Analysis of published transciptomic data studies to identify common genes and potential markers of bovine respiratory disease (BRD) and respiratory infection.

Shaun van den Hurk

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

**Authors**

* First Autor (ORCID: 0000-0000-1234-5678)
* Second Author
* Third Autor
* Last Author

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Another Affiliation, Athens, GA, USA.
3. Yet another Affiliation.

These authors contributed equally to this work.

Corresponding author: some@email.com

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

*Write a summary of your project.* Respiratory disease is an important condition that affects the health, welfare and production of cattle. Respiratory disease in cattle is often attributed to the Bovine Respiratory Disease Complex (BRDC) and can be caused by a number of different pathogens. Currently there are over 9 different viruses and bacteria that are implicated in the BRDC, with the list growing, and potentiated by factors such as the environment and management factors.

The proposed project is a type of meta-analysis of different published datasets of transcriptomic data as a means to identify some potential biomarkers (such as miRNA) that might be differentially expressed with respiratory infection in bovines. The focus will be on the gene pathways that appear to be differentially expressed. The aim is to identify if there are any associations between infection (broadly and with specific pathogens) and these transcripts, which then could be used as a predictor of disease.

# 2. Introduction

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.* The data to be used is from published studies and online data found on the NCBI Gene Expression Omnibus (GEO) database. A search was performed on this database to look for data related to bovine respiratory disease as well as transcriptome, and/or protein and/or genes. There were between 10 than 22 results search results (depending on the search) on the database, and from these projects with published data, relevant projects are being identified. The studies are only included if they have retrievable data uploaded, either on the GEO database or the corresponding published article (including supplemental data).

Currently, five potential suitable sources have been identified, one such source is listed on GEO as “GSE199108”, with the published article citation “O’Donoghue, S., Earley, B., Johnston, D., Mccabe, M.S., KIM, J., Taylor, J.F., Duffy, C., Lemon, K., Mcmenamy, M., Cosby, S.L. and Morris, D.W., Whole blood transcriptome analysis in dairy calves experimentally challenged with Bovine Herpesvirus 1 (BoHV-1) and comparison to a Bovine Respiratory Syncytial Virus (BRSV) challenge. Frontiers in Genetics, 14, p.171.”. From the mentioned resources, the data with the gene counts and other raw data will be retrieved and these Excel and Text files will provide the raw data. The researchers from these studies published the raw data and the scientific journal article, but not their analysis or their workflow for this. Additionally, each of these studies was performed with direct sampling, and I am not familiar with a study that has combined all of the data in a meta-analysis with collation of the data for a deeper analysis of these transcriptome profiles in cattle.

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

* Are there differences in the expression of genes and proteins between healthy and infected cattle that can be identified through transciptomic analysis? Can these genes be identied in this way.
* Are there any genes or biomarkers that are associated with infection with particular pathogens?
* Are there any common genes or biomarkers identified that are associated with infection with multiple different pathogens?

Thus the outcomes of interest are primarily the levels of expression of these targets and the disease status of the animals sampled.

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (1,2).

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

The raw data has samples from healthy and experimentally infected animals which will be compared to get an indication of differential expression of certain genes between these groups. This will serve as the basis for the analysis as the focus will be on the subset of differentially expressed genes. The raw data from the different sources will be cleaned up in R and analysed before they are combined in a single analysis to look for common features and to compare the expression profiles. Heatmaps will be helpful to help identify and visualise the differential expression of genes, and an approach might be to sample the top 10 genes with high and low levels of expression. Statistical analysis will focus largely on the associations between the factors, for example a t-test, and regression models might be considered too. We will look for associations between gene expression and infection status (heath vs infected).

It is expected that there will be some obvious differences in these levels of expression that will be asociated with the disease status and that many of these will be common between the different respiratory pathogens.

If potential genes or biomarkers are identified they will be run through online sources and databases (including tools such as the PubChem online tool) to identify the common names (or symbol) and proposed functions of these genes.

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 3.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.3 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ | | numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ | |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 2](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 2: Height and weight stratified by gender. |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (3) discusses types of analyses.

These papers (1,2) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

1. McKay B, Ebell M, Billings WZ, et al. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) *Open forum infectious diseases*. 2020;7(11):ofaa494.

2. McKay B, Ebell M, Dale AP, et al. [Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients.](https://doi.org/10.1098/rspb.2020.0496) *Proceedings. Biological sciences*. 2020;287(1927):20200496.

3. Leek JT, Peng RD. [Statistics. What is the question?](https://doi.org/10.1126/science.aaa6146) *Science (New York, N.Y.)*. 2015;347(6228):1314–1315.