An Identification of Metastasis Regulators in Chicken (*Gallus Gallus*) Sarcoma Cell Lines Using Transcriptomic Data

Nhu P.Y Doan[1], Adrienn Szarvas[2]

[1]*Faculty of Agriculture, University of Szeged – H6800, Hódmezővásárhely, Hungary*
[2]*Institute of Plant Sciences and Environmental Protection, Faculty of Agriculture, University of Szeged – H6800, Hódmezővásárhely, Hungary*

Abstract
Rous sarcoma virus (RSV), which is an oncovirus, can cause sarcoma and consequently induce malignant tumours in chicken *(Gallus gallus)*. Research into molecular factors that regulate the tumour-inducing ability are essential to develop prevention and curation methods against RSV. In this study, we aimed to determine candidate genes contributing to the formation of tumours through a transcriptomic analysis in R programming with GSE42516 and GSE15141, which are microarray expression dataset in GEO-NCBI database. We conducted differential expression analysis among a total of 8 metastatic samples and 5 non-metastatic samples, starting from data normalization, then creating model matrixes for pairwise comparisons and using eBayes function to calculate the log fold chance values and significance level of all genes (*p*-value). As a result, in GSE42516, we identified 295 significant (p-value ≤ 0.05) differentially expressed genes (DEGs), with 195 downregulated genes (logFC ≤ -1) and 190 upregulated genes (logFC ≥ 1). While in GSE15141, a greater list of DEGs was extracted, with 1444 downregulated genes and 1314 upregulated genes. Top 5 DEGs retrieved in GSE42516 were *TTC32, DHRS7, RARB, RSPO3, C1QB* and *RBM24, TOM1L1, LIPI, HINTW, C20orf59* were found in GSE15141. Enrichment GO (in this case, biological process - BP) analysis revealed that the DEGs are mainly enriched in *heterochromatin assembly*, *negative regulation of megakaryocyte differentiation* and *endocytosis*. The identified genes may have a vital role in elucidating the molecular metastasis mechanisms and developing effective strategies against sarcoma virus.

Keywords: candidate genes, differential expression analysis, metastasis, p-value, R, RSV.

1. Introduction

In 1908, Vilhemn Ellermann and Oluf Bang first discovered that chicken sarcoma, which is characterized by erythroleukemia transmission between chickens, was caused by an unknown virus. In 1911, Peyton Rous found that sarcoma could be transmitted by cell-free agents, and the virus isolated by Rous was called Rous sarcoma virus (RSV) [1]. This virus belongs to an endogenous retrovirus, referred to as avian sarcoma leukosis viruses (ASLV). The infection of these viruses can lead to the induction of various tumours, the most common of which is lymphoid leukosis. Symptoms of lymphoid leukosis includes the spread of cancer cells from lymphoid tissues, resulting in the tumour formation inside the liver and other abdominal organs [2], which can consequently lead to the weakness, emaciation, high mortality [3], decline in egg production [2] in infected chickens.

Attempts to elucidate complex molecular pathways underlying malignant tumour formation mechanisms induced by RSV have been made to discover therapeutic drugs over the last 50 years [4]. In 1970s, a number of research groups succeeded in identifying *v-src*, an RSV gene located inside its RNA genome, which are

* Corresponding author: Nhu P.Y Doan
nhupydoan@gmail.com

responsible for encoding a tyrosine kinase causing uncontrollable proliferation of host cells when inserted into the chicken genomes [5], [6], [7], [8], [9]. The *in vitro* introduction of *v-src* showed a relatively high metastasis potential with clonal tumour development in chicken wing web tissues [10]. The host genomes also contain a similar gene to *v-src*, *c-src*, which is also encoded non-receptor tyrosine kinase [4]. However, *c-src* seems to be weakly expressed and therefore has little tumorigenesis compared to its viral counterpart [11].

High throughput next-generation sequencing (NGS) has been an effective tool in discovering genetic factors of metastasis and developing therapeutic strategies against tumours [12]. Recent studies have successfully identified several regulators associated with metastasis progress in chicken sarcoma cell lines. A set of 176 genes were commonly expressed in embryo fibroblasts and neurotinal cells in chickens that were introduced with *v-src* [13]. Kovárová D. et al (2013) found that HOPX (homeodomain only protein X) were particularly down-regulated in *v-src* transformed cell lines compared to non-transformed clones [14]. Investigation into EGR1, a transcription factor, was shown to be specifically expressed in PR9692 metastasizing cell line [15].

In this study, we take advantage of transcriptomic data from previous two studies [14], [15] to predict a list of candidate metastasis regulators that play vital roles in tumour formation of chicken. We conducted an in-depth bioinformatics analysis to identify a set of differentially expressed genes (DEGs) by using R Bioconductor packages across all experimental samples [16], starting from data normalization, then creating model matrixes for pairwise comparisons and using eBayes function to calculate significance level (*p*-value) and log fold chance values (logFC) of all genes. The identified genes may have a pivotal role in elucidating molecular mechanisms and developing effective strategies against sarcoma virus.

2. Materials and methods

Data collection

Two microarray datasets GSE42516, GSE15141 from GEO-NCBI database were retrieved to conduct differential expression analysis by GEOquery packages [17], implemented in R version 4.3.1. A total of 13 samples were downloaded: 8 metastasis samples and 5 non-metastasis samples; GSE42516 contains 2 unexpressed metastasis and 2 control (metastasis) samples, while GSE15141 has 3 original metastasizing lines, 3 non-metastasizing lines and 3 restoredly-metastasizing lines.

Data normalization

All retrieved samples were subjected to quantile normalization by log2 transformation. Data manipulation was conducted by "dplyr" packages [18] and gene expression level was visualized with "ggplot2" packages [19].

Differential expression analysis

Differentially expressed genes (DEGs) were extracted with "limma" [20] packages. Significant DEGs were considered to have p-value ≤ 0.05 and $|\log_2 FC| \geq 1$. Volcano plots were plotted to visualize the amount of significant DEGs

Enrichment analysis

Gene ontology (Cellular component – CC, Molecular function – MF, Biological Process – BP) and KEGG were conducted by DAVID online web tool [21]. The GO or KEGG pathways that have the least significance level were reported.

Extraction of DEGs list

The list of significant DEGs was extracted with "readr" packages [22]. Top 20 most differentially expressed genes were provided in the "Results and Discussion" part.

3. Results and discussion

Data collection

The information of the samples collected in GEO database are specified in the table below:

Table 1. Information of samples collected from GEO-NCBI database

Dataset	Samples	Species	Cell lines	Experimental design	Ability to become metastatic lines
GSE42516	GSM1041352	*Gallus gallus*	PR9692	Control line with normal expression of transcription factor HOPX	Yes
GSE42516	GSM1041353	*Gallus gallus*	PR9692	Control line with normal expression of transcription factor HOPX	Yes
GSE42516	GSM1041354	*Gallus gallus*	PR9692	Experimental line with downregulated expression of HOPX	No
GSE42516	GSM1041355	*Gallus gallus*	PR9692	Experimental line with downregulated expression of HOPX	No
GSE15141	GSM378241	*Gallus gallus*	PR9692	Non-transformed, control line	Yes
GSE15141	GSM378242	*Gallus gallus*	PR9692	Non-transformed, control line	Yes
GSE15141	GSM378243	*Gallus gallus*	PR9692	Non-transformed, control line	Yes
GSE15141	GSM378244	*Gallus gallus*	PR9692-E9	Non-transformed, derived cell line	No
GSE15141	GSM378245	*Gallus gallus*	PR9692-E9	Non-transformed, derived cell line	No
GSE15141	GSM378246	*Gallus gallus*	PR9692-E9	Non-transformed, derived cell line	No
GSE15141	GSM378247	*Gallus gallus*	PR9692-E9	EGR1-tranformed line	Yes
GSE15141	GSM378248	*Gallus gallus*	PR9692-E9	EGR1-transformed line	Yes
GSE15141	GSM378249	*Gallus gallus*	PR9692-E9	EGR1-transformed line	Yes

Data normalization

Gene expression level across all samples are gone through $\log_2$ transformation to get an even distribution of expression level for all samples before conducting any type of analysis.

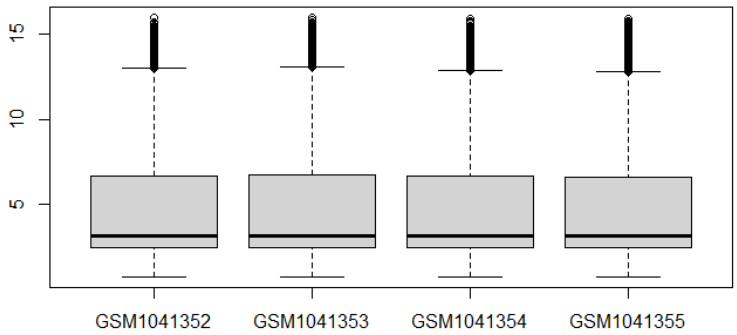

Figure 1A. Expression level of 4 samples (GSM1041352, GSM1041353, GSM1041354, GSM1041355) from GSE42516 (maximum expression values are not higher than 16)

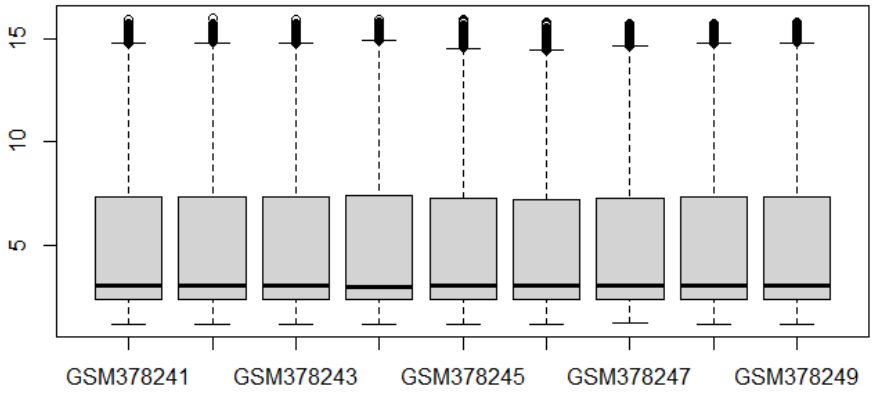

Figure 1B. Gene expression level of 9 samples from GSE15141 (maximum expression values are not higher than 16)

Differential expression analysis

Separate analysis using "limma" packages were conducted to both GSE42516 and GSE15141 to identify the list of significant DEGs. In GSE42516, we found 295 significant DEGs (*p*-value ≤ 0.05), with 195 downregulated genes (logFC ≤ -1) and 190 upregulated genes (logFC ≥ 1); while in GSE15141, a greater list of DEGs was extracted, with 1444 downregulated genes and 1314 upregulated genes. Volcano plot was plotted as a visualization of the number of insignificant and significant DEGs.

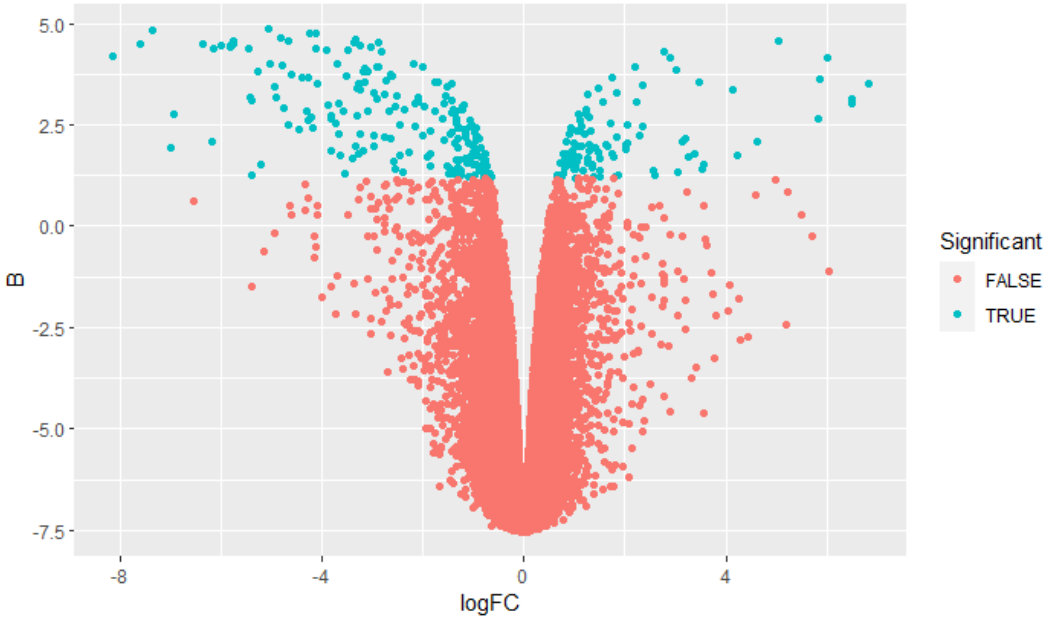

Figure 2A. Differential expression analysis of GSE42516, including insignificant DEGs with p-value ≥ 0.05 (indicated by red dots) and significant DEGs with p-value ≤ 0.05 (indicated by blue dots) (genes that have logFC ≥ 1 are considered upregulated, genes with logFC ≤ -1 are considered downregulated)

Figure 2B. Differential expression analysis of GSE15141. A total of 1444 downregulated genes and 1314 regulated genes were identified among significant DEGs

Extraction of DEGs list

The top 20 most differentially expressed genes of each dataset were reported below. These genes could be considered as potential candidate tumour-formation regulators as they show the substantially high significant level (*p-value*) between metastasis cell lines and non-metastasis cell-lines. The entire lists of DEGs were attached in the supplementary file. The molecular functions and expression pattern of these DEGs were also given here:

Table 2. Top 20 significant DEGs of each dataset and their expression pattern in metastasis samples

No.	Dataset	Gene symbol	Molecular function of gene	Expression pattern in metastasis lines
1	GSE42516	TTC32	Tetratricopeptide repeat domain 32	Upregulated
2	GSE42516	DHRS7	Dehydrogenase/reductase (SDR family) member 7	Upregulated
3	GSE42516	RARB	Retinoic acid receptor, beta	Upregulated
4	GSE42516	RSPO3	R-spondin 3	Upregulated
5	GSE42516	C1QB	Complement component 1, q subcomponent, B chain	Downregulated
6	GSE42516	ATAD1	ATPase family, AAA domain containing 1	Upregulated
7	GSE42516	SLC17A9	Solute carrier family 17 (vesicular nucleotide transporter), member 9	Upregulated
8	GSE42516	AIFM2	Apoptosis inducing factor, mitochondria associated 2	Upregulated
9	GSE42516	HTR1D	5-hydroxytryptamine (serotonin) receptor 1D, G protein-coupled	Upregulated
10	GSE42516	SRGN	Serglycin	Upregulated
11	GSE42516	PDGFRL	Platelet-derived growth factor receptor-like	Upregulated
12	GSE42516	NUDT14	Nudix (nucleoside diphosphate linked moiety X)-type motif 14	Upregulated
13	GSE42516	GATA3	GATA binding protein 3	Upregulated
14	GSE42516	GFPT2	Glutamine-fructose-6-phosphate transamine 2	Upregulated
15	GSE42516	RAB15	RAB15, member RAS oncogene family	Upregulated
16	GSE42516	RHOF	Ras homolog family member F (in filopodia)	Upregulated
17	GSE42516	APBB1IP	Amyloid beta (A4) precursor protein-binding, family B, member 1 interacting protein	Upregulated
18	GSE42516	SLC7A5	Solute carrier family 7 (amino acid transporter light chain, L system), member 5	Upregulated
19	GSE42516	ADCY7	Adenylate cyclase 7	Upregulated

No.	Dataset	Gene symbol	Molecular function of gene	Expression pattern in metastasis lines
20	GSE42516	EPB41L3	Erythrocyte membrane protein band 4.1-like 3	Downregulated
21	GSE15141	RBM24	RNA binding motif protein 24	Upregulated
22	GSE15141	TOM1L1	Target of myb (chicken)-like 1	Upregulated
23	GSE15141	LIPI	Lipase, member I	Downregulated
24	GSE15141	HINTW	Histidine trial nucleotide binding protein W	Downregulated
25	GSE15141	C20orf59	Chromosome 20 open reading frame 59	Upregulated
26	GSE15141	C16orf45	Chromosome 16 open reading frame 45	Upregulated
27	GSE15141	ROBO2	Roundabout guidance receptor 2	Downregulated
28	GSE15141	THAP7	THAP domain containing 7	Downregulated
29	GSE15141	TMEM26	Transmembrane protein 26	Downregulated
30	GSE15141	RPSA	Ribosomal protein SA	Downregulated
31	GSE15141	RCJMB04_8k11	MANSC domain containing 1	Upregulated
32	GSE15141	CCDC104	Coiled – coiled domain containing 104	Upregulated
33	GSE15141	MAP3K14	Mitogen – activated protein kinase 14	Upregulated
34	GSE15141	COL6A3	Collagen, type VI, alpha 3	Upregulated
35	GSE15141	NUDCD3	NudC domain containing 3	Downregulated
36	GSE15141	TFPI2	Tissue factor pathway inhibitor 2	Upregulated
37	GSE15141	KLHL29	Kelch like family 29	Downregulated
38	GSE15141	LASS6	LAG1 homolog, ceramide synthase 6	Upregulated
39	GSE15141	CHIC2	Cystein – rich hydrophobic domain 2	Downregulated
40	GSE15141	CALML4	Calmodulin-like 4	Downregulated

Enrichment analysis

We submitted a total of 2758 significant DEGs extracted from GSE15141 into DAVID online web tool for functional enrichment analysis. GO (Gene Ontology, including CC – Cellular Component, BP – Biological Process, MF – Molecular Function) and KEGG Pathway were conducted. Here, we presented the result of the Biological Process enrichment analysis. Other analyses (CC, MF, KEGG) were attached in the supplementary file.

Table 3. Top Biological Processes that are mainly enriched in significant DEGs list

BP term	Gene count	Percent (%)	*p*-value	Benjamini
Heterochromatin assembly	10	0.9	1.2e-4	2.6e-1
Negative regulation of megakaryocyte differentiation	5	0.4	2.4e-3	1.0e
Endocytosis	14	1.3	2.8e-3	1.0e
DNA replication – independent nucleosome assembly	6	0.5	2.9e-3	1.0e
Endosomal transport	8	0.7	8.7e-3	1.0e
Positive regulation of cell differentiation	4	0.4	1.0e-2	1.0e
Positive regulation of mitochondrial fission	5	0.4	1.1e-2	1.0e
DNA replication – dependent nucleosome assembly	5	0.4	1.4e-2	1.0e
Secretory granule organization	3	0.3	1.4e-2	1.0e

Among all BP term listed above, *heterochromatin assembly* has smallest *p-v*alue (1.2e-4), meaning that it possesses the most significant level across the mentioned term, its total gene count is 10, corresponding to 0.9% of the DEGs list. Meanwhile, *endocytosis* composes a larger proportion of gene count (1.3%), but much higher *p*-value

4. Conclusions

We demonstrated a typical workflow of *in silico* analysis to identify key genes across metastasis samples in published transcriptomics data. The list of potential candidate metastasis regulators was provided as top most significant differentially expressed genes between metastasis and non-metastasis samples. These

molecular regulators can be utilized for early diagnosis of sarcoma virus infection, as well as effective therapeutic strategies against avian leukosis. However, further *in vitro* experiments are necessary to confirm the activity and expression of discovered genes.

Acknowledgements

This study was funded and supported by Faculty of Agriculture, University of Szeged, Hungary.

References

1. A brief chronicle of retrovirology https://www.ncbi.nlm.nih.gov/books/NBK19403/
2. The current problem with avian leukosis J virus https://web.archive.org/web/20110724123631/http://animalscience.ucdavis.edu/Avian/Cpl599.htm
3. The Poultry Guide - Avian Lymphoid Leukosis https://web.archive.org/web/20171211203053/http://www.ruleworks.co.uk/poultry/Avian-Lymphoid-Leukosis.htm
4. Prakash O, Bardot SF, Cole JT. Chicken sarcoma to human cancers: a lesson in molecular therapeutics. Ochsner J. 2007 Summer;7(2):61-4. PMID: 21603517; PMCID: PMC3096390.
5. Wang LH, Duesberg PH, Kawai S, Hanafusa H. Location of envelope-specific and sarcoma-specific oligonucleotides on RNA of Schmidt-Ruppin Rous sarcoma virus. Proc Natl Acad Sci U S A. 1976 Feb;73(2):447-51. doi: 10.1073/pnas.73.2.447. PMID: 174108; PMCID:
6. Lai MM, Hu SS, Vogt PK. Occurrence of partial deletion and substitution of the src gene in the RNA genome of avian sarcoma virus. Proc Natl Acad Sci U S A. 1977 Nov;74(11):4781-5. doi: 10.1073/pnas.74.11.4781. PMID: 200931; PMCID: PMC432039.
7. Weiss SR, Varmus HE, Bishop JM. The size and genetic composition of virus-specific RNAs in the cytoplasm of cells producing avian sarcoma-leukosis viruses. Cell. 1977 Dec;12(4):983-92. doi: 10.1016/0092-8674(77)90163-5. PMID: 202396.
8. Sefton BM, Hunter T, Beemon K. Product of in vitro translation of the Rous sarcoma virus src gene has protein kinase activity. J Virol. 1979 Apr;30(1):311-8. doi: 10.1128/JVI.30.1.311-318.1979. PMID: 225522; PMCID: PMC353324.
9. Brugge JS, Erikson RL. Identification of a transformation-specific antigen induced by an avian sarcoma virus. Nature. 1977 Sep 22;269(5626):346-8. doi: 10.1038/269346a0. PMID: 198667.
10. Stoker AW, Sieweke MH. v-src induces clonal sarcomas and rapid metastasis following transduction with a replication-defective retrovirus. Proc Natl Acad Sci U S A. 1989 Dec;86(24):10123-7. doi: 10.1073/pnas.86.24.10123. PMID: 2557619; PMCID: PMC298657.
11. Ishizawar R, Parsons SJ. c-Src and cooperating partners in human cancer. Cancer Cell. 2004 Sep;6(3):209-14. doi: 10.1016/j.ccr.2004.09.001. PMID: 15380511.
12. Nagahashi M, Shimada Y, Ichikawa H, Kameyama H, Takabe K, Okuda S, Wakai T. Next generation sequencing-based gene panel tests for the management of solid tumors. Cancer Sci. 2019 Jan;110(1):6-15. doi: 10.1111/cas.13837. Epub 2018 Nov 27. PMID: 30338623; PMCID: PMC6317963.
13. Maślikowski BM, Néel BD, Wu Y, Wang L, Rodrigues NA, Gillet G, Bédard PA. Cellular processes of v-Src transformation revealed by gene profiling of primary cells--implications for human cancer. BMC Cancer. 2010 Feb 12;10:41. doi: 10.1186/1471-2407-10-41. PMID: 20152043; PMCID: PMC2837010.
14. Kovárová D, Plachy J, Kosla J, Trejbalová K, Čermák V, Hejnar J. Downregulation of HOPX controls metastatic behavior in sarcoma cells and identifies genes associated with metastasis. Mol Cancer Res. 2013 Oct;11(10):1235-47. doi: 10.1158/1541-7786.MCR-12-0687. Epub 2013 Aug 12. PMID: 23938949.
15. Cermák V, Kosla J, Plachý J, Trejbalová K, Hejnar J, Dvorák M. The transcription factor EGR1 regulates metastatic potential of v-src transformed sarcoma cells. Cell Mol Life Sci. 2010 Oct;67(20):3557-68. doi: 10.1007/s00018-010-0395-6. Epub 2010 May 28. PMID: 20505979.
16. https://www.bioconductor.org/
17. Davis S, Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics*, 14, 1846–1847.
18. Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation* https://dplyr.tidyverse.org/
19. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2016). https://link.springer.com/book/10.1007/978-3-319-24277-4
20. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, 43(7),e47.https://academic.oup.com/nar/article/43/7/e47/2414268
21. Dennis, G., Sherman, B.T., Hosack, D.A. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4, R60 (2003). https://doi.org/10.1186/gb-2003-4-9-r60
22. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the

tidyverse." *Journal of Open Source Software*, 4(43), 1686. doi:10.21105/joss.01686.
https://readr.tidyverse.org/