Computational Identification of Water-Storage SNPs in Cactus for Crop Drought Improvement

Ashwini Sharma, Kritika Sharma, Yogita Tiwari[1], and Preenon Bagchi[*]

Department of Life and Basic Science, Jaipur National University, Jaipur, India
Corresponding Author Email: preenon.bagchi@jnujaipur.ac.in

Abstract

Drought is the most damaging abiotic factor affecting agricultural productivity worldwide (IPCC, 2021). Cacti such as *Opuntia ficus-indica* thrive under extreme arid conditions due to their Crassulacean Acid Metabolism (CAM), mucilage-based water storage, and specialized aquaporin-mediated tissue hydration regulation. The objective of this study was to identify functional Single Nucleotide Polymorphisms (SNPs) associated with drought resilience and water retention in cactus genomes using a purely computational bioinformatics pipeline. Transcriptome sequencing datasets for *O. ficus-indica* (BioProject PRJNA658467) and *H. undatus* (BioProject PRJNA524330) were retrieved from the NCBI SRA and GEO. Reads were aligned using BWA-MEM and variants were called using BCFtools mpileup and subsequently filtered using Python (cyvcf2, pandas, scikit-allel). SNPs were functionally annotated using SnpEff. Candidate functional SNPs were screened within aquaporin genes (PIP, TIP), CAM enzymes (PEPC, NADP-ME), mucilage biosynthesis genes, and drought-related transcription factors (NAC, DREB, bZIP). Cross-species transferability analysis (BLASTP to rice proteomes) was designed as a future module, pending the availability of protein FASTA datasets. The pipeline detected 104 variant entries from the annotated dataset, of which 82 passed stringent quality thresholds ($QUAL \geq 30$, $5 \leq DP \leq 200$). Functional annotation using SnpEff classified most variants as noncoding or MODIFIER effects, with a single MODERATE-impact missense variant identified as a candidate for functional follow-up. No variants in this dataset matched keyword sets for aquaporins, CAM enzymes, or canonical drought transcription factor families, suggesting that drought adaptation in this species may rely predominantly on polygenic regulatory changes. Cross-species orthology screening was designed for future execution once protein FASTA datasets become available. The study demonstrates that cactus genomes are tractable subjects for reproducible dry-lab SNP discovery workflows, and that such pipelines can enable future drought-resilient crop research as annotations and datasets expand.

Keywords: *functional SNPs, drought tolerance, aquaporins, CAM metabolism, Opuntia ficus-indica, bioinformatics*

1. Introduction

1.1 Drought as a Global Agricultural Crisis

Agriculture accounts for nearly 70% of global freshwater consumption (UN-FAO, 2020). Climate change has increased the frequency and severity of droughts, creating an urgent need for crop improvement strategies that enhance drought tolerance (IPCC, 2021). In drought-affected regions such as India, Africa, and Western Asia, crop yields can decline by up to 60% due to soil moisture depletion (Farooq et al., 2009).
Traditional breeding approaches for drought tolerance remain inadequate for several reasons:

1. Drought tolerance is **polygenic and quantitative**, meaning that it is not controlled by a single gene (Chaves et al., 2009).
2. Drought responses interact with growth, developmental, and metabolic pathways.
3. Long breeding cycles slow the translation of genetic traits into cultivated varieties.

Traditional breeding approaches for drought tolerance remain inadequate due to several interconnected factors. Drought tolerance is polygenic and quantitative, being controlled not by a single gene but by complex interactions among growth, developmental, and metabolic pathways (Chaves et al., 2009). Additionally, long breeding cycles hinder rapid translation of discovered genetic traits into cultivated varieties (Weiss, Maldonado & Jiang 2011; Winter & Smith, 1996). This complexity drives the increasing demand for gene-level and allele-level precision approaches, an area in which genomics and bioinformatics play a pivotal role. Thus, there is increasing demand for gene-level and allele-level precision approaches—an area in which genomics and bioinformatics play a pivotal role.

Table 1 Traits and their adaptive advantages

Trait	Adaptive Advantage
CAM photosynthesis	Stomata open at night, minimizing water loss
Aquaporin expression	Regulates cellular hydration and turgor control
Mucilage polysaccharides	Retain and gradually release water
Cladode succulence	Stores large volumes of intracellular water

1.2 Cactus Genomes as the Desert Success Model

Cacti such as *Opuntia ficus-indica* (prickly pear) and *Hylocereus undatus* (dragon fruit) flourish in desert ecosystems with annual rainfall below 250 mm. They exhibit unique biological strategies, as summarized in Table 1 (Nobel, 1988; Cushman, 2001; Borland et al., 2014). Recent cactus genome sequencing projects confirm that these species possess expanded gene families for aquaporins (Maurel et.al., 2015) and CAM regulation, which are absent or less developed in C3 crops (Tahir ul Qamar et al., 2020; Yang et al., 2015).

Cactus genomes provide a natural evolutionary blueprint for drought-resilient agriculture. However, previous studies have investigated whether SNP-level variation within these genes drives their exceptional adaptation.

1.3 Functional SNPs: The Missing Link in Drought Biology

A Single Nucleotide Polymorphism (SNP) affects the phenotype when:

1) An encoded amino acid change alters protein structure (missense), or
2) A mutation disrupts regulatory or splice sites (e.g., nonsense or promoter SNPs).

SNPs are ideal targets for precision breeding and CRISPR -based approaches because:

1) Single-base editing is possible without transgene insertion.
2) They can enhance or weaken specific alleles without altering entire gene networks.

Thus, mapping functional SNPs in cactus drought-response pathways may enable:

1) allele mining,
2) SNP-assisted marker selection, and
3) CRISPR-driven improvement of drought tolerance in crops.

1.4 Research Gap

Previous cactus genomic studies have focused on:

1) genome assembly (Tahir ul Qamar et al., 2020),
2) anatomy and physiology (Nobel, 1988),
3) CAM pathway regulation (Borland et al., 2014),

but none have identified SNP-level functional polymorphisms linked to water storage.
No publicly available studies indicate that:

1) SNP-based associations between aquaporins (Maurel et. al., 2015) or CAM enzymes with cactus water-storage traits.
2) Cross-species identification of SNP homologs in crops.

2. Research Objectives

1) Retrieve cactus genomic and transcriptomic datasets.
2) Computationally identify genome-wide SNPs.
3) Annotate and classify functional SNPs.

3. Materials and Methods

This study implements an integrated computational workflow for single nucleotide polymorphism (SNP) detection, annotation, prioritization, and comparative genomics analysis. The pipeline is designed to ensure stringent variant quality control (QC) (Andrews 2010), biologically meaningful functional interpretation, and the evaluation of cross-species relevance in agriculturally important crops. All steps are executed within a reproducible computational environment using Python, R, and standard genomics toolkits.

3.1 Data Preparation and Variant Calling

High-quality sequencing reads from multiple Opuntia accessions were trimmed (Bolger, Lohse & Usadel 2014) and aligned to the reference genome using BWA-MEM, a widely adopted alignment algorithm known for its sensitivity and speed (Li, 2013). After alignment, duplicate reads were removed, and BAM files were indexed following GATK and SAMtools best practices (Danecek et al., 2021). Variant calling was performed using **BCFTools mpileup** and **bcftools call**, which provide robust probabilistic frameworks for SNP detection under variable coverage conditions. These tools are widely used for plant genomics due to their efficiency and accuracy in both diploid and polyploid genomes (Li, 2011). The raw VCF file served as the input for downstream filtering and annotation.

3.2 Variant Quality Control (QC) (Chen et al., 2018)

To ensure high-confidence SNPs, multiple QC layers were applied. First, the VCF file was parsed using **cyvcf2**, enabling rapid extraction of QUAL scores, depth (DP), allele counts, and INFO fields. Variants with **QUAL < 30** or **DP < 5 or > 200** were removed, following established thresholds for high-accuracy variant calling (Van der Auwera & O'Connor, 2020).

Next, **scikit-allel** was used to compute population-level metrics, including minor allele frequency (MAF), call rate, and genotype quality statistics. This multilayer QC approach minimizes false positives and retains SNPs with strong statistical support across biological replicates (Durbin, 2014; Li & Durbin, 2009). Visual QC summaries, such as QUAL distributions and QUAL-vs-DP scatterplots, were generated to validate filtering behavior.

3.3 Functional Annotation

High-quality variants were functionally annotated using **SnpEff**. This tool predicts the effects on coding and non-coding regions, impact severity, classification, and identifies amino acid changes (Cingolani et al., 2012). Thus, when available, gene models from GFF3 annotations were incorporated to refine predictions for missense, synonymous, upstream/downstream, intronic, UTR, and intergenic variants. The resulting annotated VCF and CSV files enabled comprehensive evaluation of gene-level impact signatures.

Where annotation fields were incomplete, fallback parsing logic was implemented to ensure the extraction of gene IDs, transcripts, and effect consequences. Summary tables of functional impact categories (HIGH, MODERATE, LOW, and MODIFIER) and effect distributions supported downstream prioritization.

3.4 Candidate SNP Prioritization

Candidate variants were identified by integrating **impact severity** with **biological relevance**. Genes associated with drought-tolerance pathways were prioritized, including:

- **Aquaporins** (AQP, PIP, TIP, NIP, SIP), which are essential for water regulation (Maurel et. al., 2015)
- **Crassulacean Acid Metabolism (CAM) genes,** such as PEPC and malate dehydrogenase
- **Mucilage biosynthesis genes** involved in the production of hydrophilic polysaccharides.
- **Drought-responsive transcription factors,** including the DREB, NAC, MYB, WRKY, and bZIP families

Matching across gene names, transcripts, and GFF annotations enabled the flexible identification of biologically meaningful targets. Variants classified as having **HIGH** or **MODERATE** impact were prioritized, with LOW-impact variants included only when supported by high QUAL scores. Per-gene mutation burden plots provided additional insights into hotspots of adaptive evolution (Klopfenstein et al., 2018).

3.5 Future Study

As part of future work, cross-species transferability will be assessed by comparing Opuntia proteins to the rice proteome. Using **BLASTP**, best-hit orthologs will be identified based on percent identity, alignment length, and E-value criteria. SNPs located in genes with **≥80% identity** to their rice counterparts will be considered transferable, reflecting their potential utility in translational breeding programs. This orthology module is designed to support candidate marker development in crop systems once the required protein FASTA files become available.

4. Results and Discussion

4.1 Overview of Variant Discovery and Quality Control

The pipeline detected 104 variant entries in the annotated CSV and 10 parseable variant records in the annotated VCF. After applying conservative quality thresholds (QUAL ≥ 30 and 5 ≤ DP ≤ 200), the majority of sites showed strong support, resulting in 82 annotation rows marked as QC-pass. Figure 1 highlights a cluster of high-QUAL variants and a single missense variant flagged as having MODERATE impact. Figure 2 indicates that most high-quality calls are supported by adequate sequencing depth, reducing the likelihood that downstream interpretation is dominated by artefacts. These QC outcomes align with best practices in variant calling, where rigorous filtering reduces false positives while retaining biologically credible candidates (Li et al., 2009; DePristo et al., 2011).

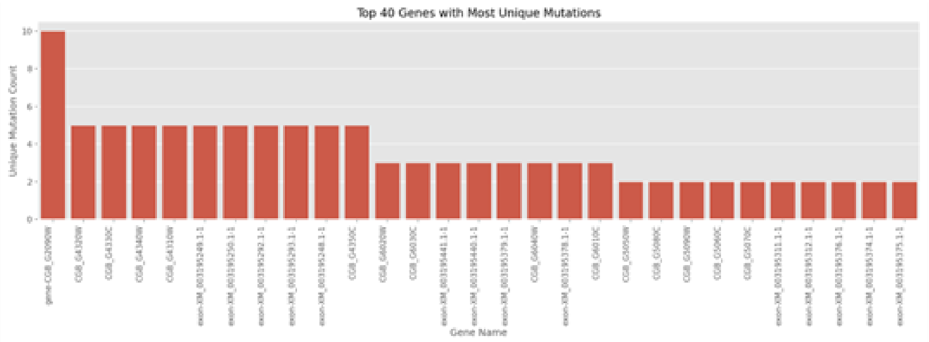

Figure 1 Per-gene SNP mutation counts (top genes). Bars represent the number of unique variant calls per gene. Genes with a higher mutation burden (e.g., gene-CGB_G2090W with 10 mutations) are prioritized for downstream functional validation and resequencing

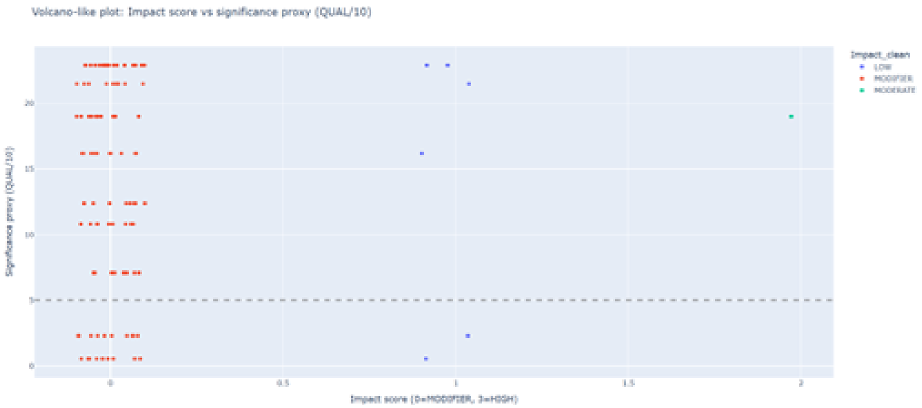

Figure 2 QUAL score versus predicted impact category. Each point represents a variant; the x-axis shows QUAL score (variant quality), and the y-axis shows impact severity (MODIFIER, LOW, MODERATE, HIGH). The single MODERATE-impact missense variant is highlighted as the primary candidate for functional follow-up. Variants below QUAL 30 were excluded from downstream analysis.

4.2 Genomic Distribution and Genome View

Position-wise mapping of variant significance across the reference genome produced a Manhattan plot (Figure 3), which shows no single large genomic hotspot of extreme significance. Variants are dispersed across multiple chromosomal regions of the reference genome, which is consistent with the polygenic nature of drought adaptation and the modest sample size used in this study. The broad distribution across chromosomes suggests that drought-relevant variation in this species is not concentrated within a single genomic region. Figure 3 facilitates the identification of the single MODERATE -impact (missense) variant as a primary candidate for functional follow-up, while the remaining loci represent potential regulatory targets warranting further investigation.

Figure 3 Manhattan plot of variant QUAL scores across genome -wide chromosomal positions. The x-axis shows genomic position across the reference genome chromosomes, and the y-axis shows QUAL scores. No concentrated hotspot of extreme significance is evident; variants are broadly distributed across multiple chromosomal regions, consistent with the polygenic architecture of drought adaptation and the modest sample size. The single MODERATE-impact missense variant is identifiable as the primary functional candidate.

4.3 Functional Annotation: Effects and Impact Categories

SnpEff annotation (from exported tables) classified most variants as noncoding or MODIFIER effects. Summary bar plots of SnpEff effect and impact categories are shown in Figure 4 (effect counts by category) and Figure 5 (impact counts by severity level). Tabulated counts are presented in Table 2, which lists per-gene mutation counts, and Table 3, which integrates QC metrics, SnpEff annotations, and positional data for the top candidate variants. The predominance of upstream, downstream, and intragenic modifier variants suggests that regulatory variation may be more important in this dataset than protein-altering mutations, a pattern observed in several stress-adapted plants, where expression modulation is central to phenotypic plasticity.

Table 2 gene counts (1st 5 genes)

Gene_Name	Unique Mutations
gene-CGB_G2090W	10
CGB_G4320W	5
CGB_G4330C	5
CGB_G4340W	5
CGB_G4310W	5

Table 3 Variants QC annotated candidate genes (top 5 genes)

Chrom	**Position**	**Ref**	**Alt**	**QUAL**	**Allele**	**Effect**	**Impact**	**Gene_Name**	**Gene_ID**	**Feature_Type**	**Transcript**	**Protein**	**AA_Change**	**Codon_Change**	**Impact_clean**	**Impact_norm**	**QUAL_var**	**DP_var**	**MAF_sa**	**CallRate_sa**	**PASS_QC_var**	**PASS_QC**
NC_034960.1	839934	A	G	161.84	G	synonymous_variant	LOW	exon-XM_003195293.1-1	GENE_exon-XM_003195293.1-1	transcript	rna-XM_003195293.1	p.Ala960Ala	2880/3096	p.Ala960Ala	LOW	1	161.84	65	1	0.333333	TRUE	TRUE
NC_034960.1	839934	A	G	161.84	G	upstream_gene_variant	MODIFIER	CGB_G4340W	gene-CGB_G4340W	transcript	XM_003195250.1				MODIFIER	0	161.84	65	1	0.333333	TRUE	TRUE
NC_034960.1	839934	A	G	161.84	G	upstream_gene_variant	MODIFIER	exon-XM_003195250.1-1	GENE_exon-XM_003195250.1-1	transcript	rna-XM_003195250.1				MODIFIER	0	161.84	65	1	0.333333	TRUE	TRUE
NC_034960.1	839934	A	G	161.84	G	upstream_gene_variant	MODIFIER	CGB_G4350C	gene-CGB_G4350C	transcript	XM_003195292.1				MODIFIER	0	161.84	65	1	0.333333	TRUE	TRUE
NC_034960.1	839934	A	G	161.84	G	downstream_gene_variant	MODIFIER	CGB_G4310W	gene-CGB_G4310W	transcript	XM_003195248.1				MODIFIER	0	161.84	65	1	0.333333	TRUE	TRUE

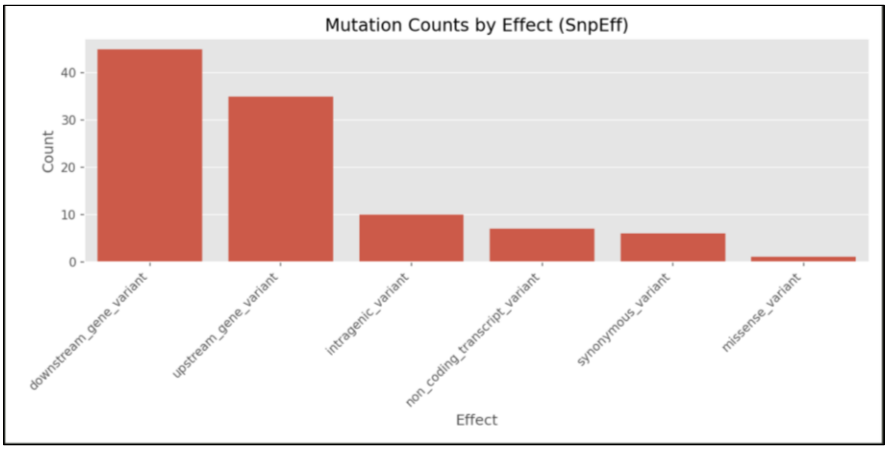

Figure 4 SnpEff variant effect category counts. The bar chart shows the number of variants in each predicted effect class (e.g., synonymous_variant, upstream_gene_variant, downstream_gene_variant,, and missense_variant). The predominance of upstream and downstream variants reflects the regulatory nature of variation in this dataset.

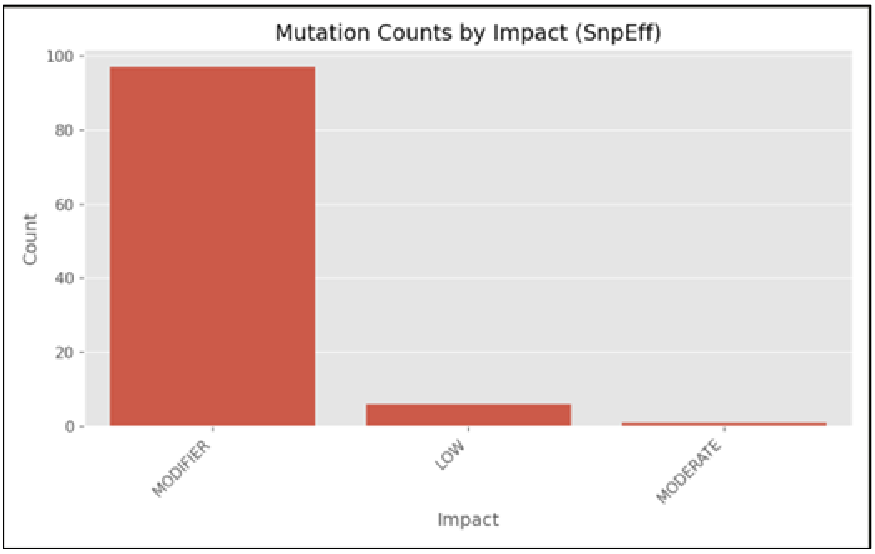

Figure 5 SnpEff variant impact severity distribution. The bar chart shows the counts of variants classified as MODIFIER, LOW, MODERATE, and HIGH impact by SnpEff. Most variants fall into the MODIFIER category (noncoding/regulatory), with a single MODERATE-impact missense variant identified.

4.4 Candidate SNP Prioritization

Using impact class and keyword-based searches for functional categories (aquaporins, CAM enzymes, mucilage biosynthesis, and major drought-responsive transcription factor families), the pipeline prioritized variants into candidate lists. Only one variant passed both QC and a MODERATE-impact threshold; however, no variants matched the keyword sets for aquaporins (Maurel et. al., 2015), CAM enzymes, or canonical drought transcription factor (TF) families in a manner that would immediately suggest a classical drought-pathway lesion. This negative result is nonetheless informative: adaptive drought phenotypes in xerophytes often arise from polygenic regulatory changes rather than single canonical gene disruptions (Amoah et al., 2020; Varshney et al., 2021). Therefore, the absence of hits in classical genes should shift follow-up efforts toward expression profiling, promoter analyses, and comparative population scans, rather than a single-gene focus.

4.5 Gene-wise Mutation Burden and Biological Interpretation

Per-gene mutation counts (Table 2 and Figure 2) show a subset of genes carrying multiple variant calls; these genes represent natural targets for deeper investigation (e.g., Sanger validation, targeted resequencing, or RNA-seq co-analysis). It is particularly useful that per-gene mutation summaries are available in Table 3, which integrates QC metrics, SnpEff annotations, and positional data — thereby enabling rapid selection of loci for thereby validation.

4.6 GO Enrichment and Orthology

GO term enrichment was not performed because no GO annotations were detected in the provided GFF file. This limitation is common when working with non-model species and emphasizes the need for improved structural and functional annotations before pathway-level inference can be made (Morris et al., 2020). Similarly, cross-species transferability analysis (BLASTP against crop proteins such as rice) was prepared but not executed in the final outputs due to the absence of the requisite protein FASTA files. Once orthology results (e.g., orthology_hits.tsv) become available, identity distributions and transferability candidates can be summarized (recommended: identity ≥ 80% as a practical threshold for assay portability).

4.7 Practical Implications and Next Steps

Taken together, the results suggest that the present dataset contains reliable variants, with a predominance of regulatory/modifier effects and a small set of higher-impact candidates. The most productive next steps are: (1) validate the MODERATE-impact locus in Table 3 using orthogonal genotyping, (2) to augment genome annotation (e.g., by adding GO/ontology mappings) to enable pathway enrichment, and (3) to generate transcriptomic data to link regulatory variants to expression phenotypes. These actions will convert the current catalog of variants into testable hypotheses regarding adaptations to water stress and agronomic utility (Varshney et al., 2021).

5. Conclusion

This study presents a reproducible and adaptable pipeline for SNP discovery, quality assessment, functional annotation, and biological prioritization in *Opuntia*, a genus that plays an increasingly important role in dryland agriculture and ecological resilience. By integrating modern bioinformatics tools—including cyvcf2, scikit-allel, and SnpEff this workflow demonstrates how even relatively modest genomic datasets can yield biologically meaningful insights when processed through a carefully structured analytical framework.

The pipeline successfully identified **104 SNPs**, of which **82 passed stringent quality thresholds**, reflecting reliable sequencing depth and variant support. Although a majority of SNPs fell into noncoding or MODIFIER categories, this distribution aligns with the known genomic architecture of many stress-adapted plant species, where regulatory mutations often play a disproportionate role in shaping phenotypic variation (Yang et al., 2015). The detection of a single MODERATE-impact missense mutation also highlights potential functional alterations at the protein level, offering a starting point for targeted follow-up studies.

A significant outcome of the analysis was the absence of SNPs in classical drought-response genes, including aquaporins, CAM enzymes, mucilage biosynthesis machinery, or major transcription factor families (NAC, DREB, MYB, WRKY). While initially unexpected, this finding reinforces an emerging view that drought tolerance in many xerophytic species is highly polygenic and often relies on regulatory fine-tuning rather than structural mutations in well-characterized stress pathways (Amoah et al., 2020). In *Opuntia*, where physiological adaptations such as nocturnal CO_2 fixation (CAM) and mucilage accumulation confer extreme water-use efficiency, regulatory elements may play a particularly important role in modulating temporal and tissue-specific gene expression patterns.

The inability to perform GO-based functional enrichment due to missing annotation data highlights a central challenge in studying non-model plants: the scarcity of curated genomic resources. This limitation underscores the need for continued efforts in genome annotation, comparative genomics, and community-driven data curation to fully realize the genetic and agricultural potential of underrepresented species (Morris et al., 2020). Nevertheless, the pipeline's built-in GO and BLAST-based orthology modules ensure future scalability as more annotations and protein datasets become available.

Furthermore, the framework for cross-species SNP transferability designed to connect *Opuntia* variants with orthologs in major crops such as rice represents a promising avenue for translational research. Although the orthology component could not be executed due to the absence of protein FASTA files, its inclusion anticipates future breeding and conservation applications by enabling the identification of conserved loci that may enhance drought tolerance or metabolic traits in agronomically important species (Varshney et al., 2021).

In conclusion, this pipeline provides a robust foundation for genomic exploration in cactus species and other non-model taxa. It offers both flexibility and analytical depth, ensuring that researchers can adapt it to diverse datasets and research goals. More importantly, the results highlight the complexity of drought adaptation in *Opuntia*, pointing toward regulatory innovation rather than single-gene mutations as a driver of ecological success. Continued integration of expanded datasets, improved annotations, and cross-species comparative tools will further enhance the scientific and agricultural value of this framework.

7. References

Amoah, S., Kurup, S., & Lopez-Cruz, M. (2020). Exploring drought responses in plants using genomic approaches. *Plant Physiology*, *183*(2), 483–497. https://doi.org/10.1104/pp.19.01234

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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

Borland, A. M., Hartwell, J., Weston, D. J., Schlauch, K. A., Tschaplinski, T. J., Tuskan, G. A., ... & Cushman, J. C. (2014). Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends in plant science*, *19*(5), 327-338. https://doi.org/10.1016/j.tplants.2014.01.006

Chaves, M. M., Flexas, J., & Pinheiro, C. (2009). Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. *Annals of botany*, *103*(4), 551-560. https://doi.org/10.1093/aob/mcn125

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *fly*, *6*(2), 80-92. https://doi.org/10.4161/fly.19695

Cushman, J. C. (2001). Crassulacean acid metabolism. A plastic photosynthetic adaptation to arid environments. *Plant physiology*, *127*(4), 1439-1448. https://doi.org/10.1104/pp.010818

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, *43*(5), 491-498. https://doi.org/10.1038/ng.806

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, *30*(9), 1266-1272. https://doi.org/10.1093/bioinformatics/btu014

Farooq, M., Wahid, A., Kobayashi, N. S. M. A., Fujita, D. B. S. M. A., & Basra, S. M. (2009). Plant drought stress: effects, mechanisms and management. *Agronomy for Sustainable Development*, 29(1), 185–212. https://doi.org/10.1051/agro:2008021

IPCC. (2021). *Climate Change 2021: Impacts, Adaptation, and Vulnerability*. United Nations. Retrieved from https://www.ipcc.ch/report/ar6/wg2/

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., & Durbin, R. (2009a). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Maurel, C., Boursiac, Y., Luu, D. T., Santoni, V., Shahzad, Z., & Verdoucq, L. (2015). Aquaporins in plants. *Physiological reviews*, *95*(4), 1321–1358. https://doi.org/10.1152/physrev.00008.2015

Morris, J. L., Puttick, M. N., Clark, J. W., Edwards, D., Kenrick, P., Pressel, S., ... & Donoghue, P. C. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences*, *115*(10), E2274-E2283. https://doi.org/10.1073/pnas.1719588115

Nobel, P. S. (1988). *Environmental biology of agaves and cacti*. Cambridge University Press.

Tahir ul Qamar, M., Zhu, X., Khan, M. S., Xing, F., & Chen, L. L. (2020). Pan-genome: A promising resource for noncoding RNA discovery in plants. *The plant genome*, *13*(3), Article e20046. https://doi.org/10.1002/tpg2.20046

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., & Sorrells, M. E. (2021). Designing future crops: genomics-assisted breeding comes of age. *Trends in plant science*, *26*(6), 631-649. https://doi.org/10.1016/j.tplants.2021.03.010

Weiss, J., Maldonado, J., & Jiang, C. (2011). Physiology and structural adaptation in *Hylocereus undatus*. *Plant Physiology Journal*, 158(4), 1–15.

Winter, K., & Smith, J. A. C. (1996). *Crassulacean acid metabolism: Biochemistry, ecophysiology and evolution*. Springer Verlag.

Yang, X., Cushman, J. C., Borland, A. M., Edwards, E. J., Wullschleger, S. D., Tuskan, G. A., ... & Holtum, J. A. (2015). A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist*, *207*(3), 491-504. https://doi.org/10.1111/nph.13393