ISSN: 2694-166X
Scientific Journal
Biology & Life Sciences
DOI: 10.33552/SJBLS.2024.03.000576
Iris Publishers

Research Article

New Trends in R Package: K-Means Analysis and Visualization

Fei Liang[1], Junyue Wang[2] and M Waqar Khan[1]*

[1]*State Key Laboratory of Crop Stress Adaptation and Improvement, Henan Joint International Laboratory for Crop Multi-Omics Research, School of Life Sciences, Henan University, Kaifeng 475004, China*

[2]*State Key Laboratory of Crop Biology, College of Agronomic Sciences, Shandong Agricultural University, Tai'an, Shandong 271018, China*

***Corresponding author:** M Waqar Khan, State Key Laboratory of Crop Stress Adaptation and Improvement, Henan Joint International Laboratory for Crop Multi-Omics Research, School of Life Sciences, Henan University, Kaifeng 475004, China

Received Date: August 05, 2024

Published Date: September 16, 2024

Abstract

An algorithm, which perfectly assign similar data points from a cluster of data to the same set, is K-mean. Different statistical software has been used for data clustering of unsupervised learning and dimensionality reduction. However, the R programing is the best suits for modern data science research. We explored recently developed R package, dedicated to K-means trend analysis and visualization. The package, aptly named "K-means Trend Analyzer" (available at https://github.com/anhuikylin/KmeansTrendAnalyzer) offers advanced tools for efficiently analyzing trends and visually representing patterns in diverse datasets. This work also provides insights into the installation process, practical usage, and emphasizes the significance of this innovative R package in applications such as sample classification and biomarker identification. The overarching goal is to present a comprehensive understanding of the capabilities of "K-means Trend Analyzer" and its potential contributions to trend analysis and visualization within the realm of data analytics.

Keywords: K-means; trend analysis; R package

Introduction

In today's field of life sciences, the rapid development of high-throughput technologies generating immense amount of omics data, encompassing genomics, proteomics, metabolomics, and many more [1–3]. However, these vast and intricate datasets require effective analytical tools to uncover hidden biological patterns and trends [3]. Against this backdrop, the K-means clustering algorithm, as an unsupervised learning method, has garnered widespread attention in recent years for its application in omics analysis [4,5]. K-means is a widely utilized clustering algorithm designed to partition a dataset into K distinct groups. The algorithm commences by initializing K cluster centroids, either randomly or through alternative methods. Each data point is then assigned to the cluster whose centroid is closest in terms of distance computation. The algorithm iteratively identifies and updates the cluster centroids by recalculating the mean of all data points within each cluster. This assignment and centroid update process repeats until convergence, where either the centroids stabilize or a predetermined number of iterations is reached. Minimizing the sum of squared distances between each data point and its assignment to cluster centroid, is

known as inertia. The algorithm's performance may be sensitive to the initial choice of centroids, prompting multiple runs to select the centroids resulting in the minimum inertia [6–8].

For handling large-scale omics data through K-means, the ggplot2 visualization is powerful tool in data exploration and result presentation [9]. Its intuitive syntax enables researchers to effortlessly create various charts, including scatter plots, line graphs, histograms, and more, facilitating a deeper understanding of data distribution, trends, and relationships. Its adaptability extends to diverse data structures and variable types, providing researchers with an intuitive and efficient data visualization solution. Multiple data analyzing platforms plays pivotal role in large scale data analysis. However, the development of R packages make the data analysis simpler, appealing and reliable. This modular programming approach allows the organization of related functionalities into reusable units, thereby enhancing code readability, and maintainability. By developing R packages, it not only facilitates code sharing and team collaboration but also standardizes the development process, supports complex analytical tasks, and fosters community innovation. The existence of R packages enables data scientists and statisticians to work more efficiently while providing robust support for education, training, and the replication and validation of research.

In the R community, package development is not only a manifestation of technological advancement but also a driving force for knowledge sharing and community growth. In this work, we have crafted an R-package leveraging the K-means algorithm and the visualization capabilities of ggplot2. This package proves adept at analytically discerning trends within diverse datasets and presenting patterns visually. Our work delineates the installation procedure and practical application of this innovative R-package, underscoring its pivotal role in sample classification and biomarker identification. The overarching objective is to comprehensively grasp the functionality of "K-means Trend Analyzer" and recognize its potential contributions to trend analysis and visualization within the realm of data analysis.

Materials and Methods

As the landscape of data analytics evolves, a novel R package (version 0.0.1), "Kmeans Trend Analyzer," has been developed to specifically cater to the needs of K-means trend analysis and visualization. Initially, we perform z-score normalization on the data row-wise. Subsequently, we identify the optimal number of clusters. We then employ the k-means analysis function from the stats package [8,10], coupled with the visualization capabilities of ggplot2, to cluster and visualize multi-sample or time-series omics data. We delve into the features of this newly minted tool, offering a roadmap for users to navigate its installation and utilization.

Results

While initiating K-mean trend analysis, we obtained R package by adopting the basics steps provided in Table 1. The installation code for the K-means Trend Analyzer package is:

Table 1: Functions found in KmeansTrendAnalyzerand their short description.

Function	Description
cncalc	This function is utilized to determine the optimal number of clusters for k-means clustering on multi-sample omics data.
Data	A time-series omics matrix.
KmeansR	This function is designed for computing k-means clustering and creating visualizations for multi-sample omics data.
Data	A time-series omics matrix.
Centers	Number of cluster.
Table	Logical value, If TRUE is selected, a kmeans_result.csv and centre_line.csv will be output. The content involves performing z-score normalization on each row of the data frame, along with information about different clusters.
Angle	The angle of rotation for x-axis labels.
Box	Logical. Should a border be drawn around the plot?
KmeansR2	This function is designed for computing k-means clustering and creating visualizations for multi-sample omics data, included a display of classification percentages.
Data	A time-series omics matrix with class and color columns.
Centers	Number of cluster.
Table	Logical value, If TRUE is selected, a kmeans_result.csv and centre_line.csv will be output. The content involves performing z-score normalization on each row of the data frame, along with information about different clusters.
Angle	The angle of rotation for x-axis labels.
Box	Logical. Should a border be drawn around the plot?
Label size	Label size on bar chart.
Legend position	Specify the desired location of legends on the Bar Chart by choosing from options such as "none," "left," "right," "bottom," "top," or a two-element numeric vector.

Devtools:install_github("anhuikylin/kmeanstrendanalyzer")

To showcase the results and graphs generated by the K-meansTrendAnalyzer, we utilized our previously obtained metabolomics data [2]. This represents a set of metabolomics data comparing the application of SPD (compound of mepiquat chloride, prohexadione-calcium and uniconazole) with the control at various time points. This figure was generated using the KmeansTrendAnalyzer package which show a high correlation with the previous results (Figure 1). Our previous results were obtained through analysis using SPSS, followed by visualizing individual clusters in Origin and eventually stitching them together using Photoshop. This process was highly complex and lacked the ability to display classification information for each cluster. In contrast, the K-means Trend Analyzer package can accomplish these tasks in a single step, streamlining the entire process and providing detailed classification information for each cluster.

After supplementing our classified data, the results still showed the stability and reliability. We used the following codes to perform the KmeansR is:

a. Library (KmeansTrendAnalyzer)

b. library(tidyverse)

c. data <- mango_FI_DAM %>%

d. select (Index, `TA-1`:`C-3`,Class = Class.I) %>%

e. column_to_rownames("Index") %>%

f. mutate (TA = rowMeans(select(., 1:3)),

g. TB = rowMeans(select(., 4:6)),

h. TC = rowMeans(select(., 7:9)),

i. A = rowMeans(select(., 10:12)),

j. B = rowMeans(select(., 13:15)),

k. C = rowMeans(select(., 16:18))) %>%

l. select(!contains("-")) %>%

m. mutate(Colour = case_when(

n. Class == "Amino acids and derivatives" ~ "#FF0000",

o. Class == "Lipids" ~ "#FFFF00",

p. Class == "Nucleotides and derivatives" ~ "#00FF00",

q. Class == "Organic acids" ~ "#00FFFF",

r. Class == "Phenolic acids" ~ "#0000FF",

s. Class == "Others" ~ "#FF00FF"

t.))

u. df_mango <- data

v. set.seed(400)

w. KmeansR2(df_mango,centers = 9)

x. # ggsave("df_mango.pdf",width = 20,height = 12)

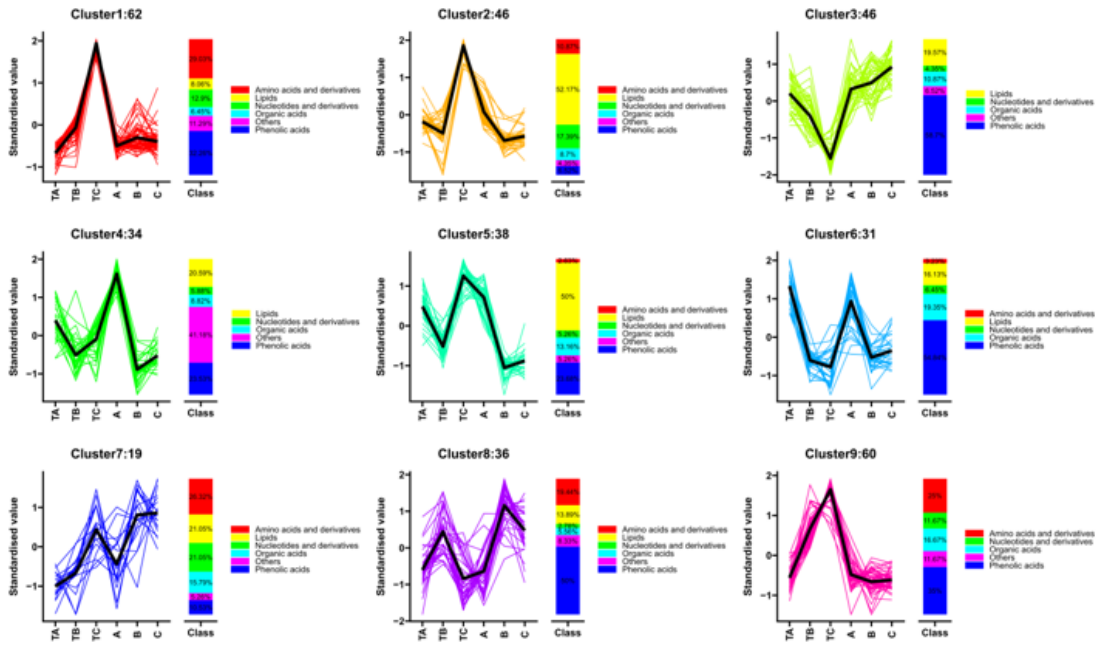

Figure 1: K-means analysis of differential metabolites. The horizontal axis (TA, TB, TC, A, B, C) represents the sample names, 'Class' signifies the proportion of each cluster in the classification, and the vertical axis represents the standardized relative amounts of metabolites. The numeric characters following each cluster indicate the quantity of metabolites in that cluster.

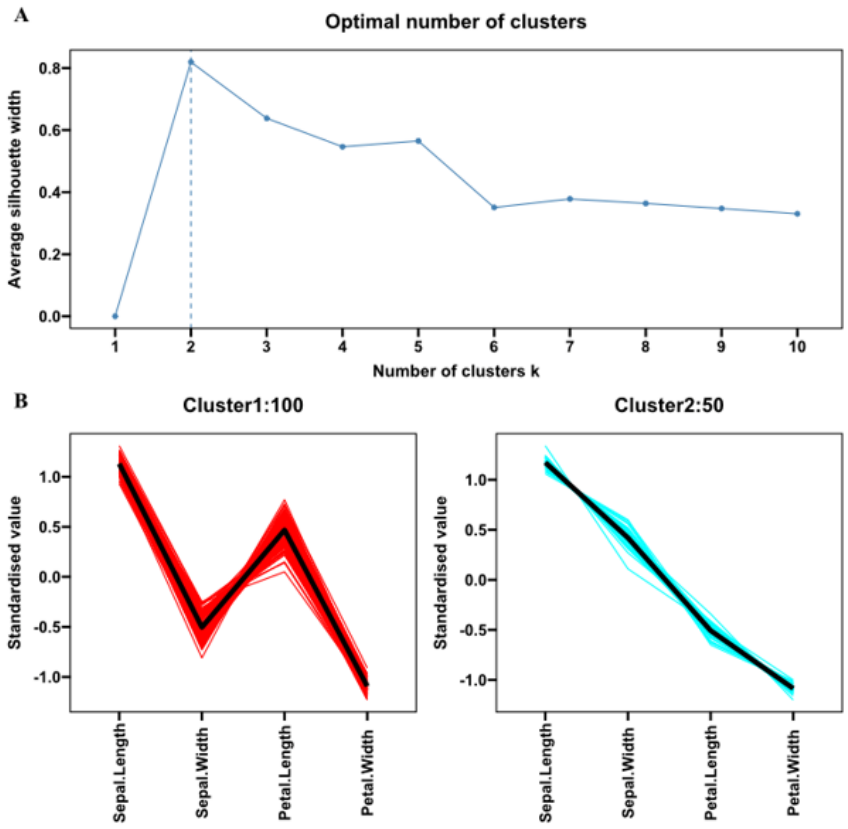

Figure 2: K-means Analysis of Iris Phenotypes. (A) The horizontal axis represents the number of clusters (k), and the vertical axis represents the average silhouette width. (B) The horizontal axis represents sample names, while the vertical axis represents the standardized relative values of phenotypic measurements.

The 'Iris' dataset is a classic benchmark dataset used for exploring and validating classification algorithms. The distinct differences in the four features among different species of iris flowers make it a commonly utilized dataset in both academic research and practical applications [11]. We conducted K-means analysis using Iris data by performing the following codes in the KmeansR package. The results was consistent with our previous findings (Figure 2).

a. library(KmeansTrendAnalyzer)

b. cncalc(iris[1:4])

c. set.seed(400)

d. KmeansR(iris[1:4],centers = 2,angle = 90,box = TRUE)

We also utilized the KmeansTrendAnalyzer package to analyze 5312 differentially expressed genes as analyzed by Cheng et al, (2005). They employed pairwise comparison and weighted gene co-expression network analysis to investigate differential gene expression. During their analysis these genes were categorized into 10 gen co-expression module. In the MElightcyan module, 46 potential genes associated with the regulation of cotton flower bud differentiation were pinpointed. Notably, these genes exhibited expression during the flower bud differentiation stage. Within the MElightcyan module, GhCAL, a novel key regulatory gene linked to flower bud differentiation, was specifically identified [12]. However, while using K-meanTrendanalyzer for the mention data, we determining the optimal number of 4 clusters. We observed a similarity in the expression patterns between cluster 4 and the MElightcyan module. Subsequently, upon comparing cluster 4 and the MElightcyan module, 93% of the genes have been sorted in the MElightcyan module and were identified in cluster 4 (Figure 3). Notably, GhCAL was also present in cluster 4. This indicates that accurate results can still be obtained using KmeansTrendAnalyzer. However, utilizing WGCNA requires researchers to possess a strong programming background, and the program involves lengthy computations, which can be unfriendly for many researchers. In contrast, the analysis process with Kmeans Trend Analyzer only requires two functions, and the plotting process utilizes the map function, ensuring a simple, fast, and accurate computation result.

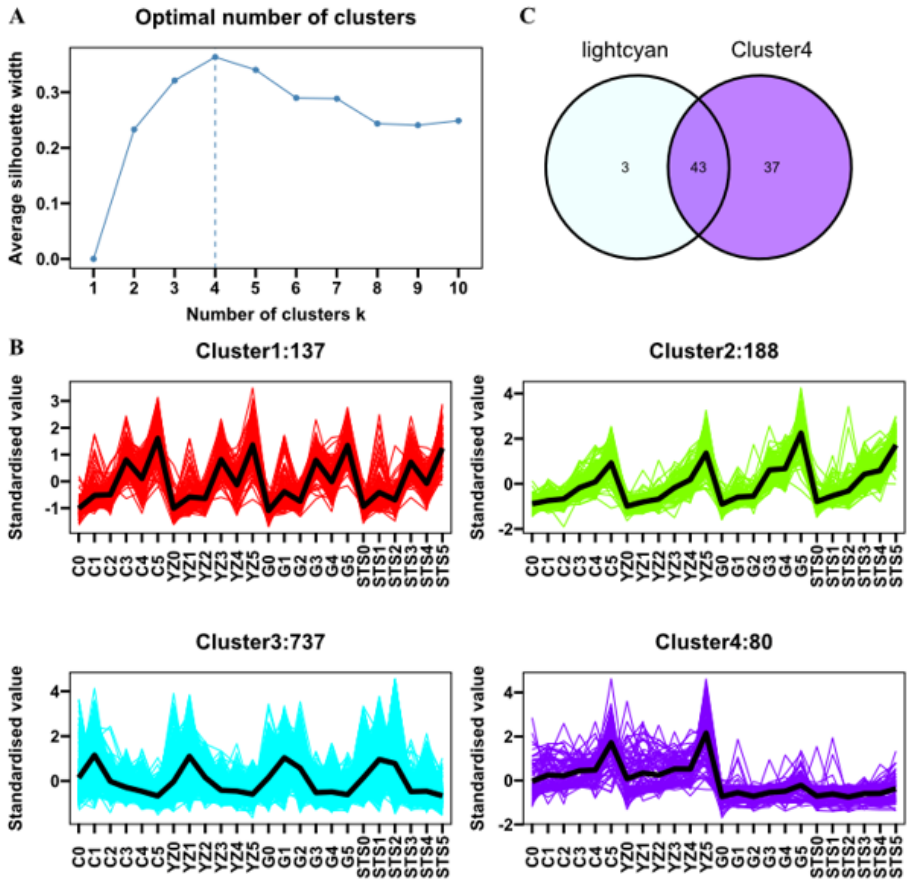

Figure 3: K-means Analysis of cotton transcriptome data. (A) The horizontal axis represents the number of clusters (k), and the vertical axis represents the average silhouette width. (B) The horizontal axis represents the sample names, and the vertical axis represents the standardized relative amounts of genes. (C) The light cyan circle represents the results obtained from WGCNA analysis, while Cluster 4 represents the results from the analysis conducted with the K-meansTrendAnalyzer package.

Discussion

Multi-omics is a comprehensive approach that integrates various biological techniques and informatics methods to conduct a thorough study of biological systems. It encompasses research across multiple levels, including genomics, transcriptomic, proteomics, metabolomics, and more. It enables a more comprehensive understanding of the structure and functionality of biological systems. Presently, multi-omics has been applied across various domains, including medical research, agricultural science, microbiology, environmental science, anthropology, and beyond [13–18]. Trend analysis finds broad applications in plant omics. To thoroughly examine the metabolic shifts occurring throughout the rice growth cycle, Yang et al. employed the k-means clustering algorithm to categorize all 825 annotated metabolites into 12 clusters based on their accumulation patterns. Through the analysis of these 12 clusters, Yang et al. successfully identified metabolites that exhibited abundance in specific tissues. This approach enables a detailed exploration of the dynamic changes in metabolite profiles over the course of rice development, shedding light on tissue-specific metabolic variations [3]. Li et al. conducted a K-Means clustering analysis to unveil genes exhibiting analogous expression patterns. This analysis serves to streamline the identification of potential transcription factors (TFs) that may regulate structural genes related to sugar metabolism and transport [19]. Our previous research unveiled distinct expression patterns of diverse categories of metabolites during the induction phase of mango flowering [2]. Furthermore, an assessment was conducted between Kmeans Trend Analyzer and WGCNA. The Kmeans Trend Analyzer package demonstrated the ability to analyze data simply, quickly, and accurately, showcasing significant potential.

Conclusion

Cutting-edge K-means Clustering Algorithm: The package incorporates a state-of-the-art K-means clustering algorithm, ensuring swift and accurate data classification. Advanced Trend Analysis and Visualization Tools: "KmeansTrendAnalyzer" (https://github.com/anhuikylin/KmeansTrendAnalyzer) is equipped with sophisticated tools designed for in-depth trend analysis and visually appealing data representation. This paper sheds light on the

capabilities of the "KmeansTrendAnalyzer" R package, showcasing its potential to become an indispensable asset in the toolkit of data analysts and researchers.

Author contributions

Liang Fei, Wang Junyue, and M. Waqar Khan collaboratively designed this experiment. Liang Fei was responsible for coding, code refinement, and article composition, while Wang Junyue conducted code reviews. M. Waqar Khan played a role in editing the article.

Declaration of Competing Interest

The authors declare no conflict of interest.

Funding

The author(s) declare that financial support was received for conducting the research, contributing to the authorship, and/or facilitating the publication of this article.

References

1. Schneider MV, Orchard S (2011) Omics Technologies, Data and Bioinformatics Principles. Bioinformatics for Omics Data 719(1): 3-30.
2. Liang F, Xu W, Wu H, Bin Zheng, Qingzhi Liang, et al. (2022) Widely targeted metabolite profiling of mango stem apex during floral induction by compond of mepiquat chloride, prohexadione-calcium and uniconazole. PeerJ 10(1): e14458-e14465.
3. Yang C, Shen S, Zhou S, Yufei Li, Yuyuan Mao, et al. (2022) Rice metabolic regulatory network spanning the entire life cycle. Molecular Plant 15(2): 258-275.
4. Kaur P, Singh A, Chana I (2021) Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions. Archives of Computational Methods in Engineering 28(6): 1-37.
5. Arjmand B, Hamidpour SK, Tayanloo Beik A, Parisa Goodarzi, Hamid Reza Aghayan, et al. (2022) Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. Frontiers in Genetics 13(1): 824451-824455.
6. Ikotun A M, Ezugwu A E, Abualigah L, Belal Abuhaija, Jia Heming (2023) K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences 622(11): 178-210.
7. Esteves RM, Hacker T, Rong C (2013) Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets. 2013 IEEE 5th International Conference on Cloud Computing Technology and Science. Bristol, United Kingdom IEEE 1(1): 17-24.
8. Hartigan JA, Wong MA (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics 28(1): 100-108.
9. Villanueva RAM, Chen ZJ (2019**)** ggplot2: Elegant Graphics for Data Analysis (2nd ed.). Measurement: Interdisciplinary Research and Perspectives 17(3): 160-167.
10. Lloyd S (1982) Least squares quantization in PCM. IEEE Transactions on Information Theory 28(2): 129-137.
11. Fisher RA (1936) THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. Annals of Eugenics 7(2): 179-188.
12. Cheng S, Chen P, Su Z, Liang Ma, Pengbo Hao, et al. (2021) High-resolution temporal dynamic transcriptome landscape reveals a GhCAL-mediated flowering regulatory pathway in cotton (Gossypium hirsutum L.). Plant Biotechnology Journal 19(1): 153-166.
13. Wang P, Clark NM, Nolan TM, Gaoyuan Song, Parker M Bartz, et al. (2022) Integrated omics reveal novel functions and underlying mechanisms of the receptor kinase FERONIA in Arabidopsis thaliana. The Plant Cell 34(7): 2594-2614.
14. Li JX, Li RZ, Sun A, Hua Zhou, Erwin Neher, et al. (2021) Metabolomics and integrated network pharmacology analysis reveal Tricin as the active anti-cancer component of Weijing decoction by suppression of PRKCA and sphingolipid signaling. Pharmacological Research 171(1): 105574-105579.
15. Li P, Yan MX, Liu P, Dan Jie Yang, Ze Kun He, et al. (2023) Multiomics Analyses of Two Leonurus Species Illuminate Leonurine Biosynthesis and Its Evolution. Molecular Plant 17(1): 158-177.
16. Chen S, Wang P, Kong W, Kun Chai, Shengcheng Zhang, et al. (2023) Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant Camellia sinensis. Nature Plants 9(12): 1986-1999.
17. Spänig S, Eick L, Nuy J K, Daniela Beisser, Margaret Ip, et al. (2021) A multi-omics study on quantifying antimicrobial resistance in European freshwater lakes. Environment International 157(1): 106821-106828.
18. Viana J N, Pilbeam C, Howard M, Brett Scholz, Zongyuan Ge, et al. (2023) Maintaining High-Touch in High-Tech Digital Health Monitoring and Multi-Omics Prognostication: Ethical, Equity, and Societal Considerations in Precision Health for Palliative Care. OMICS: A Journal of Integrative Biology 27(10): 461-473.
19. Li L, Wu HX, Ma XW, Wen Tian Xu, Qing Zhi Liang, et al. (2020) Transcriptional mechanism of differential sugar accumulation in pulp of two contrasting mango (Mangifera indica L.) cultivars. Genomics 112(6): 4505-4515.