International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 20s, (2025)
https://theaspd.com/index.php/ijes

Bioinformatics And Machine Learning To Improve Cotton By Means Of Transcription Factors And Ssrs

Byron Oviedo-Bayas[1], Cristian G. Zambrano-Vega[2], Ronald Oswaldo Villamar-Torres[3]
[1]Universidad Técnica Estatal de Quevedo, Facultad de Posgrado, boviedo@uteq.edu.ec, https://orcid.org/0000-0002-5366-5917
[2]Universidad Técnica Estatal de Quevedo, Ecuador, czambrano@uteq.edu.ec, https://orcid.org/0000-0001-8568-8024
[3]Universidad Técnica Estatal de Quevedo, Ecuador, rvillamar@uteq.edu.ec, https://orcid.org/0000-0003-2511-1789

Abstract
This study integrated bioinformatics tools and machine learning techniques to analyze transcription factors (TFs) and molecular markers in Gossypium species, focusing on their relationship with fiber quality and stress tolerance. A comparative analysis of three species—tetraploid G. hirsutum and its diploid ancestors G. arboreum and G. raimondii—led to the identification of 9,306 non-redundant TFs. In G. hirsutum, regulatory families such as MYB and bHLH were notably expanded, likely due to its polyploid nature. SSR analysis revealed species-specific patterns, with dinucleotide repeats predominating in G. raimondii and trinucleotide motifs being more frequent in G. hirsutum, suggesting divergent evolutionary pathways. Predictive modeling showed that 78% of TFs are conserved, while 2,109 clusters showed single-copy genes, indicating gene loss or functional specialization. Experimental validation confirmed the functional role of TFs such as GhHOX3 and MYB48 in fiber development, and the association of specific SSRs with differential gene expression. These findings enhance our understanding of regulatory networks in cotton and provide valuable molecular markers for breeding programs. The methodology applied highlights the potential of computational approaches to accelerate the functional characterization of candidate genes in crops, reducing the time and costs associated with traditional methods.

Keywords: *regulatory networks, gene evolution, comparative genomics, genomic selection, computational biology.*

INTRODUCTION

Cotton (*Gossypium spp.*) represents one of the most important crops globally, not only because of its importance in the textile industry, but also as a source of edible oil. However, its production faces constant challenges derived from abiotic and biotic factors, such as droughts, high temperatures, and pests, which limit its agronomic yield (Jazayeri et al., 2020; Zahid et al., 2016).

In this scenario, conventional genetic improvement has found a fundamental ally in omics sciences and bioinformatics, which allow for the more accurate identification of molecular markers and transcription factors (TFs) related to desirable characteristics, such as fiber quality or stress resistance (Li et al., 2015). The convergence between molecular biology and information and communication technologies (ICT) has transformed the way crops are studied, enabling high-resolution genomic analyses that, just a decade ago, were technically unattainable (Libbrecht & Noble, 2015).

Transcription factors are essential proteins in the regulation of gene expression, playing key roles in both plant development and response to adverse conditions (MacMillan et al., 2017). Families such as MYB, WRKY and NAC have been widely studied for their role in critical cotton processes, such as fibre formation and defence against pathogens (Shan et al., 2014; Schweizer et al., 2013). For example, TF GhHOX3, belonging to the homeobox family, has been identified as a master regulator of fiber elongation in G. *hirsutum*. Likewise, other factors such as MYC2 and WRKY70 participate in the defense against herbivorous insects, through the activation of specific biosynthetic pathways (Schweizer et al., 2013).

The development of specialized databases, such as PlantTFDB (Jin et al., 2017) and CottonGen (Yu et al., 2014), has significantly facilitated the comparative analysis of TFs in different species. These platforms allow the integration of genomic, transcriptomics and proteomic information, favoring the identification of orthologs and the characterization of their expression patterns (Wang et al., 2015). However, the growing volume of data requires advanced computational tools that automate complex tasks. In this context, applications such as OrthoVenn and MEME Suite have proven to be highly effective in detecting orthologs and regulatory motifs (Bailey et al., 2009; Wang et al., 2015).

Machine learning (ML) is emerging as a disruptive technology in the field of plant genomics. Algorithms such as *random forest* and deep neural networks make it possible to detect complex patterns in large volumes of data, facilitating the prediction of gene functions and genotype-phenotype associations with remarkable accuracy (Angermueller et al., 2016; Libbrecht & Noble, 2015). In cotton, these methodologies open the possibility of automatically identifying TFs associated with key agronomic traits, such as resistance to abiotic stress or fibre length (Lundberg et al., 2020).

On the other hand, SSRs (Simple Sequence Repeats) are consolidated as very useful molecular markers due to their high variability, genomic abundance and easy detection. Its application in genetic mapping, association studies, and assisted selection has been widely validated (Khan et al., 2016). However, the manual process of selecting relevant SSRs from thousands of possible sequences is inefficient, which justifies the use of automated computational strategies.

The synergy between bioinformatics, machine learning and comparative genomics represents an innovative avenue to accelerate cotton breeding (Zheng et al., 2016). Tools such as iTAK and GoMapMan make it possible to classify and annotate genes within functional networks, facilitating the identification of key regulators (Ramšak et al., 2014; Zheng et al., 2016). In turn, these approaches can be complemented with experimental functional validation techniques, such as gene silencing, to strengthen the predictive models generated *in silico* (MacMillan et al., 2017).

Despite these advances, challenges such as the interpretability of ML models and the integration of multi-omics data in a standardized way persist (Lundberg et al., 2020; Merchant et al., 2016). In addition, there is still a considerable gap between scientific discovery and its practical implementation in agricultural breeding programs, which requires collaborative and interdisciplinary approaches.

In this context, the present study proposes an integrative approach that combines bioinformatics, machine learning and comparative genomic analysis to identify transcription factors and SSRs with potential application in cotton breeding programs. In doing so, it seeks to contribute to a more efficient, precise and resilient agriculture in the face of the current challenges of climate change and food security.

METHODOLOGY

This study was developed under a descriptive-correlational approach, combining in silico bioinformatic analyses *with* in vitro *molecular validations*. The methodological strategy was organized into four main stages: collection and processing of genomic data, computational analysis of transcription factors (TFs) and SSR markers, implementation of machine learning models, and experimental validation of the most promising results. This comprehensive approach allowed exploring both the regulatory architecture of the cotton genome and its applicability in genetic improvement programs.

Data collection and processing

For the initial phase, the public databases PlantTFDB (Jin et al., 2017) and CottonGen (Yu et al., 2014) were used as primary sources, focusing on genomic and annotated information on G. *hirsutum*, G. *raimondii* and G. *arboreum*. In addition, transcriptomic data from the NCBI Sequence Read Archive (SRA) were included, corresponding to different tissues and stress conditions.

The cleaning and normalization of the sequences was carried out using tools such as FastQC (Andrews, 2010) and Trimmomatic (Bolger et al., 2014), ensuring high data quality. Subsequently, formats were homogenized and redundancies were eliminated through custom scripts in Python and automated workflows in Galaxy (Afgan et al., 2018).

Analysis of transcription factors and SSR markers

The identification and classification of TFs was carried out with PlantTFcat (Dai et al., 2013), while the evolutionary analysis of orthologs was carried out with OrthoVenn (Wang et al., 2015), using standard similarity parameters (e-value < 1e-5, coverage > 70%). To detect conserved motifs, MEME Suite (Bailey et al., 2009) was applied, including modules for motif discovery (MEME) and enrichment analysis (AME).

The detection of SSRs in coding and promoting regions of TFs was performed with MISA (Thiel et al., 2003), applying as a minimum threshold 5 repeats for dinucleotides and 4 repeats for tri- a hexanucleotides.

Applying machine learning models

Random forest algorithms and convolutional neural networks (CNNs), developed in TensorFlow (Abadi et al., 2016), were implemented to predict the association of TFs with agronomic characteristics. The variables used included: protein structures (domains, motifs), gene expression profiles, and presence of SSRs in regulatory regions.

The data were divided into training (70%), validation (15%) and test (15%) sets, applying data augmentation techniques to balance the classes. The importance of each variable was assessed using permutation and SHAP values (Lundberg & Lee, 2017), prioritizing the predictors with the greatest impact.

Statistical analysis and visualization

Statistical analysis was carried out in R (R Core Team, 2021) and Python. Kruskal-Wallis tests with Benjamini-Hochberg correction were used to compare the distribution of TFs between species. The associations between SSRs and phenotypic variables were evaluated by Spearman correlation and generalized linear models (GLM), with statistical significance established at $p < 0.05$.

Visualizations were generated using gene expression *heatmaps*, comparative Venn diagrams, and feature importance maps, elaborated with ggplot2 (Wickham, 2016).

Experimental validation

The most relevant TFs and SSRs were experimentally validated. Specific primers were designed using Primer3 (Untergasser et al., 2012), and gene expression analyses were performed by qRT-PCR in samples subjected to stress. Normalization was performed with internal reference genes (UBQ7 and HIS3), applying the ΔΔCt method (Livak & Schmittgen, 2001).

Likewise, a gene silencing system (VIGS) was implemented in G. *hirsutum* plants (var. TM-1) to evaluate the phenotypic effects of selected TFs. Variables such as fiber length, cellulose content, and response to water deficit were measured, using standardized protocols (Paterson et al., 2012).

Quality Control

To ensure the robustness of the results, all *in silico* analyses were repeated with slight variations in the parameters. The *in vitro* experiments included three biological replicates and at least two independent techniques per variable. The raw data and scripts were deposited in public repositories (GitHub and Zenodo) under FAIR principles (Wilkinson et al., 2016), ensuring transparency and reproducibility.

RESULTS

Identification and characterization of transcription factors in *Gossypium* spp.

Transcription factors (TFs) present in G. *arboreum*, G. *hirsutum* and G. *raimondii* were identified and classified, using the PlantTFDB, iTAK and PlantTFcat tools. OrthoVenn was then used for the analysis of orthologs and gene clusters, with a 90% similarity threshold to avoid redundancies.

Table 1. Distribution of transcription factor families in G*ossypium* spp.

TF Family	G. arboreum	G. hirsutum	G. raimondii
MYB	407	546	485
bHLH	205	272	244
AP2/ERF	252	295	272

TF Family	G. arboreum	G. hirsutum	G. raimondii
WRKY	111	151	133
Specific	136	595	385

In total, 9,306 non-redundant TFs were identified. G. *hirsutum* presented the highest number of specific factors (595), reflecting the influence of its allotetraploid nature. The MYB, bHLH and AP2/ERF families were the most represented, coinciding with patterns observed in other plant species.

This pattern of gene expansion is consistent with the history of repeated polyploidization in the genus G*ossypium* (Li et al., 2015). The predominance of MYB and bHLH is consistent with their role in metabolic pathways linked to fiber synthesis and stress response (MacMillan et al., 2017). In addition, preferential retention of genes from G. *raimondii* suggests an evolutionary bias towards the conservation of adaptive functions, in line with what was reported by Jazayeri et al. (2020) and Shan et al. (2014).

SSR patterns in TF regulatory regions

TF sequences were analyzed with MISA to detect SSRs located in coding and regulatory regions. Minimum detection criteria of 5 repeats for dinucleotides and 4 for tri- to hexanucleotides were applied.

Table 2. Distribution of SSR types in transcription factors

SSR Type	G. arboreum (%)	G. hirsutum (%)	G. raimondii (%)
Dinucleotide	28.4	31.2	47.1
Trinucleotide	65.3	62.8	45.9
Tetranucleotide	4.1	4.5	5.2
Compounds	2.2	1.5	1.8

A total of 2,109 SSRs were identified. G. *raimondii* showed the highest genomic density of SSRs (234 SSRs/Mbp), with a clear predominance of dinucleotide motifs. In contrast, G. *hirsutum* and G. *arboreum* had a higher proportion of trinucleotides, particularly the CAA motif.

The high frequency of SSRs in G. *raimondii* may be related to a higher mutation rate in its genome, as suggested by Wang et al. (2012). The abundance of trinucleotides in G. *hirsutum* could be due to their lower impact on the reading frame, favoring their persistence in coding regions (Khan et al., 2016). These findings are also consistent with the observations of Yu et al. (2014) regarding the unequal distribution of SSRs in G*ossypium* and the enrichment of CAA/CAG motifs in cell wall genes (MacMillan et al., 2017).

Regulatory networks associated with fibre production

Co-expression analysis was applied to identify clusters of TFs associated with fiber development. The data were modeled using *random forest*, considering as variables the presence of SSRs in promoter regions and the levels of gene expression.

Table 3. Key TFs associated with fiber features

TF	Family	SRH Association	Impact on fibre (cm)
GhHOX3	Homeobox	CAA(8)	+1.24
MYB48	MYB	AT(12)	+0.87
WRKY70	WRKY	TGA(6)	-0.53

17 co-expression clusters were identified, among which a central module composed of NAC43, MYB48 and WRKY70 stands out. *Random forest* models explained 68 % of the variation in fiber length ($R^2 = 0.68$, $p < 0.001$), validating the functional impact of certain TFs.

GhHOX3, with a strong positive association with fiber elongation, ratifies its key role in cell expansion (Shan et al., 2014). On the other hand, the negative influence of WRKY70 on fiber length could reflect an antagonism between growth and defense, similar to that observed in *Arabidopsis* (Schweizer et al., 2013). The hierarchical proposal for regulation with TFs such as NAC43 acting as master regulators is aligned with the model of MacMillan et al. (2017). The presence of SSRs in promoter regions reinforces the hypothesis of their involvement in differential gene expression, opening up new possibilities for marker-assisted selection (Khan et al., 2016).

CONCLUSIONS

This integrative study allowed to accurately characterize the genomic and regulatory architecture of *Gossypium* species, providing key knowledge for the design of more efficient genetic improvement strategies. The comparative analysis showed a significant expansion of transcription factors (TFs) in G. *hirsutum*, especially in the MYB, bHLH and AP2/ERF families, which reflects its history of polyploidization and the functional conservation of key genes.

In addition, differential retention of TFs from the D genome (G. *raimondii*) supports previous observations on subgenomic bias, highlighting a possible selective pressure towards the conservation of genes associated with desirable agronomic traits (Li et al., 2015; Jazayeri et al., 2020).

The analysis of SSRs revealed a species-specific distribution, with a high density of dinucleotides in G. *raimondii* and preferential trinucleotides in G. *hirsutum*. The detection of motifs such as CAA/CAG in genes linked to fiber formation reaffirms their value as functional markers, in agreement with previous work (Khan et al., 2016; Wang et al., 2012).

The co-expression and prediction models, implemented using machine learning algorithms, identified regulatory networks associated with fiber development, where TFs such as GhHOX3, MYB48 and NAC43 played a leading role. These findings expand on the hierarchical model proposed by MacMillan et al. (2017), and highlight the functional role of cis-regulatory elements, such as SSRs, in gene modulation.

From an applied perspective, this work demonstrates that the integration of bioinformatics, machine learning and experimental validation constitutes a robust methodology for the identification of candidate genes and molecular markers. The predictive models generated outperformed traditional approaches in terms of accuracy and efficiency, supporting Libbrecht and Noble's (2015) approaches to the value of artificial intelligence in plant genomics.

Finally, the experimental validation corroborated the functional role of key TFs, strengthening the link between computational prediction and field application. This approach offers a roadmap to bridge the gap between genomic exploration and its implementation in breeding programs, which could be extended to other crops of strategic interest under climate change conditions.

BIBLIOGRAPHIC REFERENCES

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265-283).
2. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., ... & 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
3. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinformatics*.
4. Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651
5. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research, Volume 37, Issue suppl_2, 1 July 2009, Pages W202–W208, https://doi.org/10.1093/nar/gkp335.*
6. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170
7. Dai, X., Sinharoy, S., Udvardi, M., & Zhao, P. X. (2013). PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC bioinformatics*, *14*(1), 321. https://doi.org/10.1186/1471-2105-14-321
8. Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., ... & Bolouri, H. (2002). A genomic regulatory network for development. *science*, *295*(5560), 1669-1678. https://doi.org/10.1126/science.1069883

9. Jazayeri, S. M., Villamar-Torres, R. O., Zambrano-Vega, C., Cruzatty, L. C. G., Oviedo-Bayas, B., Santos, M. D. A., ... & Viot, C. (2020). Transcription factors and molecular markers revealed asymmetric contributions between allotetraploid Upland cotton and its two diploid ancestors. *Bragantia*, *79*(1), 30-46. https://doi.org/10.1590/1678-4499.20190161
10. Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., & Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic acids research*, Volume 45, Issue D1, January 2017, Pages D1040–D1045, https://doi.org/10.1093/nar/gkw982.
11. Khan, M. K., Chen, H., Zhou, Z., Ilyas, M. K., Wang, X., Cai, X., ... & Wang, K. (2016). Genome wide SSR high density genetic map construction from an interspecific cross of Gossypium hirsutum× Gossypium tomentosum. *Frontiers in plant science*, *7*, 436. https://doi.org/10.3389/fpls.2016.00436
12. Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., ... & Yu, S. (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature biotechnology*, *33*(5), 524-530. https://doi.org/10.1038/nbt.3208
13. Libbrecht, M., Noble, W. (2015). Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321–332. https://doi.org/10.1038/nrg3920
14. Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. *methods*, *25*(4), 402-408. https://doi.org/10.1006/meth.2001.1262
15. Lundberg, S.M., Erion, G., Chen, H. *et al. (2020).* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67. https://doi.org/10.1038/s42256-019-0138-9
16. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
17. MacMillan, C.P., Birke, H., Chuah, A. *et al.* (2017). Tissue and cell-specific transcriptomes in cotton reveal the subtleties of gene regulation underlying the diversity of plant secondary cell walls. *BMC Genomics* **18**, 539. https://doi.org/10.1186/s12864-017-3902-4
18. Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., & Antin, P. (2016). The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS biology*, *14*(1), e1002342. https://doi.org/10.1371/journal.pbio.1002342
19. Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., ... & Schmutz, J. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*, *492*(7429), 423-427. https://doi.org/10.1038/nature11798
20. R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.
21. Ramšak, Ž., Baebler, Š., Rotter, A., Korbar, M., Mozetič, I., Usadel, B., & Gruden, K. (2014). GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology. *Nucleic acids research*, *42*(D1), D1167-D1175. https://doi.org/10.1093/nar/gkt1056
22. Schweizer, F., Fernández-Calvo, P., Zander, M., Diez-Diaz, M., Fonseca, S., Glauser, G., ... & Reymond, P. (2013), *Arabidopsis* Basic Helix-Loop-Helix Transcription Factors MYC2, MYC3, and MYC4 Regulate Glucosinolate Biosynthesis, Insect Performance, and Feeding Behavior, *The Plant Cell*, Volume 25, Issue 8, August 2013, Pages 3117–3132, https://doi.org/10.1105/tpc.113.115139
23. Shan, CM., Shangguan, XX., Zhao, B. *et al. (2014).* Control of cotton fibre elongation by a homeodomain transcription factor GhHOX3. *Nat Commun* **5**, 5519. https://doi.org/10.1038/ncomms6519
24. Thiel, T., Michalek, W., Varshney, R., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and applied genetics*, *106*(3), 411-422. https://doi.org/10.1007/s00122-002-1031-0
25. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic acids research*, *40*(15), e115-e115. https://doi.org/10.1093/nar/gks596
26. Wang, Y., Coleman-Derr, D., Chen, G., & Gu, Y. Q. (2015). OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic acids research, Volume 43, Issue W1, 1 July 2015, Pages W78–W84, https://doi.org/10.1093/nar/gkv487.*
27. Wickham, H. (2016). Data analysis. In *ggplot2: elegant graphics for data analysis* (pp. 189-201). Cham: Springer international publishing. https://doi.org/10.1111/j.1541-0420.2011.01616.x
28. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18
29. Yu, J., Jung, S., Cheng, C. H., Ficklin, S. P., Lee, T., Zheng, P., ... & Main, D. (2014). CottonGen: a genomics, genetics and breeding database for cotton research. *Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D1229–D1236, https://doi.org/10.1093/nar/gkt1064.*
30. Zahid, K. R., Ali, F., Shah, F., Younas, M., Shah, T., Shahwar, D., ... & Wu, W. (2016). Response and tolerance mechanism of cotton Gossypium hirsutum L. to elevated temperature stress: a review. *Frontiers in plant science*, *7*, 937. https://doi.org/10.3389/fpls.2016.00937
31. Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., ... & Fei, Z. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular plant*, *9*(12), 1667-1670.