Analysis of ATAT, ACAC and AGAG Microsatellite from the Worldwide *Cactaceae Barcodes*

Alejandra Guzmán-Cepeda[1], Argelia Francisca Tapia-Canseco[2], Ernesto Rios-Willars[3]

[1]Universidad Carolina. Calzada Antonio Narro 707, Zona Centro C.P. 25000, Saltillo, Coahuila, México.

[2]CBTis 235. Calle 20 Sin Número. Saltillo Coahuila México

[3]Facultad de Sistemas de laUniversidad Autónoma de Coahuila. Ciudad Universitaria, Fundadores km 13, Zona Centro C.P. 25350. Arteaga, Coahuila.

Resumen

Los códigos de barras genéticos son herramientas esenciales en biotecnología y biología evolutiva, utilizadas para la identificación de especies y el estudio de la biodiversidad. Sin embargo, la presencia de microsatélites en estos códigos puede influir en la precisión de la información genética que representan. En este estudio, se realizó un análisis bioinformático de microsatélites en códigos de barras de la familia *Cactaceae*, utilizando la herramienta MFind. Se analizaron 3,203 secuencias obtenidas de la base de datos BOLD, abarcando registros de 26 países. Se identificaron tres tipos principales de microsatélites: ACAC, AGAG y ATAT, observando que ATAT presentó la mayor frecuencia de repeticiones. Además, los análisis de correlación revelaron relaciones significativas entre las repeticiones de estos microsatélites (p = 0.00), sugiriendo patrones de distribución que podrían afectar la identificación de especies mediante códigos de barras genéticos. Los resultados resaltan la importancia de considerar la presencia de microsatélites en estudios filogenéticos y de conservación.

Palabras Claves: *Códigos de barras genéticos, microsatélites, Cactaceae, bioinformática, MFind, biodiversidad.*

Abstract

Genetic barcodes are essential tools in biotechnology and evolutionary biology, used for species identification and biodiversity studies. In this study, we conducted a bioinformatics analysis of microsatellites in Cactaceae barcodes using the powerful MFind tool. This tool allowed us to analyze a total of 3,203 sequences from the BOLD database, covering records from 26 countries. Three main microsatellite types were identified: ACAC, AGAG, and ATAT, with ATAT showing the highest frequency of repetitions. Correlation analyses revealed significant relationships between microsatellite occurrences (p = 0.00), suggesting distribution patterns that may influence species identification using genetic barcodes. The results underscore the importance of considering microsatellite presence in phylogenetic and conservation studies.

Key Words: *Genetic barcodes, microsatellites, Cactaceae, bioinformatics, MFind, biodiversit*

Introduction

Genetic barcodes, particularly DNA barcodes, are powerful tools in evolutionary biology, providing insights into species identification, lineage tracing, and evolutionary processes. They are short gene sequences that identify species and understand ecological and evolutionary dynamics. This is particularly useful in species identification and evolutionary studies: DNA barcodes are crucial for identifying species and understanding species boundaries, community ecology, and functional trait evolution. They help in studying trophic interactions and conserving biodiversity by providing a standardized method for species identification across various taxonomic groups (Gostel & Kress, 2022). The analysis of microsatellites in the genetic barcodes of the family Cactaceae is essential due to the uniqueness and diversity of this plant group, which includes numerous endemic species adapted to extreme environments. Microsatellites, due to their high variability and distribution in the genome, can influence the accuracy of molecular identification based on barcodes. Studying these repetitive elements in the barcodes of Cactaceae allows us to detect possible sources of error or ambiguity in the delimitation of species, which is especially relevant in a group with a high incidence of hybridization and morphological convergence. The Cactaceae family has significant ecological, economic, and cultural importance, being key in the conservation of arid and semi-arid ecosystems, as well as in the provision of resources for human communities.

A genetic barcode is a short DNA sequence that identifies and differentiates between species or genetic material. It serves as a unique identifier, much like a product barcode, allowing for the rapid and accurate classification of organisms or genetic elements. However, barcodes are susceptible to repetitions in the form of microsatellites that influence the genetic information they represent. The most common types of genetic barcodes are the Genome Barcodes: These are unique k-mer frequency distributions across a genome, used to identify and differentiate genomes. They are particularly useful in metagenome binning and identifying horizontally transferred genes. The DNA Barcodes are standardized short sequences of DNA, typically 400-800 base pairs long, used for species identification. They are widely used in ecology, evolution, and conservation to understand species boundaries and interactions (Kress, 2017). More specifically, genetic barcodes are useful in species identification, where the DNA barcodes are extensively used to identify species, aiding in biodiversity conservation and ecological studies. Also, genome barcodes help in studying genome structure and function, particularly in determining phylogenetic relationships and gene transfer events (Zhou et al., 2008).

Some vegetable genes are commonly used as barcodes.

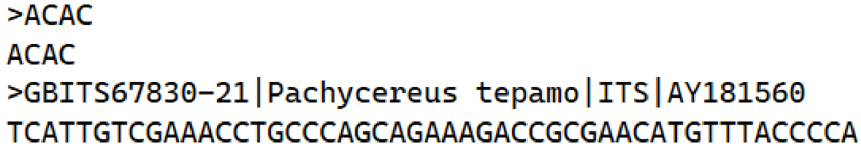
Genes used as plant barcodes mainly include rbcL, matK, trnH-psbA, ITS2, and ycf1. These genes have been developed and tested to identify plant species and address questions in systematics, ecology, evolutionary biology, and conservation (Dong et al., 2015). The rbcL and matK are two geneswidely recommended as the central barcode for plants due to their species-discrimination ability and universal applicability (Hollingsworth et al., 2009). The combination of rbcL and matK provides a universal framework for plant identification. The nuclear marker ITS2 is commonly used in plant phylogenetic investigations and shows high levels of interspecific divergence. On the other hand, ycf1 is one of the most variable loci in the plastid genome that is more effective than other plastid candidates for species identification (Li et al., 2021). However, species-level resolution can be challenging, especially in regions with low floristic diversity (Braukmann et al., 2017).

Dinucleotide microsatellites, such as AT repeats, are common in plants and animals. In plants, AT repeats are the most frequent, while in animals, AC/TG repeats are more common. Trinucleotide repeats, such as TAT, are also prevalent and have been observed in various plant species (Morgante & Olivieri, 1993). Also, tetranucleotide and lengthier microsatellites exist and are used in genetic studies, especially in rice, where between 5,700 and 10,000 microsatellites have been identified (McCouch et al., 1997). In this context, other studies have been carried out on the genetic diversity of cacti in Mexico: 10 microsatellite loci were developed for *Pachycereus pringlei* from the Sonoran Desert, showing polymorphism with an average of 6.3 alleles per locus. These loci help investigate population structure and genetic diversity (Gutiérrez Flores et al., 2014). Likewise, in another study, eight microsatellite markers were developed for *Mammillaria crucífera*, endemic of Mexico. These markers help to describe population structure and support conservation (Solórzano et al., 2009).

Materials and Method

The MFind tool (Rios-Willars & Chirinos-Arias, 2024) was used for microsatellite analysis, and a query was made to the BOLD database (Ratnasingham, 2007) with the classifier "Cactaceae[tax]" in February 2025. With this, 3,203 barcodes were found in records from 26 countries deposited in 27 institutions worldwide. For processing, computer equipment with the following characteristics was used: Processor 11th Gen Intel(R) Core™ i9-11900K @ 3.50GHz Installed RAM 64.0 GB (63.7 GB usable). The system type is a 64-bit operating system with an x64-based processor. The barcode database and the MFind output are available in the galaxy (Afgan et al., 2018) platform link: https://usegalaxy.org/u/rioswillars/h/cactaceacedb Regarding the use of the MFind tool, a download of the Java code corresponding to the software was made. Subsequently, the database of cactus barcodes was incorporated for processing. It is essential to ensure that the database is in multifasta format and that the first sequence is incorporated as a "problem sequence". The first position is the microsatellite to be searched for throughout the rest of the sequences, for example, the "ATAT" microsatellite. Figure 1 is a screenshot of the beginning of the multifasta file where the first sequence to be searched is found.

The graphs and correlation calculations were performed using the statistical tool MiniTab16.



```
>ACAC
ACAC
>GBITS67830-21|Pachycereus tepamo|ITS|AY181560
TCATTGTCGAAACCTGCCCAGCAGAAAGACCGCGAACATGTTTACCCCA
```

Figure 1. Rendering the sequences in the multifasta file. The first sequence is the problem sequence “ACAC”.

Results

For the ACAC microsatellite, 924 records corresponding to two hits and 468 records for one hit were found in the database. Figure 2A illustrates these findings. A higher prevalence was found for the AGAG microsatellite than the previous one, with the highest recorded being one hit per genetic marker in the database, 998. The records with two hits for that microsatellite follow this count. Figure 2B illustrates the counting values for this variable. Finally, the ATAT microsatellite had the highest number of hits compared to the previous two. In this case, the count of six repetitions stands out, and 576 represents the highest bar in Figure 2C. It should be noted that the counts for the cases of zero repetitions were included in the graphs to contrast the results. This count represents the absence of the microsatellite in question for genetic markers. The representation of the count of the three different microsatellites in surface graph format clearly illustrates a genetic picture of each microsatellite's distribution and prevalence compared to the others. It also provides current information on the genetic characteristics of these plants. Figure 3 is a surface illustration of the ATAT vs ACAC and AGAG microsatellites. There are peaks in prevalence in the central part and towards the back of the upper left corner. This represents relevant information for genetic markers, such as identifiers of cacti of particular interest. Finally, we found that the correlation between the variables is positive. When ATAT increases, AGAG also increases (P=.62; p=0.00), as does ACAC (P=0.65; p=0.00). In this sense, when ACAC increases, ATAT increases (P=0.70; p=0.00). Table 1 illustrates the Pearson correlation P-values for the three variables. The set of results in .csv format is also available at the following link on the Galaxy platform.

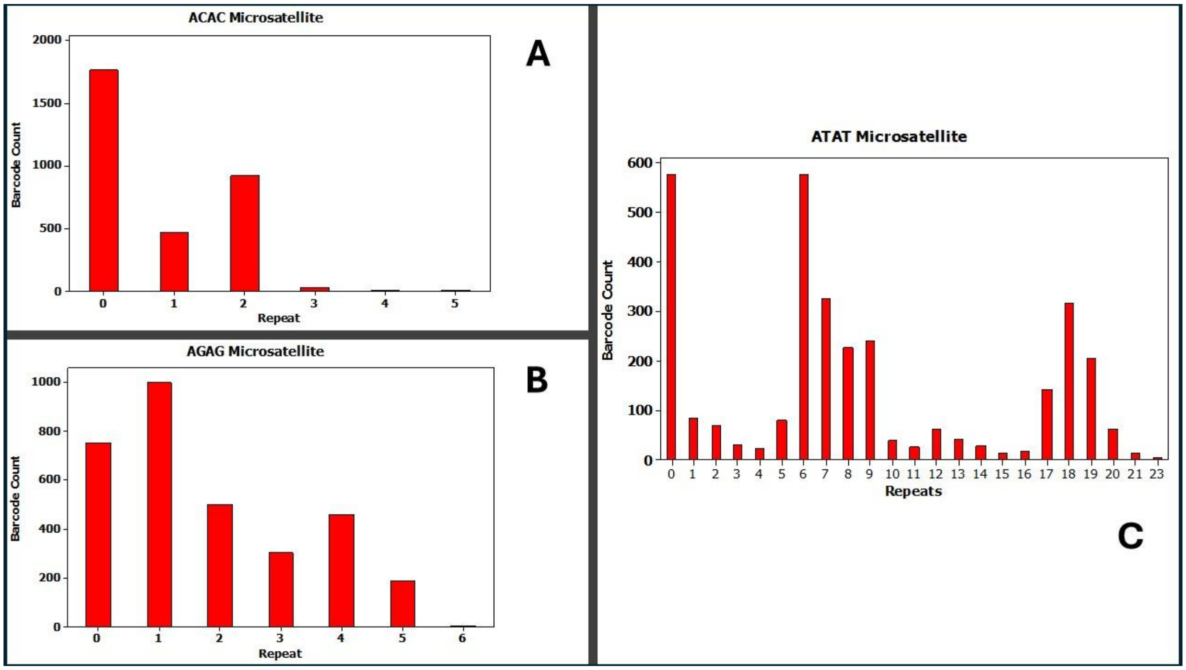


Figure 2. Description of microsatellite findings on the analyzed database. The figures describe the count of barcodes vs microsatellite repeat. Section A depicts the ACAC microsatellite. Section B depicts the AGAG microsatellite, and Section C depicts the ATAT microsatellite.

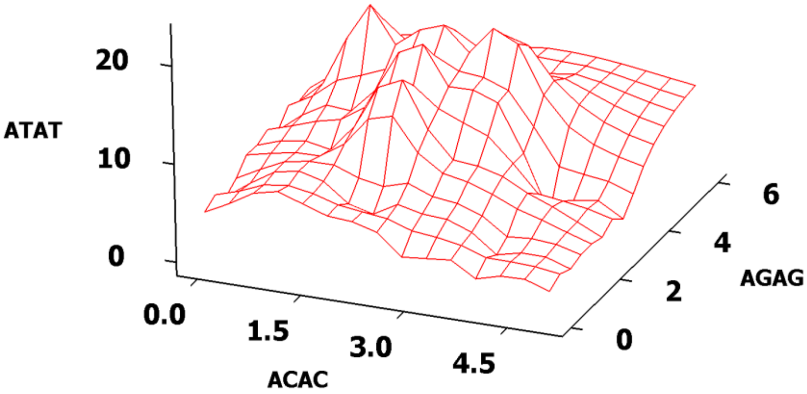


Figure 3. It shows the surface of the ATAT vs ACAC and AGAG microsatellites. The noticeable peaks describe the prevalence of ATAT microsatellites vs AGAG and ACAC.

Table 1. Correlation between the variables ATAT, ACAC, and AGAG with p = 0.00

	AGAG	**ATAT**
ATAT	0.625	
ACAC	0.653	0.705

Discussion

The analysis of microsatellites in the genetic barcodes of the Cactaceae family reveals the predominant presence of the ATAT, ACAC, and AGAG motifs, with the first showing a notably higher frequency. This finding suggests that certain microsatellites may be associated with conserved or functionally relevant regions within the barcodes, which could influence the accuracy of molecular species identification. The significant correlation between the repetitions of these microsatellites indicates the existence of non-random distribution patterns, possibly related to the adaptive evolution of the family in extreme environments and its high genetic diversity. These results highlight the importance of considering the presence and distribution of microsatellites when using genetic barcodes for species delimitation, especially in complex taxonomic groups such as Cactaceae. The variability generated by microsatellites can be a source of error in identification, but it also represents an opportunity to explore additional markers of genetic diversity and phylogeny. Furthermore, detailed knowledge of these patterns can help improve conservation strategies by enabling more precise identification of threatened species and facilitating decision-making in biodiversity management and protection programs.

Finally, this study provides a novel perspective by integrating bioinformatics tools for the analysis of microsatellites in barcodes, which can be replicated in other plant families of biotechnological and ecological interest. The methodology and results obtained lay the groundwork for future research aimed at understanding the functional role of microsatellites in plant evolution and adaptation, as well as their impact on the effectiveness of molecular identification systems used in conservation biology.

Conclusions

According to the literature, the results obtained in this work are congruent with other related works. The presence of microsatellites in genetic markers can influence the information of the barcodes themselves. Likewise, the MFind tool showed results efficiently, and it is recommended to use it for future research in biotechnology and its disciplinary areas. The count of ATAT-type microsatellites stood out for their prevalence throughout the database analyzed, representing an opportunity for their use as a resource for constructing surface graphs such as the one presented here. In this sense, it was observed that the count of microsatellites such as AGAG and ACAC had a lower presence and, at the same time, a direct correlation with ATAT. This suggests that the barcodes used for cactus analysis may be saturated with microsatellites, affecting their efficiency as genetic identifiers.

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J. , Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible, and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Braukmann, T. W. A., Kuzmina, M. L., Sills, K. , Zakharov, E. V., & Hebert, P. D. N. (2017). Testing the Efficacy of DNA Barcodes for Identifying the Vascular Plants of Canada. *PLOS ONE*, *12*(1), e0169515. https://doi.org/10.1371/journal.pone.0169515

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., Cheng, T., Guo, J., & Zhou, S. (2015) . ycf1, the most promising plastid DNA barcode of land plants. *Scientific Reports*, *5*(1), 8348. https://doi.org/10.1038/srep08348

Gostel, M. R., & Kress, W. J. (2022). The Expanding Role of DNA Barcodes: Indispensable Tools for Ecology, Evolution, and Conservation. In *Diversity* (Vol. 14, Issue 3). https://doi.org/10.3390/d14030213

Gutiérrez Flores, C., Lozano Garza, O. A., León de la Luz, J. L., & García de León, F. J. (2014). Development and characterization of 10 microsatellite loci in the giant cardon cactus, *Pachycereus pringlei* (Cactaceae). *Applications in Plant Sciences*,*2*(2). https://doi.org/10.3732/apps.1300066

Hollingsworth, P. M., Forrest, L. L., Spouge, K. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S. W., James, K. E., Kim, K.-J., Kress, W. J., Schneider, H., van AlphenStahl, J., Barrett, C.H., van den Berg, C., Bogarin, D., … S. Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Kress, W. J. (2017). Plant DNA barcodes: Applications today and in the future. *Journal of Systematics and Evolution*, *55*(4), 291–307. https://doi.org/10.1111/jse.12254

Li, H., Xiao, W., Tong, T., Li, Y., Zhang, M., Lin, X., Zou, X., Wu, Q., & Guo, X. (2021). The specific DNA barcodes based on chloroplast genes for species identification of Orchidaceae plants. *Scientific Reports*, *11*(1), 1424. https://doi.org/10.1038/s41598-021-81087-w

McCouch, S. R., Chen, X., Panaud, O., Temnykh, S., Xu, Y., Cho, Y. G., Huang, N., Ishii, T., & Blair, M. (1997). Plant Molecular Biology. *Plant Molecular Biology*, *35*(1/2), 89–99. https://doi.org/10.1023/A:1005711431474

Morgante, M., & Olivieri, A. M. (1993). PCR-amplified microsatellites as markers in plant genetics. *The Plant Journal*, *3*(1), 175–182. https://doi.org/10.1046/j.1365-313X.1993.t01-9-00999.x

Ratnasingham, S. , & E. P. (2007). *The Barcode of Life Data System*. Molecular Ecology Notes 7, 355 - 364.

Rios-Willars, E., & Chirinos-Arias, M. C. (2024). Mfind: a tool for DNA barcode analysis in angiosperms and its relationship with microsatellites using a sliding window algorithm. *Planta*, *259*(6), 134. https://doi.org/10.1007/s00425-024-04420-3

Solórzano, S., Cortés-Palomec, A., Ibarra, A.,Dávila, P., & Oyama, K. (2009). Isolation, characterization, and cross-amplification of polymorphic microsatellite loci in the threatened endemic *Mammillaria crucigera* (Cactaceae). *Molecular Ecology Resources*, *9*(1),156–158. https://doi.org/10.1111/j.1755-0998.2008.02422.x

Zhou, F., Olman, V., & Xu, Y. (2008). Barcodes for genomes and applications. *BMC Bioinformatics*, *9*(1), 546. https://doi.org/10.1186/1471-2105-9-546