating novelties in the field and accessing new sequencing data [Danchin et al. 2018]. The main biological databases used include GenBank, UCSC Genome Browser, and Ensembl [Baxevanis and Bateman 2015]. These databases contain biological data from various organisms and types of sequences. According to the National Center for Biotechnology Information (NCBI), there are two main types of biological databases: comprehensive and specialized. The comprehensive ones encompass the databases mentioned earlier; they contain abundant data on various topics and are widely disseminated and used. On the other hand, specialized databases contain data specific to particular organisms and sequencing types [Villalba and Matte 2021].

Biological databases face some unique challenges that are not as common in other applications. The first is the constant discovery of new relationships among genetic data, leading to potential changes in data interpretation and consequently in the database structure. Another issue is the scarcity of professionals who can model a database considering both the biological and programmatic aspects.

Hence, one advantage of specialized biological databases is the greater flexibility in modeling and curating data. In recent years, several databases of this type have emerged and gained momentum due to specific motivations, either focusing on studying a particular biological problem or better serving a segment of the biological community. In this regard, computational analysis plays a pivotal role in interpreting the data derived from RNA-Seq technique, proving indispensable in addressing significant biological queries [Deshpande et al. 2023]. In this context, several databases and computational tools have been developed with the aim of investigating this kind of information and also establishing connections with other study modalities, even extending to the utilization of RNA-Seq data for characterizing both individual adaptive immune repertoire and microbiome [Deshpande et al. 2023].

A prominent example is RNA CoMPASS, a database focused on providing insights into gene expression of pathogens and hosts. It concurrently analyzes transcriptome and metatranscriptome, enhancing the understanding of intricate interactions [Xu et al. 2014]. Additionally, the BEAVR tool stands out for simplifying the visualization of RNA-Seq data and allowing local execution [Perampalam and Dick 2020]. The GTEx reference platform also leverages RNA-Seq data, aiming to correlate genetic variations with changes in gene expression in human individuals [Lonsdale et al. 2013]. The evolution of computational tools within the realm of RNA-Seq sequencing showcases a promising landscape for research and hypothesis development. Some of these tools broaden the scope of investigation, as seen with RNA CoMPASS, while others, such as BEAVR, play a crucial role in streamlining researcher interaction with various programming languages. Furthermore, there are those that establish interdisciplinary connections, like GTEx, by linking RNA-Seq data to other sequencing modalities.

4. Data Modeling

The data for this work was generated based on the division into groups, studying the behavior of each bovine RNA gene in three different samples when subjected to metabolism-altering substances. These samples were sequenced using the RNA-Seq method. All results were stored in a spreadsheet. The generated file contains 24,616 rows and 37 columns. In the columns, there are two main types of data: information about the genes (name, type, ID, location, among others) and the results of each of the samples. The indicators considered in this study are the RPKM and the total number of reads. There are 28 columns containing the results, which are divided by indicator, sample (1, 2, or 3), development speed (fast, slow, or in vivo), and cell type (embryo or blastocyst). Considering that the main objective is to compare the results between samples and within each sample, the first step was to manipulate this data to facilitate such comparisons in an easier and more efficient way. Therefore, from the raw data, which was all in a single spreadsheet, two separate spreadsheets containing this data were created. The *gene_data* spreadsheet contains the descriptive gene data. It has 14 columns, with the first being a unique identifier for each row (column *id*). The second spreadsheet created was the *sample_data*, which contains the indicator data. Instead of creating a column for each combination of sample, speed, and cell type, as structured in the main spreadsheet, only three columns (*sample*, *speed*, and *type*) were created, and the combinations with the corresponding gene ID were listed. After this structuring, the data was exported to the CSV format to be later inserted into database tables.

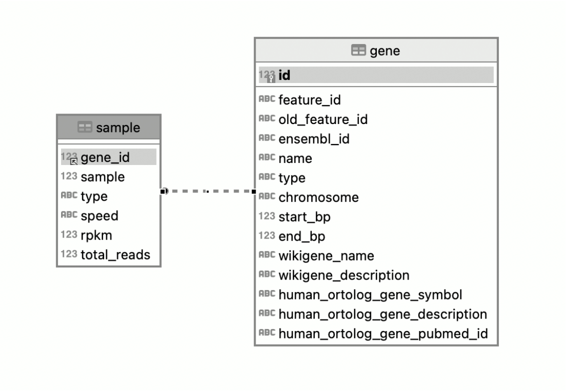
4.1. Database implementation

The database management system used was MariaDB (https://mariadb.org/). For the manipulation of the database and tables, both the terminal and the DBeaver software (https://dbeaver.io) were used, which assists in database administration. The model chosen for the biological database is the relational model. Therefore, the step before creating the database itself was to structure the data and understand the relationship between them. The main relationship is based on genes; thus, two different tables were created and related through the *id* field to handle the data.

Firstly, the database *rnaseq_data* was created. The next step was to create the tables. For the *gene* table, containing the data from the *gene_data* spreadsheet, the *id* field (a unique identifier for each record in the table) was set as the *primary key*, which indicates that this field must have only unique values. In the *sample* table, containing the information from the *sample_data* spreadsheet, the *gene_id* field was set as a *foreign key*, indicating that this column will relate to other tables in the database. After these definitions and the determination of the data types for each column, the *rnaseq_data* database was created locally (Figure 1).

By setting up the database and tables, the next step will be to populate the tables with the data from the CSV files generated during the data manipulation step. This will involve inserting the data into the corresponding tables and ensuring that the relationships between the *gene* and *sample* tables are maintained correctly.

Figure 1. Architectural overview of the database structure. The database is currently undergoing expansion to accommodate growing needs.



5. Web platform

As previously mentioned, the web platform aims to provide data from RNA-Seq research in an automated and graphical manner, comparing the results obtained from *RPKM* and total read count. The web interface was developed based on the "EmbryoGene Profiler" work (https://emb-bioinfo.fsaa.ulaval.ca/IMAGE/index.html). The HTML was designed so that all pages contain a main menu with the project name (TranscriptomicsSeqDB) and links to the four existing pages. The CSS, which determines the visual aesthetics of the pages, was created using existing fonts available online (https://templatemo.com/tm-516-known). The connection to the local database was established with the help of phpMyAdmin (https://www.phpmyadmin.net/), a software used to manage data in relational databases. With the database already created, the integration of the *rnaseq_data* database with the platform was performed. In parallel, the development was carried out using a Python3 script to integrate the platform data with the HTML. The pymysql library (https://pypi.org/project/pymysql/) was used for the database connection in the script, while the Flask library (https://flask.palletsprojects.com/en/2.3.x/) was selected for connecting the script with the HTML. Through these libraries, the successful connection of the database with the HTML was achieved.

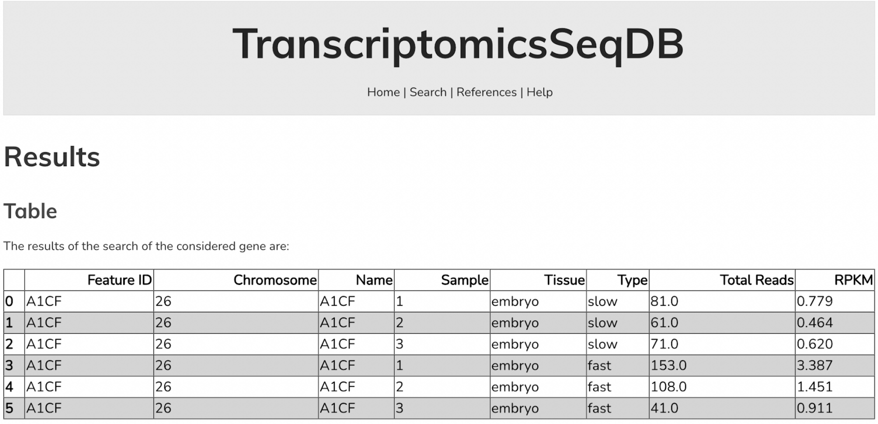
5.1. Search page

In the TranscriptomicsSeqDB web platform, the main goal is to provide results based on specific queries. To achieve this, the *Search* page includes fields in the HTML where users can input the necessary filters for their queries, such as *feature_id* and *tissue* (equivalent to the *type* field in the database). In the backend *script*, these fields are used to filter the data with a pre-defined SQL query. The query joins data from both tables using an *inner join* between the *id* field in the *gene* table and the *gene_id* field in the *sample* table. The returned data is at the sample level (by type and speed), providing the main information regarding gene features and expression. The filters for *feature_id* and *tissue* in the query are exact, meaning that if the user enters a non-existent *feature_id*, no results will be returned. To address this, a redirect to the *Help* page is implemented, where users can download a spreadsheet containing an updated list of all *feature_ids* along with indications of whether there is an *RPKM* greater than zero in any of the samples. The results page presents the retrieved data in two types of visualizations: a table containing all the data and a graph. The graph is generated using the plotly library (https://plotly.com/), with the X-axis representing the sample speed and the Y-axis showing the RPKM values. The graph calculates the average RPKM across all samples, as well as the standard deviation.

6. Results and discussion

The results page provides two distinct visualization views. The table (Figure 2) presents the data in the same structure as stored in the database. Its primary objective is to offer a detailed view per sample, enabling a comprehensive understanding of RPKM results for each tested sample. This granular perspective is essential for a thorough analysis of specific gene expressions in different morphokinetic groups.

Figure 2. Table in results page when filtered *2Cells* (*embryo*) and *feature_id* A1CF.



TranscriptomicsSeqDB

Home | Search | References | Help

Results

Table

The results of the search of the considered gene are:

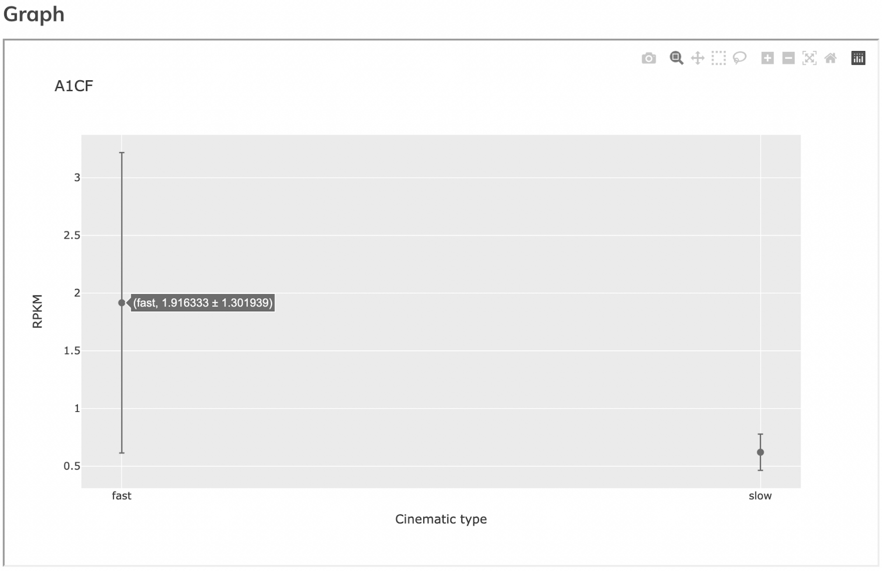
	Feature ID	Chromosome	Name	Sample	Tissue	Type	Total Reads	RPKM
0	A1CF	26	A1CF	1	embryo	slow	81.0	0.779
1	A1CF	26	A1CF	2	embryo	slow	61.0	0.464
2	A1CF	26	A1CF	3	embryo	slow	71.0	0.620
3	A1CF	26	A1CF	1	embryo	fast	153.0	3.387
4	A1CF	26	A1CF	2	embryo	fast	108.0	1.451
5	A1CF	26	A1CF	3	embryo	fast	41.0	0.911

The graph is located below the table and provides an analysis view categorized by type (*slow* or *fast*). This allows for a straightforward comparative analysis. As previously mentioned, the graph calculates the average RPKM for all groups, as well as the standard deviation. Users have the option to download the graph in PNG format, and the image is interactive, enabling zooming in or out. Additionally, users can view the values of the mean and standard deviation for each type of sample (Figure 3). This interactive and visual representation facilitates a comprehensive understanding of the variation in gene expression between different morphokinetic groups.

6.1. Data analysis

The database that feeds the web platform contains a total of 24,615 records in the *gene* table. Among these, 21,454 (87%) are protein-coding genes, while the remaining 13% are non-coding genes. Out of the total, 21,308 genes (87%) have a record of their human ortholog. Regarding their location, there is a proportional distribution of genes across all chromosomes. In the *sample* table, there are 344,610 records. These records are divided in such a way that 196,920 (57%) are related to tests performed on blastocysts, and 147,690 (43%) on embryos (*2Cells*). For blastocysts, sequencing was carried out on samples treated with substances to increase and decrease metabolic speed, as well as

Figure 3. Graph in results page when filtered *2Cells* (*embryo*) and *feature_id* A1CF.



samples without any substances. For the rapid and slow speed groups, three tests were conducted, resulting in 73,845 records each (with all 24,615 genes sequenced in each test). For the group without any substances (*in vivo*), two tests were performed, resulting in a total of 49,230 records. For embryos (*2Cells*), tests were conducted to increase and decrease the metabolic speed. Three tests were performed for each group, resulting in a total of 73,845 records per group.

7. Conclusions

TranscriptomicsDB emerges as a valuable resource tailored to meet the specific demands of the laboratory's RNA-Seq experiments conducted at the Laboratory of Embryonic Metabolism and Epigenetics, UFABC. Its primary focus is to facilitate comparative analysis of gene expression by providing a user-friendly platform to explore and compare results of *total reads* and *RPKM* across diverse samples and experimental groups. TranscriptomicsDB's intuitive web interface empowers researchers to efficiently navigate vast amounts of data, gaining valuable insights into gene expression patterns within different morphokinetic groups. The platform's objective visualizations further enhance the analysis, enabling users to discern trends and patterns effectively. However, as we expand TranscriptomicsDB to create an invaluable repository for bovine embryonic cell RNA-Seq data, the database is currently unavailable for public access. At this moment, it is functioning locally only for testing purposes. In parallel, we are diligently working to address its shortcomings and improve usability. By incorporating provenance information and conducting comparisons with other databases, we are committed to ensuring that TranscriptomicsDB becomes a vital tool for advancing research in the field of bovine embryonic cell analysis. By offering a centralized repository for standardized and organized gene expression data, TranscriptomicsDB streamlines the research process, allowing researchers to focus more on interpreting the results and generating meaningful conclusions. Overall, this database and its associated web platform play a pivotal role in advancing the understanding of embryonic metabolism and epigenetics, empowering researchers in their quest for innovative discoveries in bovine biology.

References

Baxevanis, A. D. and Bateman, A. (2015). The importance of biological databases in biological discovery. *Current Protocols in Bioinformatics*, 50(1).

Chitwood, J. L., Rincon, G., Kaiser, G. G., Medrano, J. F., and Ross, P. J. (2013). Rna-seq analysis of single bovine blastocysts. *BMC Genomics*, 14:350.

Cánovas, A., Rincon, G., Islas-Trejo, A., Wickramasinghe, S., and Medrano, J. F. (2010). Snp discovery in the bovine milk transcriptome using rna-seq technology. *Mammalian Genome*, 21(11-12):592–598.

Danchin, A., Ouzounis, C., Tokuyasu, T., and Zucker, J.-D. (2018). No wisdom in the crowd: genome annotation in the era of big data – current status and future prospects. *Microbial Biotechnology*, 11(4):588–605.

Desai, N., Ploskonka, S., Goodman, L., Austin, C., Goldberg, J., and Falcone, T. (2014). Analysis of embryo morphokinetics, multinucleation and cleavage anomalies using continuous time-lapse monitoring in blastocyst transfer cycles. *Reprod Biol Endocrinol*, 12:54.

Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P. Łabaj, P., and Mangul, S. (2023). Rna-seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, 14.

Hrdlicková, R., Toloue, M., and Tian, B. (2016). Rna-seq methods for transcriptome analysis: Rna-seq. *Wiley Interdisciplinary Reviews: RNA*, 8.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., and et al. (2013). The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585.

Milazzotto, M. P., Goissis, M. D., Chitwood, J. L., Annes, K., Soares, C. A., Ispada, J., Assumpção, M. E. O. , and Ross, P. J. (2016). Early cleavages influence the molecular and the metabolic pattern of individually cultured bovine blastocysts. *Mol Reprod Dev.*, page 54.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5.

Perampalam, P. and Dick, F. A. (2020). Beavr: A browser-based tool for the exploration and visualization of rna-seq data. *BMC Bioinformatics*, 21(1).

Ullah, S., Rahman, W., Ullah, F., Ahmad, G., Ijaz, M., and Gao, T. (2022). Dbhr: a collection of databases relevant to human research. *Future Science OA*, 8(3):FSO780.

Villalba, G. C. and Matte, U. (2021). Fantastic databases and where to find them: Web applications for researchers in a rush. *Genetics and Molecular Biology*, 44(2):e20200203.

Xu, G., Strong, M. J., Lacey, M. R., Baribault, C., Flemington, E. K., and Taylor, C. M. (2014). Rna compass: A dual approach for pathogen and host transcriptome analysis of rna-seq datasets. *PLoS ONE*, 9(2).