UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

ALEXANDRA RUSSOLO CARDELLI

BIOINFORMATICS ANALYSIS OF AGAVE SPECIES: EXPLORING GENETIC DIVERSITY AND ADAPTIVE TRAITS FOR SUSTAINABLE BIOFUEL PRODUCTION

ANÁLISE BIOINFORMÁTICA DE ESPÉCIES DE AGAVE: EXPLORANDO A DIVERSIDADE GENÉTICA E CARACTERÍSTICAS ADAPTATIVAS PARA A PRODUÇÃO SUSTENTÁVEL DE BIOCOMBUSTÍVEIS

CAMPINAS
2025

ALEXANDRA RUSSOLO CARDELLI

BIOINFORMATICS ANALYSIS OF AGAVE SPECIES: EXPLORING GENETIC DIVERSITY AND ADAPTIVE TRAITS FOR SUSTAINABLE BIOFUEL PRODUCTION

ANÁLISE BIOINFORMÁTICA DE ESPÉCIES DE AGAVE: EXPLORANDO A DIVERSIDADE GENÉTICA E CARACTERÍSTICAS ADAPTATIVAS PARA A PRODUÇÃO SUSTENTÁVEL DE BIOCOMBUSTÍVEIS

Dissertation presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Genetics and Molecular Biology, with focus on Bioinformatics

Dissertação apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do Título de Mestra em Genética e Biologia Molecular, na área de Bioinformática

Orientador: PROF. DR MARCELO FALSARELLA CARAZZOLLE

Co-Orientador: DR. LUCAS MIGUEL DE CARVALHO

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA ALEXANDRA RUSSOLO CARDELLI E ORIENTADA PELO MARCELO FALSARELLA CARAZZOLLE

CAMPINAS

2025

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

Cardelli, Alexandra Russolo, 1992-
C178b Bioinformatics analysis of agave species : exploring genetic diversity and adaptive traits for sustainable biofuel production / Alexandra Russolo Cardelli. – Campinas, SP : [s.n.], 2025.

Orientador: Marcelo Falsarella Carazzolle.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Biologia.

1. Bioinformática. 2. Genômica comparativa. 3. Polimorfismo de nucleotídeo único. 4. Metagenômica. 5. Mutação Indel. 6. Agave. I. Carazzolle, Marcelo Falsarella, 1975-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Biologia. III. Título.

Informações complementares

Título em outro idioma: Análise bioinformática de espécies de agave : explorando a diversidade genética e características adaptativas para a produção sustentável de biocombustíveis
Palavras-chave em inglês:
Bioinformatics
Comparative genomics
Single nucleotide polymorphism
Metagenomics
Indel mutation
Agave
Área de concentração: Bioinformática
Titulação: Mestra em Genética e Biologia Molecular
Banca examinadora:
Marcelo Falsarella Carazzolle [Orientador]
Jorge Maurício Costa Mondego
Elisson Antonio da Costa Romanel
Data de defesa: 20-02-2025
Programa de Pós-Graduação: Genética e Biologia Molecular

Objetivos de Desenvolvimento Sustentável (ODS)
ODS: 7. Energia acessível e limpa
ODS: 1. Erradicação da pobreza

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-8951-9594
- Currículo Lattes do autor: http://lattes.cnpq.br/7837704663093434

Prof. Dr. Marcelo Falsarella Carazzolle

Prof. Dr. Jorge Maurício Costa Mondego

Prof. Dr. Elisson Romanel

Prof. Dr. Renato Vicentini dos Santos

Prof. Dr. Antonio Figueira

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa Mestrado em Genética e Biologia Molecular do Instituto de Biologia (IB/Unicamp) na Universidade Estadual de Campinas.

AGRADECIMENTOS

Primeiramente, agradeço à minha família: meus pais, minha irmã, minhas sobrinhas Manu e Elena, minhas primas e minhas tias. Agradeço também aos amigos próximos que conheci ao longo dos anos e que sempre me apoiaram: Dora, Dan, Ana David, Lethicia e todos os demais.

Agradeço ainda à Shell Brasil e à ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) pelo apoio estratégico, por meio de incentivos regulatórios para Pesquisa, Desenvolvimento e Inovação. Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

E, finalmente, agradeço ao Prof. Dr. Gonçalo Pereira e aos meus orientadores: Lucas Carvalho, que esteve comigo desde o início da minha iniciação científica e me apoiou em todos os momentos; e Tchelos, meu melhor amigo do LGE, que mais me apoiou e me orientou ao longo deste Mestrado.

RESUMO

O uso de biocombustíveis como estratégia de curto prazo tem sido uma ótima estratégia para diminuir a dependência de combustíveis fósseis para geração de energia. A biomassa tem um papel fundamental nessa cadeia de produção e a expansão de plantações de culturas bioenergéticas para regiões semiáridas é essencial para a aceitação dessa estratégia mundialmente. Nesse contexto, é importante estudar culturas que podem prosperar em áreas semiáridas com altas temperaturas, alta insolação e baixa disponibilidade de água, especialmente sem irrigação, como é o caso das agaves. No Brasil, as agaves são cultivadas principalmente por sua fibra de sisal, que é extraída das folhas de certas espécies, como *Agave sisalana* e genótipos híbridos (H11648 e H400L). *A. tequilana* é outra espécie importante amplamente usada para bebidas alcoólicas (tequila, mezcal etc.) no México e notável por sua adequação na produção de biocombustíveis devido ao seu alto teor de açúcar e ciclo de vida mais curto. No contexto deste projeto de mestrado, usamos genômica populacional, análise de SNPs/Indels e genômica comparativa para entender os processos biológicos subjacentes aos mecanismos relacionados aos fenótipos de interesse para a produção, como produtividade de biomassa, resistência a doenças, tolerância à seca e teor de açúcar/fibra, que podem ser alvos para engenharia genética e melhoria da produtividade das culturas. O sequenciamento ddRAD e a análise de 10 indivíduos H11648, 8 H400L e 77 A. sisalana revelaram 15.887 SNPs e 5.302 Indels para H11648; 15.105 SNPs e 5.235 Indels para H400L; e 12.168 SNPs e 9.003 Indels para *A. sisalana*. As análises genômicas populacionais mostraram uma clara diversidade genética entre H11648 e H400L e entre as subpopulações de *A. sisalana* de 3 diferentes regiões produtoras. O enriquecimento de GO identificou processos biológicos relacionados aos processos catabólicos de glutationa, resposta de defesa a fungos, respostas celulares à privação de ferro e resposta à estrigolactona. Na análise genômica comparativa, identificamos 28.719 ortogrupos, dos quais 6.818 específicos de cada espécie e 9.460 ortogrupos com todas as espécies presentes. Para a análise de famílias gênicas, obtivemos 1.545 famílias expandidas e 2.788 famílias contraídas para *A. sisalana*; 3.515 famílias expandidas e 751 famílias contraídas para H11648; 1.678 famílias expandidas e 856 famílias contraídas para *A. tequilana*. A análise de GO mostrou ortogrupos exclusivos e famílias gênicas associados aos fenótipos de interesse para produção. Nossas descobertas aumentam nossa compreensão sobre

Agaves e nos dão uma base para uma possível engenharia genética que visa melhorar a produtividade das culturas, contribuindo para o avanço da produção sustentável de biocombustíveis e reduzindo a dependência de combustíveis fósseis.

ABSTRACT

The use of biofuels as a short-term strategy has been a great strategy to diminish the reliance on fossil fossil fuels for energy generation. Biomass has a fundamental role in this production chain and the expansion of bioenergetic crops plantations to semiarid regions is a way of increasing the production of biofuels around the world. In this context, it is important to study crops that can thrive in such semiarid areas with high temperatures, high insolation, and low water availability, especially without irrigation, as is the case for agaves, drought-resistant plants. In Brazil, agaves are primarily grown for their sisal fiber, which is extracted from the leaves of certain species, like *Agave sisalana* and Hybrid genotypes (H11648 and H400L). *A. tequilana* is another important species that is widely used for sprits (tequila, mezcal etc) in Mexico and notable for its suitability in biofuel production due to its high sugar content and shorter life cycle. In the context of this master's project, we used population genomics, SNPs/Indels analysis and comparative genomics to understand the biological processes underlying mechanisms related to phenotypes of interest for production, such as biomass productivity, disease resistance, drought tolerance and sugar/fiber content, which can be targets for further genetic engineering and crop productivity improvement. ddRAD sequencing and analysis of 10 H11648, 8 H400L, and 77 *A. sisalana* individuals revealed 15,887 SNPs and 5,302 Indels for H11648; 15,105 SNPs and 5,235 Indels for H400L; and 12,168 SNPs and 9,003 Indels for *A. sisalana*. Population genomics analyses showed a clear genetic diversity between H11648 and H400L and between the subpopulations of *agave sisalana* from 3 different production regions. GO enrichment identified biological processes related to glutathione catabolic processes, defense response to fungus, cellular responses to iron starvation, and response to strigolactone. In the comparative genomics analysis, we identified 28,719 orthogroups, of which 6,818 species-specific and 9,460 orthogroups with all species present. For the gene families analysis, we obtained 1545 expanded families and 2788 contracted families for *A. sisalana*; 3515 expanded families and 751 contracted families for H11648; 1678 expanded families and 856 contracted families for *A. tequilana*. The GO analysis showed exclusive orthogroups and gene families associated with the phenotypes of interest for production. Our findings enhance our understanding of Agave and give us a base to future genetic engineering aimed at improving crop productivity, contributing to the advancement of sustainable biofuel production and reducing reliance on fossil fuels.

SUMMARY

1. INTRODUCTION 11
1.1 Characteristics of Agave 11
1.2 Agave as biofuel feedstock 12
1.3 Important Phenotypes of Agave species 13
1.4 Genetics 16
1.5 Reproduction and Propagation of Cultivated Agave Species 17
1.6 Bioinformatics analysis for plant breeding 18
2. MOTIVATION 19
3. CHAPTER I: ANALYSIS OF POPULATION GENOMICS AND NON- SYNONYMOUS SNPS RELATED TO HIGHLY PRODUCTIVE AGAVE GENOTYPES 21
3.1 Introduction 21
3.1.1 Molecular markers 21
3.1.2 SNPs and Indels 22
3.1.3 Population Genomics 23
3.2 Objectives 25
3.2.1 Specific Objectives: 25
3.3 Material and Methods 25
3.3.1 ddRAD-seq of fiber-producing cultivars 25
3.3.2 Publicly available genome data 27
3.3.3 Alignment and Variant Calling 27
3.3.4 Population Genomics 28
3.3.4.1 Neutral SNPs 29
3.3.4.2 Population Structure Analysis 29
3.3.4.3 Phylogenomics 29
3.3.5 Selection of exclusive and high-impact SNPs and Indels in H400L, H11648 and A. sisalana populations 30
3.3.6 Annotation and Gene Ontology (GO) 31
3.4. Results 32
3.4.1 Alignment and Variant Calling 32
3.4.2 Population Genomic Analysis of H400L and H11648 33
3.4.3 Population genomic analysis of A. sisalana individuals from different production regions 35
3.4.4 Functional Annotation and Genotype Comparison 38
3.4.5 GO Enrichment of exclusive gene lists 40
3.5. Conclusion 48
4. CHAPTER II: COMPARATIVE GENOMICS BETWEEN HIGH FIBER CONTENT AGAVES (H11648 AND A. SISALANA) AND HIGH SUGAR CONTENT AGAVES (A. TEQUILANA) 50
4.1 Introduction 50
4.1.1 Plant Genomes 50
4.1.2 Comparative Genomics 51
4.2 Objectives 52
4.2.1 Specific Objectives: 53
4.3 Material and Methods 53

4.3.1 Data and Publicly available genome data 53
4.3.2 Genome Assembly 53
4.3.3 Genome Structural Annotation 53
4.3.4 Comparative and evolutionary analysis 54
4.3.5 Expanded and contracted gene families 54
4.3.6 Functional Annotation and Gene Ontology (GO) 54
4.4 Results 55
4.4.1 Genome Assembly 55
4.4.2 Genome Annotation 55
4.4.3 Comparative Genomics 56
4.4.3.1 Orthofinder Results 56
4.4.3.2 Exclusive Orthologs 57
4.4.3.3 Expanded and contracted families 62
4.5 Conclusion 69
5. GENERAL DISCUSSION AND CONCLUSION 70
6. REFERENCES 73
7. SUPPLEMENTARY MATERIAL 85
ANEXOS 107

1. INTRODUCTION

The use of fossil fuels to obtain energy is responsible for a large part of greenhouse gas emissions into the atmosphere. To mitigate this problem, some countries are focusing on the use of biofuels as a short-term strategy (EPA, 2024). Biomass plays a fundamental role in this production chain, but the limitations imposed by the use of arable land for energy crop plantations undermine the acceptance of this solution by many countries. Thus, the expansion of bioenergetic crops plantations to semiarid regions is a way of increasing the production of biofuels around the world (Raya et al., 2021; Pérez-Pimienta et al., 2017). In Brazil, the total area of the semiarid regions is around 83 million hectares, and the majority of this land is unoccupied, mainly because of its climate (Projeto MapBiomas, 2019). In Europe, aside from the several semiarid regions spread over the South of Portugal, Southeast and central area of Spain, Greece, there is the possibility to benefit from semiarid regions in North Africa and the Middle East (Raya et al., 2021). In this context, it is important to study crops that can thrive in such semiarid areas with high temperatures, high insolation, and low water availability, especially without irrigation, as is the case for agaves (Pérez-Pimienta et al., 2017).

1.1 Characteristics of Agave

Agaves are members of the Agavaceae family, and the Agave genus includes more than 200 species (Eguiarte et al., 2021). They are monocotyledonous, predominantly monocarpic, succulent, and xerophytic plants, and usually reach maturity in approximately 10 years (Raya et al., 2023, Sarwar et. al, 2019). They are native to arid and semiarid regions of North America, particularly Mexico, and they have been used and cultivated by mesoamericans for at least 9000 years (Vargas-Ponce et al 2009).

Agaves are known for their resilience to water stress (Eguiarte et al., 2021; Raya et al., 2023). This resilience is attributed to their photosynthetic pathway known as crassulacean acid metabolism (CAM) (Davis et al., 2010; Yin et al., 2018). CAM plants have minimal water requirements and thrive in semiarid environments because they absorb carbon dioxide during the night, which causes a reduction in transpirational water loss and, therefore, enhances water use efficiency. The efficacy of CAM photosynthesis in water conservation is important in regions with significant

temperature fluctuations between day and night, because opening stomata during cooler nighttime periods minimizes water loss per unit of carbon dioxide assimilated (Davis et al., 2010; Yin et al., 2018).

This efficient use of CAM allows for growth in semiarid regions, but the association of this characteristic with biomass accumulation is something that happens in only a few plants, which makes Agaves good candidates for bioenergy production (Borland et al., 2009; Yin at al., 2018).

1.2 Agave as biofuel feedstock

Several factors should be considered when analyzing biofuel feedstock: potential yield per hectare, adaptability to climatic conditions, agricultural inputs, biomass characteristics, and potential applications (Davis et al., 2010; Pérez - Pimienta et al., 2017). Agave stands out across many of these criteria, being, currently, used to produce spirits (tequila, mezcal, etc) and fibers in different regions of the world. In the case of fibers, only 4% of the harvested leaves are converted into commercial fiber, producing a huge of residual bagasse rich in carbohydrates and organic acids, raw material for biorefineries (Pérez - Pimienta et al., 2017; Raya et al., 2021). In the case of spirits, only agave pineapple is used to produce fermentable sugars, generating a large amount of bagasse in this process, including the leaves that are left in the field.

Numerous studies have showcased the bioenergetic potential of Agaves, highlighting the impressive productivity of *A. tequilana*, exclusively used for tequila production in Mexico, despite their minimal water and agricultural management requirements. Depending on the region, *A. tequilana* can yield between 8.5 to 22 Mg ha-1 year-1 of dry biomass, with theoretical analyses suggesting even greater potential, up to approximately 38 Mg ha-1 year-1. In comparison, traditional semiarid crops such as cotton typically yield only 1.5 Mg ha-1 year-1 while demanding higher water resources (Sarwar et. al, 2019). *A. tequilana* is a domesticated specie of *A. angustifolia* with a focus on high sugar content (called inulin, a fructose polymer) and high productivity, measured by pineapple weight and sugar concentration. Although this strategy has been used for beverages, this domesticated plant has great potential for biofuel production.

In Brazil, agaves are primarily grown for their sisal fiber used in the production of ropes, strings, etc. (Raya et al., 2021; Raya et al., 2023), which is extracted from the leaves of certain species, such as *A. sisalana* and two hybrids (called H11648 and

H400L). In general, fiber-producing agave species have a low sugar content in the pineapple (compared to *A. tequilana*) and the fiber productivity is determined by the number of leaves. In this model, the biofuel production could be generated by these fiber extraction residues.

In 2020, Brazil exported 60 thousand tons of fiber-derived products, which means around 1.500 thousand tons of unused bagasse, generating $78 million in revenue for the semiarid region of Bahia, the primary producer. As the world's largest producer and exporter of sisal fiber, Brazil commands 70% of global exports and contributes 58% to global production (Davis and Long, 2015; FAO, 2020). Unfortunately, Brazil's sisal production operates through a decentralized, semi-manual extraction process and farmers primarily rely on undefined *A. sisalana* plants with minimal agricultural management, exacerbating challenges posed by diseases like sisal bole rot, which threaten production (Pérez-Pimienta et al., 2017; Raya et al., 2022).

Thus, the organization and structuring of this sector of the economy has the potential to improve the quality of life of families that produce sisal fiber, in addition to allowing the reuse of leaf bagasse to produce biofuels, such as ethanol, biogas, among others. Furthermore, introducing *Agave tequilana* in these regions has the potential to increase biofuel production capacity due to the high sugar accumulation in their pineapple. In the context of this master's project, the focus is on understanding the molecular bases of these plants through comparative and population genomic analyses, aiming to apply the findings in breeding programs for these cultivars.

1.3 Important Phenotypes of Agave species

When aiming to improve the efficiency of Agave for biofuel production, three main phenotypes should be considered: biomass productivity, drought resistance and disease resistance. In the case of fiber-producing agave, biomass productivity is represented by the number of leaves present in each individual that can be used in the production of fibers and biofuel (due to waste). During the lifetime, *A. sisalana* produces around 250 leaves, the hybrid H11648 around 350 leaves, while the hybrid H400L reaches 400 leaves (Figure 1), which represents around 57, 90 and 100 tons/ha/year of leaves, respectively.

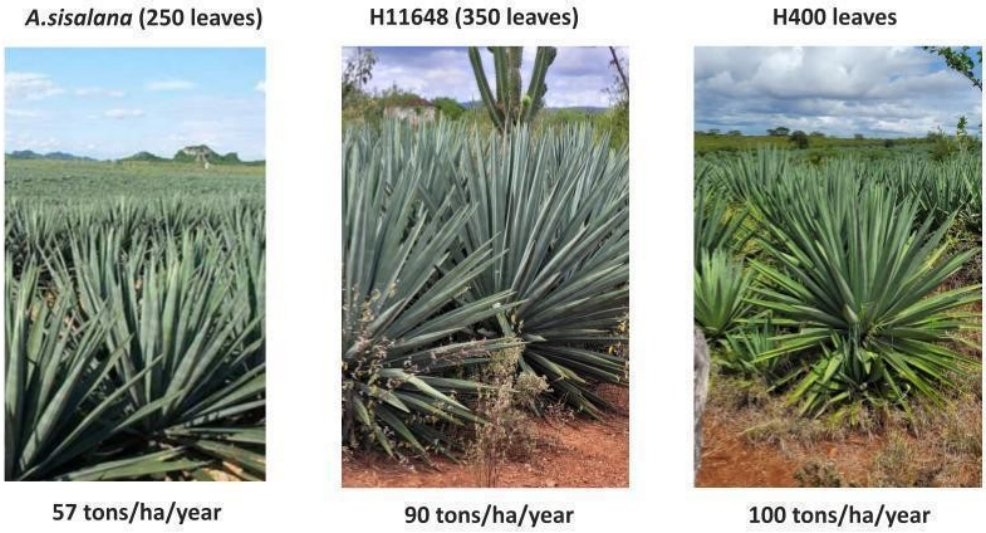

Figure 1 - Biomass productivity (tons/ha/year) of *A. sisalana*, H11648 and H400L

Agave sisalana is named after the port, Sisal in Yucatán, Mexico where it was originally exported for fiber production in Africa, India and Brazil (Davis et al 2019). This cultivar was predominant in East Africa, but it was then replaced by H11648 in Tanzania and other regions, because H11648 contains more fiber per leaf and produces a higher amount of leaves per year (Zhang et al 2013). H11648 ((*A. amaniensis* x *A. angustifolia*) x *A. amaniensis*) was developed in 1948 in a breeding program in Tanzania/Africa aimed to improve fiber-producing cultivars. The other hybrid, H400L, also has a higher fiber content and produces more leaves per year than both H11648 and A. sisalana. However, the origin of H400L is unknown, but is believed to be closely related to the H11648 (Souza et al., 2018).

Since 2012, Brazil has witnessed a significant decline in sisal fiber production, primarily attributed to the prevalence of bole rot disease, which stands as the principal phytosanitary challenge in the country's sisal fields (Raya et al, 2023; Abreu, 2010; Soares et al., 2020). This disease is instigated by *Aspergillus welwitschiae*, a saprotrophic fungus that transitions to a necrotrophic existence upon invading injured sisal boles, causing destruction to the parenchymal tissue (Duarte et al., 2018). A difference in symptomatology between common sisal and hybrid cultivars (Table x), especially H400L, was observed, and the latter did not present any symptoms (Raya et al., 2023).

Regarding drought resistance, field observation implies that hybrid H11648 thrives better in those stress environments than *A. sisalana*, as can be seen in Figure 2 during a long period of drought stress in semiarid of Bahia, in which H11648 appears to be healthier.

Figure 2 - Two varieties of agave demonstrating differing drought resilience: H11648 (left) thrives despite prolonged dryness, while *A. sisalana* (right) shows signs of stress under the same conditions.

Overall, we can compare production efficiency of the different genotypes of Agave used in Fiber production based on the following characteristics: biomass productivity (number of leaves), disease resistance (Bole rot disease), and drought tolerance (Table 1).

Table 1 - Phenotypes of Fiber Producing Agaves: H11648, H400L and *A. sisalana*

	Biomass Productivity	**Disease Resistance (Bole rot)**	**Drought Tolerance**
A.sisalana	250 leaves 57 tons/ha/year	High incidence High symptoms	Moderate
H11648	350 leaves 90 tons/ha/year	High incidence Very low symptoms	High
H400L	400 leaves 100 tons/ha/year	High incidence No symptoms	NA

Regarding ethanol production, *A. tequilana* (domesticated species of *A. angustifolia)* is the most significant variety to focus on. Evidence shows that this variety has been selected for traits favorable to tequila production and cultivation, such as high sugar content, a short life cycle around 5 years, minimal obstruction from teeth and spines, and a higher number of rhizomes. For tequila production, the agave is harvested at maturity, then leaves are removed, and the stem (pineapple) and leaf base are processed (Figure 3). The carbohydrates in the agave stem are broken down into sugars through heat treatment, and the resulting juice is subsequently fermented. The residues (leaves and bagasse) from this process can also be utilized in the production of second-generation biofuels, which gives this variety an even greater economic importance (VALENZUELA et al, 2011).

Figure 3: The harvesting of *Agave tequilana* for tequila production in Mexico

1.4 Genetics

Agaves are characterized by their complex and large genomes with polyploidy and hybridization being widespread in the genus (Raya et al., 2023; Robert et al., 2008). Genome sizes range between 2.9 and 12.2 Gbp (1C) (Raya et al., 2023). For commercially utilized species, the haploid genome sizes are: 3.75 Gbp for *A. tequilana* (2n), 3.68 Gbp for *A. sisalana* (5n), 4.25 Gbp for the hybrid 11648 (2n) and unknown for H400L (2n) (Robert et al., 2008; YANG et al, 2024).

The number of chromosomes can vary between species, because of aneuploidies or duplications (Simpson et al., 2011). They can form a euploid series of 2x, 3x, 4x, 5x and 6x with basic chromosome number n = x = 30 (Palomino et al., 2003). It is also known that during metaphase, their mitotic chromatids tend to display a distinctive pattern where larger chromosomes are located at the periphery of the metaphase plate and smaller ones are found in the central region, regardless of the level of ploidy (Granick, 1944). Regarding their ploidy, research shows that *A. sisalana* is a pentaploid with 5n=150 chromosomes, and *A. tequilana* and H11648 are diploid with 2n=60 chromosomes.

Currently, there are complete genomes for *A. tequilana* and H11648, and raw DNA reads for *A. angustifolia*, publicly available. The *A. tequilana* genome was sequenced, assembled and annotated by The Joint Genome Institute (JGI)/USA, which also made the *A. angustifolia* raw reads available. The final genome assembled had size 3.75Gb organized in chromosomes (2n = 2x = 60 chromosomes) and scaffolds (N50 of 282 Mb). The annotation process showed 42.199 protein coding-genes and 73.837 protein coding-transcripts (Yang, 2024).

The second complete genome of the agave genus is the hybrid H11648, with 30 pseudo-chromosomes and estimated size 4.25Gb, was assembled and annotated by YANG et al, 2024. The genome was composed of 80.29% repeats, heterozygosity rate of 0.42%, and karyotype determination typical of agave species, with 25 small and 5 large pairs of homologous chromosomes. (YANG et al, 2024).

1.5 Reproduction and Propagation of Cultivated Agave Species

Agave species propagate through two main strategies: sexual reproduction, via seeds, and asexual reproduction, through rhizome offsets and bulbils. Sexual reproduction occurs when seeds form from pollinated flowers, which grow on the floral stalk (scape) of the agave plant. Pollination in agaves is facilitated by various animals, including birds, rodents, and insects, attracted by the flowers' nectar and pigmentation (Gentry, 1982; Queiroga et al., 2020). Among these pollinators, bats (Chiroptera) play a particularly vital role (Queiroga et al., 2020). Their migratory behavior and ability to traverse distances of up to 35-50 km allow them to visit multiple plants, effectively transferring pollen and promoting genetic diversity (Fleming et al, 2009).

Agave sisalana, however, is considered sterile under natural conditions (Queiroga et al., 2020). This sterility is attributed to mechanical factors, specifically the cessation of abscission layer activity at the junction of the flower and pedicel. Despite this, it was observed that under certain environmental conditions, *A. sisalana* could produce fruits and viable seeds (Nutman et al, 1931). Nevertheless, sexual propagation is rarely used due to the species' low seed germination rate and the lengthy period—approximately three years—required for seedlings to reach a plantable size (Queiroga et al., 2020).

Vegetative reproduction, bulbils and rhizome offsets are the primary form of propagation used in cultivated agave species, which indicates that these populations may be mostly clones (Nobel, 1994). Between these methods, rhizome offset is the most widely used, as plants derived from rhizomes grow more quickly and robustly compared to those propagated from bulbils (Queiroga et al., 2020).

1.6 Bioinformatics analysis for plant breeding

Next-Generation Sequencing (NGS) enabled a higher amount of data to be sequenced at a lower cost allowing us to perform more robust bioinformatics analysis and generate a better insight into genetic diversity, gene expression, and molecular mechanisms of important traits such as drought resistance, disease tolerance, and enhanced yield (Novogene, 2024). Through comparative genomics, genetic markers, and population genomics we can deepen our understanding of plant biology and apply this knowledge to crop improvement. Comparative genomics offers insights into traits and genome evolution through the analysis of genomes differences and similarities (Hardison et al, 2003). Genetic markers, such as single nucleotide polymorphisms (SNPs) and microsatellites (SSRs), can be used to assess genetic variability and associate it to traits of interest in crop productivity (Andrews et al., 2016; Amiteye, 2021). Population genomics examines genetic variation within and between populations, giving us a better understanding on domestication processes and the genetic basis of agronomically important traits, and enabling the development of more resilient and productive crops (Novogene, 2024). There are still challenges brought by the complexity of plant genomes, such as polyploidy and large structural variations, however advances in long-read sequencing technologies are addressing these obstacles, which further advances plant genomics research and its applications in sustainable agriculture.

2. MOTIVATION

Biofuels are a promising alternative to mitigate the emission of greenhouse gas emissions in the atmosphere. However, for this alternative to be viable, there is a need to identify biomass that does not impose limitations in the use of arable land. Agave is the perfect example of a bioenergetic crop that can thrive in semiarid regions characterized by high temperatures, intense sunlight, and low water availability and therefore can be used as a sustainable option for biofuel production.

Agave species, as most plants, have complex genomes with high ploidy and repetitive regions, making it challenging for understanding the molecular mechanisms behind interesting traits for crop productivity. However, with the advance in sequencing technologies and bioinformatics tools we can use comparative genomics, population genomics and high-impact genetic mutations analysis, to discover valuable genetic targets for crop improvement. These insights are critical for advancing biofuel production and supporting the global transition to more sustainable energy systems.

In this master's thesis, we use population genomics, SNPs/Indels analysis and comparative genomics to understand the biological processes underlying mechanisms related to phenotypes of interest in agave, such as biomass productivity, disease resistance, drought tolerance and sugar/fiber content, which can be targets for further genetic engineering and crop productivity improvement.

In the first chapter, we focus on population genomics and SNPs/Indels applied to ddRADseq data from 3 populations: H11648, H400L and *A. sisalana*. Population genomics is used to analyze the structure and diversity between the populations of H11648 and H400L, and the subpopulations of *A. sisalana* in different geographical locations. SNPs and Indels analysis is employed to understand what similarities and differences between the populations of H11648, H400L and *A. sisalana* can be associated with the different phenotypes important in fiber production in Brazil: biomass productivity, disease resistance and drought tolerance.

In the second chapter, we use whole genome data to perform comparative genomic analysis between H11648, *A. sisalana* and *A. tequilana*, which are genotypes of great importance in fiber and biofuel production. The focus is to expand the analysis made on chapter 1 about the differences and similarities in the fiber production genotypes H11648 and *A. sisalana*. Additionally, we introduce *A. tequilana*, enabling a comparison between genotypes with high fiber content (H11648

and *A. sisalana*) and those with high sugar content (*A. tequilana*), a key trait for biofuel production.

3. CHAPTER I: ANALYSIS OF POPULATION GENOMICS AND NON-SYNONYMOUS SNPS RELATED TO HIGHLY PRODUCTIVE AGAVE GENOTYPES

3.1 Introduction

Agave species show different phenotypes that can be of great economic value in industries like sisal fibers and biofuel production. Through population genomics and SNP/Indels analysis of several individuals representing fiber-producing genotypes in Brazil (*A. sisalana*, H11648 and H400L), we can gain valuable insights into the genetic diversity of these varieties and identify potential targets for genetic improvement. The process used to produce fibers generates a lot of biomass residues that can be converted into bioethanol or biogas, therefore the improvement of fiber-producing cultivars has the potential to increase the income of local populations in a short period of time, in addition to the future bioenergy sector. By contrasting the SNPs and Indels present in each of these genotypes, we aim to identify the genomic markers responsible for the most interesting phenotypes for production: number of leaves, disease resistance and drought tolerance. In addition, using population genomics tools, we can uncover the genetic changes that have occurred during domestication and breeding. This information is crucial for developing new breeding strategies that leverage genetic diversity. Ultimately, this data can be used to pinpoint targets for genetic engineering, leading to improved crop productivity and resilience in fiber production.

3.1.1 Molecular markers

With next-generation sequencing (NGS) technologies at lower costs, whole-genome sequencing (WGS) became a powerful tool for population genomics. WGS provides a complete genetic profile of an organism, which enables the identification of genes linked to desired traits like disease resistance, productivity, and yield, that can be used for genetic engineering and crop improvement (Novogene, 2024).

NGS has also allowed the improvement of genetic markers analysis (Alves-Pereira et al, 2020). Genetic markers, like microsatellites, RAPD, ISSR, IRAP, and AFLP, are important tools for identifying genetic variability and distinguishing between genotypes

(Amiteye, 2021). Currently, the most commonly employed markers are derived from SNPs (single nucleotide polymorphisms) or microsatellite sequences (also known as simple sequence repeats, SSRs) (Andrews et al., 2016; Amiteye, 2021).

Since Agaves have complex and large genomes, SNPs are generally obtained from methods that use restriction enzymes and Next-Generation Sequencing (NGS), allowing the comparison of several sequenced regions of the genome in several individuals at a more affordable cost and less complexity (Andrews et al., 2016; Peterson et al., 2012). Among these methods, the best known are GBS (Genotyping By Sequencing) and derivatives of RAD-seq (Restriction Associated DNA sequencing), such as ddRAD-seq (Double Digestion RAD-seq). Similar to other reduced-representation sequencing techniques, RADseq focuses on a subset of the genome, offering some advantages over whole-genome sequencing like higher coverage depth per locus, leading to increased confidence in genotype calls, and the ability to sequence more samples within a given budget. Consequently, RAD-seq has emerged as the predominant genomic approach for high-throughput SNP discovery and genotyping in ecological and evolutionary studies involving non-model organisms (Andrews et al., 2016; Peterson et al., 2012).

Although WGS is a more complete genome-wide analysis, ddRAD sequencing is also an efficient strategy for identifying SNPs. This strategy focuses on genomic regions with high polymorphism, providing sufficient SNP data for population studies and can be performed at a lower cost using less computational resources. Furthermore, as shown in previous studies, ddRAD's targeted approach addresses the limitations of WGS in areas with low variant frequency, which makes it a greater choice for SNP identification (Boatwright et al., 2022).

3.1.2 SNPs and Indels

As mentioned, genetic markers are important tools in understanding genetic variability. In this chapter, we will focus on SNPs (single-nucleotide polymorphism) and Indels (insertion/deletion polymorphism). A SNP is a variation in the DNA sequence where a single nucleotide (A, T, C, or G) differs among individuals within a species or between paired chromosomes in an individual. These variations can happen in coding and noncoding regions in the genome. When this variation causes a change in the amino acid in a coding region, we call it missense SNPs. This type of SNPs can impact the final translated protein activity by changing its folding patterns,

catalytic functions, allosteric regulation, localization, post-translational modifications, aggregation, or half-life. In noncoding regions, SNPs can affect transcription factor binding, gene splicing, or mRNA degradation (MARWAL et al, 2020).

An insertion/deletion polymorphism (Indel) is a variation where a nucleotide sequence is inserted or deleted. Although Indels are less common than SNPs, they can significantly influence gene function and protein structure. In coding regions, Indels that are not in multiples of three nucleotides cause frameshift mutations, which alter the reading frame of the messenger RNA, which can lead to truncated proteins, destabilization, misfolding, or complete loss of function In-frame Indels (multiples of three nucleotides) lead to the insertion or deletion of amino acids within the protein sequence without disrupting the reading frame. While these changes are less disruptive they can still affect protein stability and function, particularly if they modify key structural regions, active sites, or ligand binding sites (RODRIGUEZ-MURILLO et al., 2020). Indels can also affect regulatory regions, altering protein interactions, localization, or enzymatic activity, thereby influencing cellular processes and phenotype. Indels in noncoding regions can disrupt transcription factor binding sites, alter mRNA processing, or lead to alternative splicing, which can produce truncated or functionally distinct protein isoforms.

3.1.3 Population Genomics

WGS and SNPs are valuable tools for understanding the history and evolution of crops. By comparing the genomes of different varieties and their wild relatives, we can learn about the genetic changes that have occurred during domestication and breeding. This information can be used to develop new breeding strategies that take advantage of the genetic diversity that exists within wild crop relatives (Novogene, n.d.).

As mentioned above, the Hybrids H11648 and H400L have great potential in fiber production due to their phenotypes (greater number of leaves, better resistance to disease and drought) compared to *A. sisalana.* Therefore, it is important to gain a better understanding of their origins and evolution. We know that H11648 ((*A. amaniensis* x *A. angustifolia*) x *A. amaniensis*) was developed in 1948 in a breeding program in Tanzania/Africa aimed to improve fiber-producing cultivars. However, the origin of H400L is unknown. Since this variety is believed to be closely related to the H11648 (Souza et al., 2018), we can use phylogenomics tools to infer its origin. And,

in addition to a phylogenetic tree construction, the analysis of population structure and genetic variation metrics, such as expected and observed heterozygosity (He and Ho), FST (fixation index), PCA (principal component analysis) and admixture can provide a better understanding of the relationship between the different species and their evolutionary history.

In addition to comparing populations across different species, examining populations of the same species in distinct geographic locations provides valuable insights into how environmental, geographic, and anthropogenic factors shape genetic variation and adaptability. On the anthropogenic front, for instance, the process of cultivation significantly influences the structure and diversity of populations within a species. In cultivated agave species, as previously noted, propagation primarily occurs through rhizome offsets—a form of asexual reproduction—raising the question of whether these individuals are genetically identical clones (Nobel, 1994). Furthermore, the primary pollinator of agave, bats, typically don't forage for distances longer than 30-50 km, whereas the plantations under study are over 120 km apart (Fleming, et al 2009). Therefore, by employing population metrics and comparing agave sisalana populations across different plantations, we can uncover patterns of gene flow, local adaptation, and population differentiation. Such insights are critical for guiding conservation strategies, improving breeding programs, and enhancing agricultural productivity in diverse environments.

For genomics population analyses, the use of neutral SNPs is more common because they accumulate at a uniform rate and provide unbiased estimates of random processes (Luikart et al., 2003; Storz & Nachman, 2003). In analyzing population structure, we can identify the pattern of genetic relations between populations and subpopulations. A common metric applied in this analysis is FST, which is a measure that reflects the proportion of genetic variation found within a subpopulation compared to the total genetic variation across all populations (Weir, 1996). An FST value over 0.15 is generally considered significant for distinguishing between populations (Frankham et al., 2002). Another commonly used metric is principal component analysis (PCA), which is an unsupervised learning technique that identifies population structure based on genetic variation (He et al., 2024).

Genetic variation within populations can also be analyzed through observed and expected heterozygosity. Expected heterozygosity is calculated from allele frequencies, while observed heterozygosity is derived from the real observed

heterozygosity of the population (Ritland, 1996). When expected heterozygosity exceeds observed heterozygosity, we can interpret it as evidence of local inbreeding (Hoffmann et al., 2021).

3.2 Objectives

In the context of this chapter, we aim to use population genomics and non-synonymous SNP/Indel analysis to understand the biological processes underlying phenotypes of interest for sisal fiber production, such as biomass productivity and disease and drought resistances, in *A. sisalana*, H11648, and H400L. These analyses could enhance our understanding of the genes responsible for each phenotype, potentially identifying targets for genetic engineering and crop productivity improvement.

3.2.1 Specific Objectives:

Compare populations of genotypes relevant to biomass production in Brazil, specifically *A. sisalana*, H11648, and H400L and identify the genomic markers responsible for phenotypes with greater productivity: number of leaves, disease resistance and drought tolerance.

- Identify neutral SNPs and non-synonymous mutations (SNPs and Indels) in the different populations.
- Identify the structure and diversity of the H400L, H11648 and *A. sisalana* populations through genomic metrics: phylogenetic tree, Fst, PCA, He, and Ho.
- Analyze the list of genes with mutations focusing on the correlations of productivity and resistance phenotypes through enrichment analyses (GO and KEGG)

3.3 Material and Methods

3.3.1 ddRAD-seq of fiber-producing cultivars

The individuals were collected (small piece from the leaves) on a field trip to the Bahia semiarid in August 2019 by Marina Püpke Marone during her PhD thesis. The focus was the specie *A. sisalana*, and the hybrids H11648 and H400L distributed in the four cities (more than 100 kilometers among them) that representing the most producing regions: Itiúba, Conceição do Coité, Campo Formoso, and Várzea Nova. Many interesting phenotypes were sampled, among them, the cultivars presenting

long leaves (up to 1.7m length), apparent resistance to the bole rot disease (healthy plants beside sick ones), and absence/presence of spikes in leaf borders. In the first area, Itiúba (IT), 46 individuals were sampled, in Conceição do Coité (CC) 59 individuals were sampled, in the area Campo Formoso (CF) 53 individuals and in the Várzea Nova (VN) 56 individuals, resulting in a total of 214 samples of individuals of *A. sisalana* or hybrids (Figure 4). From the 214 samples, 95 individuals were chosen in order to represent equally the different sampling locations and phenotypes of interest. The 95 individuals had DNA extracted, 77 of *A. sisalana* and 18 of hybrids (10 H11648 and 8 H400L), and submitted to the company Floragenex/USA to perform the ddRAD-seq (double digest RAD-seq) protocol and library preparation. The restriction enzymes used were PstI and MseI, generating fragments with an estimate of 5K-15K bp. After preparing the libraries, sequencing was done with Illumina HiSeq technology. Sequencing generated 715,237,767 single reads of 138 bp after removing the adapters. A demultiplexing step to remove the barcodes was performed with sabre v.1.0 software (https://github.com/najoshi/sabre) using default parameters, generating reads with 127 bp. The reads without an association with the individual, 32%, were subjected to another round of demultiplexing, now allowing a mismatch ("-m 1") in the barcode sequence, which generated an increase in the number of reads for some individuals, but still some samples presented low coverage. In total, 188,607,081 (26%) reads remained unassigned to any individual, probably due to some error in the DNA extraction or sequencing step.

Figure 4 - Summary of location of the samples collected. Map of the Bahia semiarid region where the samples were collected and the number of samples in each location. Each spot is a different farm.

3.3.2 Publicly available genome data

In this chapter, we used the *A. tequilana* genome, retrieved from JGI, as the reference genome for our variant calls and population genomics analysis. Additionally, Illumina raw DNA reads of *A. angustifolia* (100x coverage) were obtained from JGI repository (Yang, 2024) to complement the population analysis, as it is the progenitor of H11648.

3.3.3 Alignment and Variant Calling

The analyses were performed for each genotype *A. sisalana*, H11648, H400L and *A. angustifolia* following the same methods. First, the *A. tequilana* genome was downloaded from JGI, and the FASTq files for each individual were aligned against it using BWA-MEM v.0.7.17 (Li et al, 2010). The SAM files were converted to BAM with SAMTools view v.1.6 (Li et al, 2009). Read mapping quality was calculated using SAMTools view and the samples with a total alignment under 95% were removed. The BAM files were ordered by alignment position using Picard v2.27.5 (Broad Institute, 2019). The Variant calling was performed using GATK v4.3 (Van der Auwera et al, 2020), using the modules ‘HaplotypeCaller’, ‘CombineGVCFs’,

‘GenotypeGVCFs’. SNPs and Indels were separated using the module ‘SelectVariants’. SNPs and Indels were filtered with the following parameters: "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" and "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0", respectively, using GATK module ‘VariantFiltration’. The ploidy parameter was set to diploid for hybrids and A. angustifolia, and to pentaploidy for *A. sisalana*.

We used BCFTools v.1.9 and an in-house bash script, for further filtration with focus on population metrics. For the H400L (5 individuals) and H11648 (7 individuals) populations, we kept homozygous genotype calls if they have at least 5 supporting reads; heterozygous calls are accepted if they are supported by at least 10 reads. In the context of population, SNPs with a minor allele frequency (minmaf) below 20% or less than 30% present calls (minpresent) were discarded (DP > 100, --dphom 5 --dphet 10 --minmaf 0.2 --minpresent 0.3). For the *A. sisalana* population (29 individuals), we kept homozygous genotype calls if they have at least 12 supporting reads; heterozygous calls are accepted if they are supported by at least 24 reads, and SNPs with a minor allele frequency below 20% or less than 20% present calls were discarded (DP > 100, --dphom 12 --dphet 24 --minmaf 0.2 --minpresent 0.2). These values were chosen considering ploidy (hybrids are diploid and *A. sisalana* is pentaploid) and the difference between the number of individuals in the populations and were select to make sure that the mutations appear consistently in at least 20% of the individuals of the two hybrids and 20% of the individuals of *A. sisalana*.

3.3.4 Population Genomics

The population genomics analysis was divided into two main areas:

(1) Comparison of the H400L and H11648 populations to understand their diversity and structure and infer the unknown origin of H400L. To make this analysis more robust, we included one of the parentals of H11648, *A. angustifolia*.

(2) comparison of the populations of *agave sisalana* from different production locations to understand the genetic diversity and structure present in the cultivated crops. For this analysis, we used the samples collected from production areas that had 6 or more individuals sampled (total of 22 individuals): 9 individuals from Campo Formoso (CF: 10°30'13.5"S, 40°19'12.6"W), 6 individuals from Lajes do Batata Jacobina (LB: 11°3'9.1"S, 40°46'48.5"W), and 7 individuals from Valente (VA: 18°49'27.1"S, 45°13'33.5"W).

3.3.4.1 Neutral SNPs

The phylogenomic and population structure analysis were performed using only neutral SNPs identified for each genotype: populations of H400L, H11648 and *A. sisalana*, in addition to the individual *A. angustifolia*. For the identification of neutral SNPs, we utilized BayScan v2.1 with 20 pilot runs, 5,000 of length of pilot runs, 20 thinning intervals, 5,000 sample size, burn-in of 50,000 iterations, and 100 prior odds for the neutral model (Foll et al, 2008). BayeScan is a tool used to identify outlier loci based on Wright's fixation index (FST). It decomposes locus-population FST into two components: beta, which is population-specific, and alpha, which is locus-specific, using logistic regression. Positive alpha values point to diversifying selection, whereas negative alpha values indicate balancing or purifying selection (Foll et al., 2008). The results were processed in R and neutral SNPs were identified by q-value>0.05. The VCF files for H11648, H400L, *A. sisalana* and *A. angustifolia* were filtered with BCFTools v.1.12 to maintain only the neutral SNPs for further analysis. For the population genomics analysis between populations of *agave sisalana* from different locations, we kept only the Neutral SNPs present in the 22 individuals of the 3 locations chosen for the analysis.

3.3.4.2 Population Structure Analysis

We estimated pairwise genetic differentiation (Fst) with the R package hierfstat v5.11 and the expected (He) and observed heterozygosity (Ho) with adegenet v2.1.10 (Jombart, 2011). The PCA analysis was performed with SNPRelate v1.36 in R (Zheng et al, 2012). And the population structure analysis was performed using ADMIXTURE v1.3.0 through several runs with different values for K. We used K =3 for the H400L, H11648 and A. *angustifolia* analysis (1); and K = 4 for A sisalana populations analysis (2) (Alexander et al, 2009).

3.3.4.3 Phylogenomics

All phylogenomic analysis were performed following the steps: (1) VCF files of neutral SNPs for each genotype were merged with BCFTools v.1.12 (Li et al, 2011), (2) converted to PHYLIP format with vcf2phylip v2.8 (Ortiz, EM 2019), and (3) the phylogenetic tree was constructed by IQ-TREE v2.0.3 (Minh et al, 2020) with Bootstrap 1,000 and GTR+ASC mode, recommended for SNP data. The results were

visualized with Figtree v1.4.4 (Rambaut et al, 2010). Two separated phylogenetic analyses were performed:

(1) Using all the individuals from the populations of H400L and H11648, and the exclusive individual of *A. angustifolia*. As an outgroup, we chose one individual of *A. sisalana*, that had better alignment quality and a greater number of reads.

(2) Using the 22 individuals of *A. sisalana* from the 3 different locations. As the outgroup, we chose one individual of H11648, the one with better alignment quality and a greater number of reads.

3.3.5 Selection of exclusive and high-impact SNPs and Indels in H400L, H11648 and *A. sisalana* populations

An in-house R script was used to identify exclusive SNPs and Indels for each genotype H400L, H11648 and *A. sisalana*. Here, we focused on comparing the population of each genotype to verify the genetic markers that are specific to each one of them and gain a better insight of the biological processes responsible for their different phenotypes: number of leaves, disease resistance and drought tolerance. We used cutoff parameters to maintain only SNPs and Indels representative of each population in general and

First, we compared the 7 individuals of H11648 with the 29 of *A. sisalana* and then the 5 H400L individuals with the 29 individuals of *A. sisalana*. To be exclusive of H11648, the SNPs had to be in 100% of the H11648 population and not be present in the *A. sisalana* population. We considered the genotypes ./. and 0/0 for individuals that do not present the variant (SNP or Indel), and 1/1 and 0/1 for individuals that had the variant (Table 2). To be included in the list of H400L exclusives, the SNPs had to be in 100% of the H400L population and not be present in the *A. sisalana* population. And to be included in the *A. sisalana* exclusive list, the SNPs had to be present in at least 7 individuals of the *A. sisalana* population, but not be present in the H11648 or the H400L population.

Table 2 - VCF files annotation labels

VCF Genotype annotation	Meaning
./.	There are no reads supporting the variant in this individual

0/0	The individual is homozygous and has the same alleles as the reference
0/1	The individual is heterozygous, and has 1 copy with the reference allele and 1 copy with an alternative allele
1/1	The individual is heterozygous and has alleles different from the reference

Then, to ensure a higher quality in our data analysis, we applied an additional filtering step. Although an initial filtering had been conducted after the GATK Variant Calling, we used this additional step to confirm that the position where each variant (SNP or Indel) was in the reference genome, was also adequately covered with a reasonable amount of reads in the individuals from the other two populations. This ensured a higher confidence in our data, as we are aware of the limitations with coverage depth in ddRAD sequencing. Thus, we retained only the variants meeting the following cross-population coverage criteria: each position required support from at least 4 reads in a minimum of 2 individuals from the H400L population, 4 reads in at least 2 individuals from the H11648 population, and 10 reads in at least 6 individuals from the *Agave sisalana* population. This filtering was executed using an in-house script in conjunction with the samtools depth coverage tool. After this additional screening, we obtained 3 lists: SNPs exclusive of H400L, SNPs exclusive of H11648 and SNPs exclusive of *A. sisalana*. We used the same pipeline for indels.

SNP and Indel annotation for each list was performed with snpEff v4.5 (Lu et al, 2012). Then, SNPs were filtered selecting only the missense mutations and Indels were filtered selecting only HIGH Impact Indels: frameshift_variant, splice_acceptor_variant, splice_donor_variant, stop_gained, stop_lost and start_lost.

3.3.6 Annotation and Gene Ontology (GO)

The exclusive and high-impact gene sets of H400L, H11648 and *A. sisalana* were submitted to GO (Gene Ontology) enrichment analysis using topGO package v2.48.0 (Alexa et al., 2006). Since the *A. tequilana* genome was used as a reference, all gene sets are related to *A. tequilana* proteins. Therefore, for enrichment analyses, all *A. tequilana* proteins were annotated using PANNZER2 (Törönen et al., 2018), and the

Biological Processes (BP) GO terms were selected from the results to be used as background for enrichment analysis.

3.4. Results

3.4.1 Alignment and Variant Calling

We performed the analysis for *A. sisalana*, H11648 and H400L, separately. First, we aligned the ddRAD-seq read of each genotype to the reference genome of *A. tequilana*. After analyzing the quality of each alignment, we selected the samples with good alignment statistics resulting in 7 individuals for H11648, 5 for H400L and 29 for *A. sisalana*. Samples with under 3 million reads and total alignment under 95% were discarded. As a result, for the H11648, we had an average number of reads of 13,8 million and an average total alignment of 98,3%. For the H400L individuals, the average number of reads was 13,6 million and the average total alignment was 97,5%. Finally for the *A. sisalana* individuals, the average number of reads was 10,2 million and the average total alignment was 98,2%.

For the H400L and H11648 individuals, the SNP calling generated a total number of SNPs of 262.955 and 262.884 for H11648 and H400L, respectively. After filtering, it was reduced to 15,877 and 15,105 SNPs for H11648 and H400L, respectively (Table 2). For Indels the total number found without any filters was 31.451 for H11648 and 31.589 for H400L. After filtering, we found 5,302 Indels for the H11648 and 5,235 Indels for the H400L (Table 3). For the 29 individuals of *A. sisalana*, a total of 385,423 SNPs were found. After filtering, 12,168 SNPs remained (Table 3). The total number of Indels found without any filters was 44,397 and after filtering it was reduced to 9.003 Indels (Table 3).

Table 3 - Total of SNPs and Indels for each genotype, H11648, H400L and *A. sisalana*.

Genotype	SNPs	Indels
H11648 (7 individuals)	15,877	5,302
H400L (5 individuals)	15,105	5,235
A. sisalana (29 individuals)	12,168	9,003

3.4.2 Population Genomic Analysis of H400L and H11648

As mentioned above, data from the literature indicates that H400L is closely related to H11648 (Souza et al., 2018). So, to gain a better understanding of the origin of the H400L genotype, we performed phylogenomics and population structure analysis of the H400L, H11648 population adding data from one of the H11648 parentals, *A. angustifolia*. We performed the He, Ho, Fst, PCA, Admixture and phylogenetic tree analyses considering only the 6.116 putatively neutral SNPs obtained through a Bayescan analysis with q<0,05.

The genetic diversity between H11648 and H400L was analyzed by estimating the expected and observed heterozygosity with adegenet. Our results showed that H11648 had greater expected and observed heterozygosity (He = 0.292 and Ho = 0.2095) than H400L (He = 0.248 and Ho = 0.136), showing that the H11648 population is slightly more diverse than the H400L population.

Then, we proceeded to analyze the population structure and phylogenetic trees. Using the hierfstat R package, we obtained a Fst of 0.25 between H11648 and H400L which shows a significant divergence between these two genotypes, since an Fst value greater than 0.15 can be considered as significant in differentiating populations (Frankham et al., 2002). This result indicates that despite their close genetic relationship, they have distinct evolutionary trajectories. This divergence is further supported by PCA analysis (Figure 5), which showed that H400L formed a very distinct cluster from H11648 with EV1 (47,92%). In addition, the Admixture analysis (k = 3) shows a clear structured difference between all 3 populations, H11648, H400L and *A. angustifolia* (Figure 6).

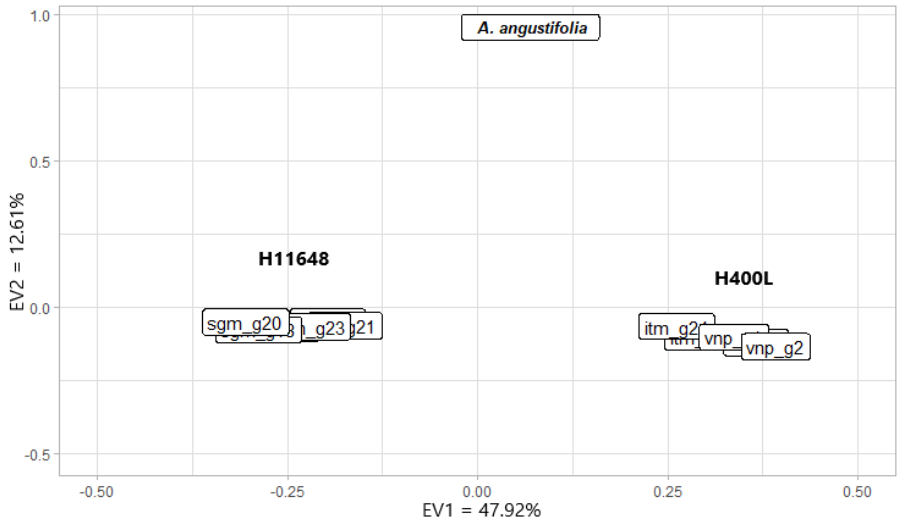

Figure 5 - PCA analysis showing the separation between the H11648 and H400L population

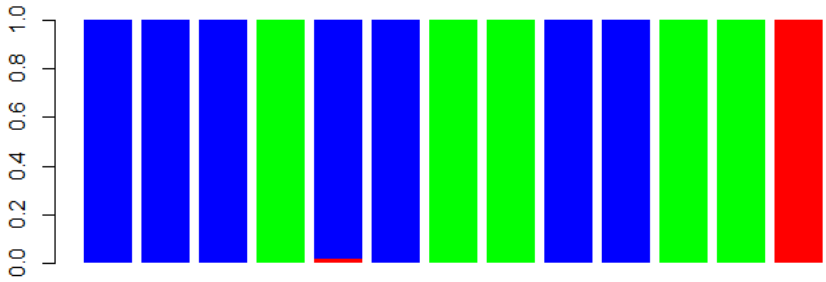

Figure lation are represente red).

We also performed a phylogenetic analysis of these 3 populations using only one individual of *A. sisalana*, for simplification, to be an outgroup (Figure 7).

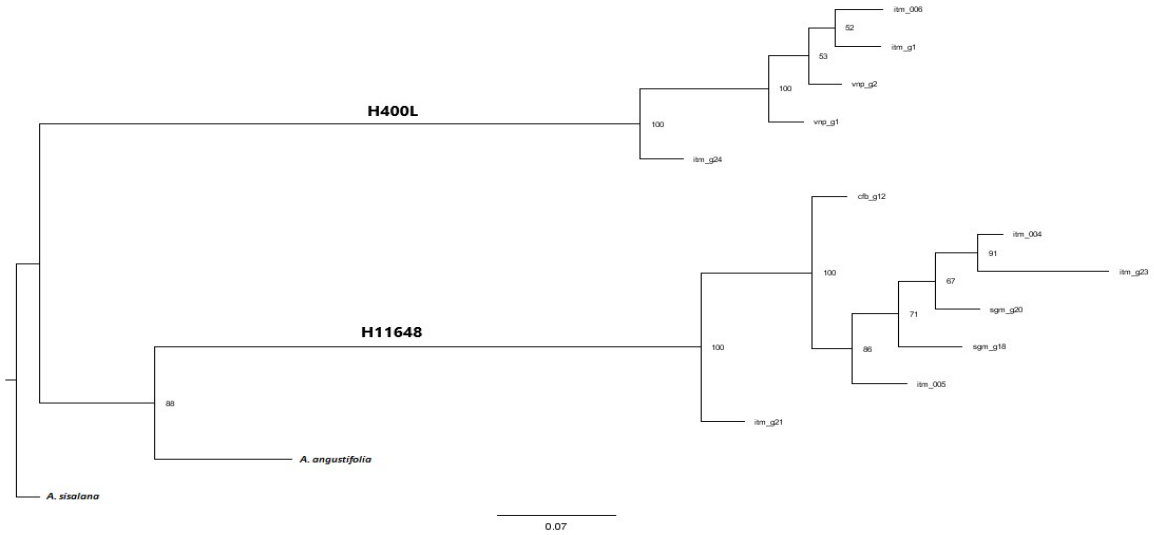

Figure 7 - Phylogenetic tree created with individuals of H11648, H400L, *A.*

angustifolia and *A. sisalana* as an outgroup. The tree was constructed using IQ-TREE v2.0.3 with Bootstrap 1,000 and GTR+ASC mode.

These results demonstrate that H11648 and H400L are closely related to each other, when compared to *A. sisalana*, as demonstrated by Souza et al, 2018, but they have a notable distinction between them, which leads us to believe that they have different parentals. Additionally, H11648 individuals have a higher proximity to *A. angustifolia* than those of the H400L population, which is expected since *A. angustifolia* is one of H11648 parentals.

3.4.3 Population genomic analysis of *A. sisalana* individuals from different production regions

In this parallel, we conducted a population genetics analysis of *Agave sisalana* cultivated populations from distinct geographic regions to assess their genetic diversity and population structure. We focused on cultivated populations separated by over 120 km within the Bahia semiarid primary production regions. A total of 22 individuals were analyzed: 9 from Campo Formoso (CF), 6 from Lajes do Batata (LB), and 7 from Valente (VA) (Figure 8). Genetic diversity and structure analyses were performed using heterozygosity (He, Ho), pairwise Fst, principal component analysis (PCA), admixture analysis, and phylogenetic tree construction. Only neutral SNPs, identified through Bayescan ($q < 0.05$), were included in these analyses, yielding a total of 4,474 neutral SNPs.

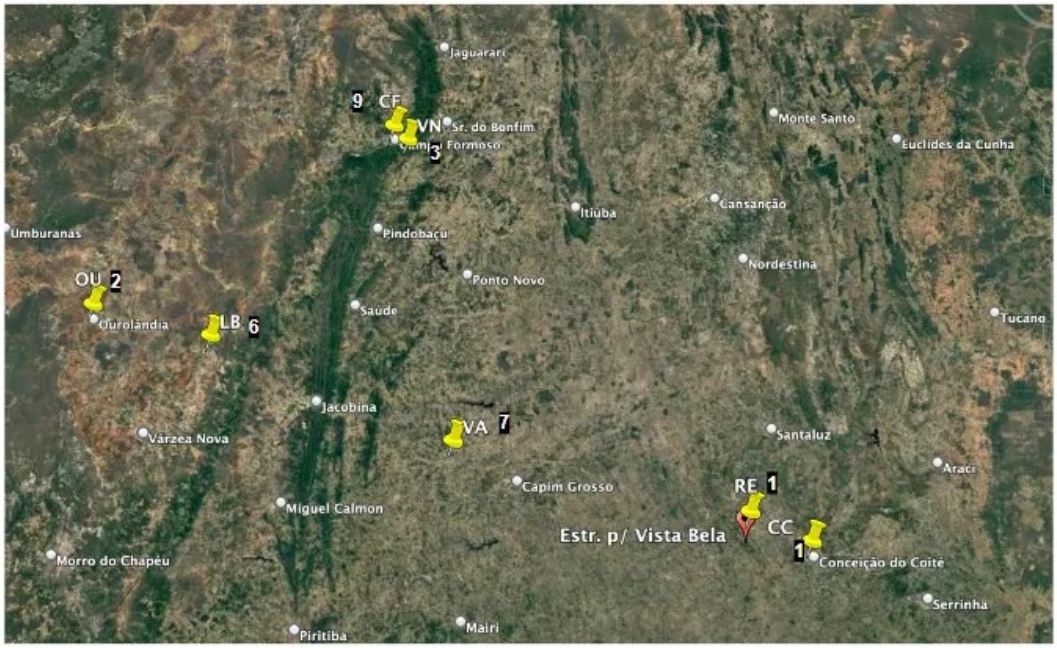

Figure 8 - Geographic location of a. sisalana samples collected in the Bahia semiarid region. In the map, we see the 29 samples used for the SNP/Indel analysis. For our population genetics analysis, we chose 22 samples from three different locations: 9 individuals from Campo Formoso (CF), 6 individuals from Lajes do Batata Jacobina (LB)) and 7 from Valente (VA).

The genetic diversity analysis revealed moderate diversity in the CF population (He = 0.399, Ho = 0.156), whereas the LB population exhibited the lowest diversity (He = 0.270, Ho = 0.104). The VA population showed genetic diversity similar to CF (He = 0.391, Ho = 0.191). These differences suggest that the CF and VA populations maintain slightly higher genetic variability than LB.

Pairwise Fst values supported these findings, indicating significant genetic differentiation among populations. The Fst between CF and LB was 0.2649, highlighting substantial divergence, while CF and VA had a lower but still moderate differentiation (Fst = 0.1747). The greatest divergence was observed between LB and VA (Fst = 0.2698). Generally, Fst values above 0.15 indicate significant differentiation, confirming that all three populations are genetically distinct to varying degrees.

The PCA (Figure 9) and admixture analysis (k = 4) further underscored the genetic distinction among the populations, with PCA's first eigenvector explaining 11.5% of the variance. These results are consistent with the phylogenetic tree (Figure 10), highlighting clear genetic separation between populations from different regions. This pattern aligns with our previous knowledge: in cultivated *A. sisalana*, propagation occurs primarily through rhizomes, an asexual reproduction method, resulting in clonal individuals within production regions. Moreover, the primary pollinators of Agave species, such as bats, typically do not forage beyond 50 km, further limiting gene flow between regions separated by over 120 km (Fleming et al, 2009). Additionally, *A. sisalana* is sterile under natural conditions, restricting sexual reproduction without human intervention (Queiroga et al., 2020). Interestingly, the admixture analysis identified a minor genetic mixture in one individual. This anomaly warrants further investigation; one hypothesis is that farmers may have attempted artificial sexual reproduction between cultivated populations.

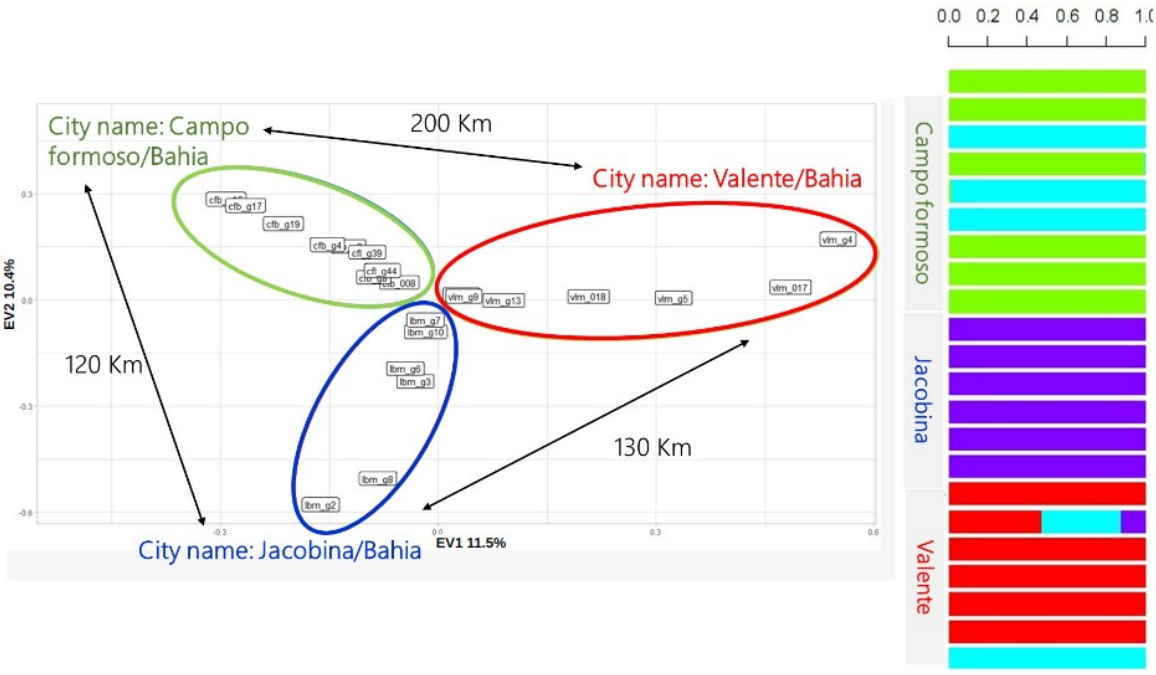

Figure 9 - PCA analysis EV1 = 11.5% and EV2 = 10.4%. ADMIXTURE profile at K = 4. The individuals of each population are represented by the colors: CF (green), JB (purple) and VA (red).

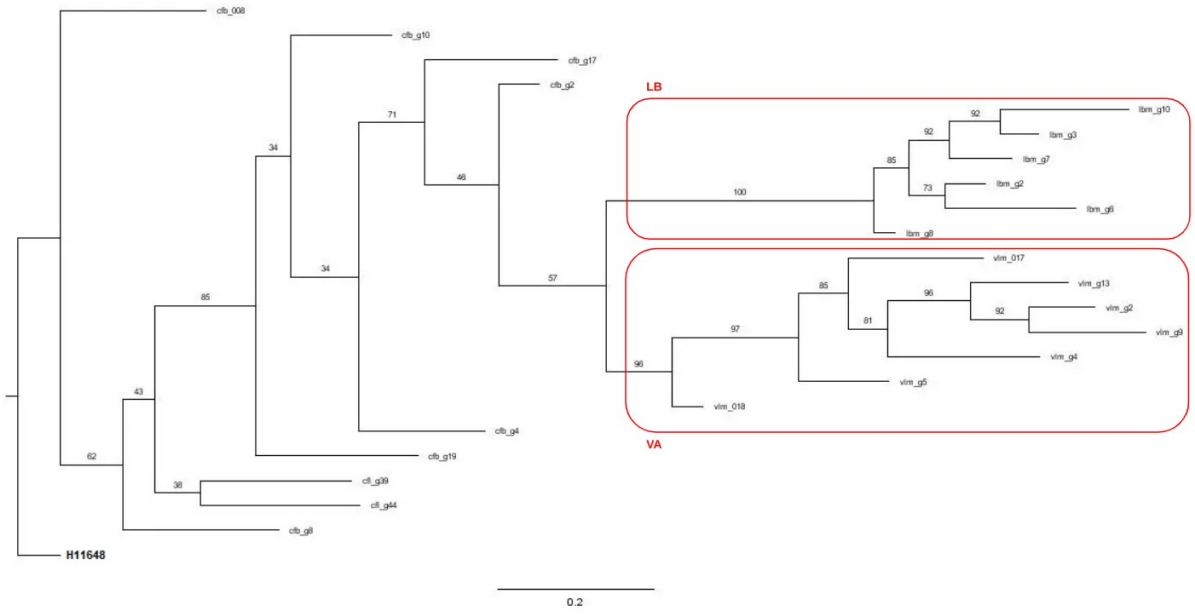

Figure 10 - Phylogenetic tree created with individuals of A. sisalana from 3 different regions (CF,VA and LB) and one individual of H11648 as an outgroup. The tree was constructed using IQ-TREE v2.0.3 with Bootstrap 1,000 and GTR+ASC mode.

Examining genetic variation among geographically isolated populations provides critical insights into the impact of environmental and spatial factors on gene flow and adaptability. These findings enhance our understanding of local adaptation, population differentiation, and genetic diversity, which are essential for conservation strategies, breeding programs, and improving agricultural productivity in diverse environments.

3.4.4 Functional Annotation and Genotype Comparison

As mentioned, the hybrids (H400L and H11648) exhibit phenotypes that can be more interesting to fibers and biofuel production than *A. sisalana*: greater number of leaves, greater disease resistance and drought tolerance. As said, during their lifetime, *A. sisalana* produces around 250 leaves, H11648 around 350 leaves, while the H400L reaches 400 leaves. In addition, the hybrids are more tolerant to drought stress and fungal diseases. By comparing the SNPs and Indels present in each genotype, we can hopefully identify the genomic markers responsible for each phenotype and through genetic engineering improve production.

In addition to comparing the genotypes between each other, we can deepen the analysis by comparing the phenotypes that each individual in the same population exhibit. However, after cleaning and refining the data, we did not have a dataset that was representative for each phenotype and each location. Therefore, we decided not to pursue the analysis within each population and keep only the comparison between them.

A R script was used to identify exclusive SNPs and Indels present in each genotype. To ensure robust analysis, we applied an additional filtering step and kept only the variants in which the position in the genome was also adequately covered by a threshold of reads in the other populations (see methods). Then, a functional annotation of these SNPs and Indels was performed to identify in which genes these SNPs and Indels were present and classify the types of SNPs and Indels, so we could select only the missense SNPs and High Impact Indels. As a result, for the list of genes with missense mutation SNPs, we have a total of 163 H400L, 117 H11648 and 983 A. sisalana genes, respectively (Figure 11).

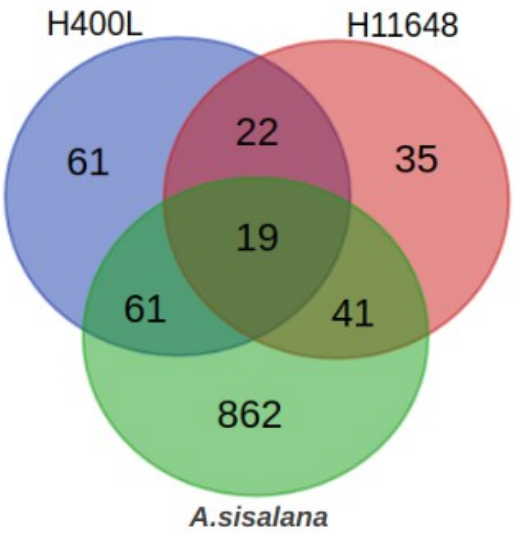

Figure 11 - Venn diagram of the lists of genes with missense mutation SNPs.

For the lists of genes with high impact Indels we have a total of 17 H400L, 13 H11648 and 591 sisalana genes (Figure 12). The results show a higher number of exclusive genes with missense SNPs and Indels for the *A. sisalana* individuals. This can be explained by two main factors: the higher number of *A. sisalana* individuals in the analysis compared to H11648 and H400L; and the fact that *A. sisalana* is a pentaploid (5n), while the hybrids are diploid (2n). The identification of SNPs in polyploid species is more challenging than in diploids for various reasons. In polyploids, there is the need to distinguish between homeologous SNPs (polymorphic positions occurring across subgenomes within and among individuals) from allelic SNPs (polymorphic positions occurring within a single subgenome among individuals). This means that there is more chance of SNP calling softwares to make errors and find false positives SNPs in polyploid species than in diploids (Clevenger et al 2015).

Additionally, we observed a higher number of SNPs compared to indels, with this disparity increasing after applying an additional filter to ensure that variant positions were confidently covered in all populations. This observation aligns with previous studies, which show that Indels are more strongly influenced by purifying selection than SNPs, because they have more potential to disrupt the biological function of the proteins, and that the sensitivity and specificity for detecting indels are limited when coverage depth is below 400x (Perini et al 2025; Sehn et al, 2015).

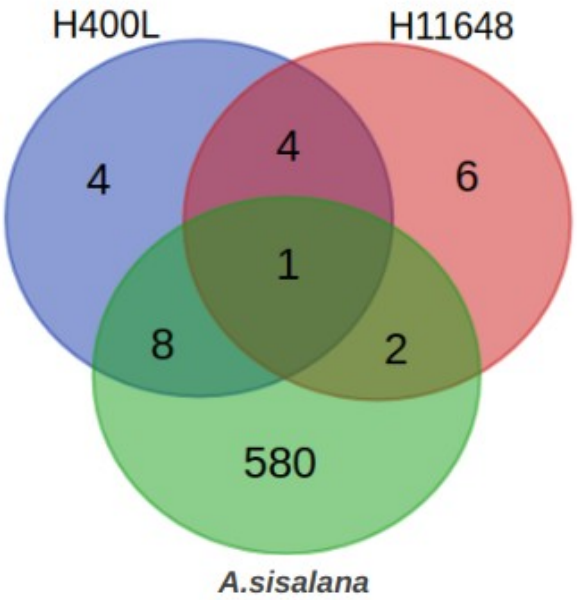

Figure 12 - Venn diagram of the lists of genes with High Impact Indels.

To further analyze the difference between genotypes and try to understand if the different phenotypes of interest were the result of modifications caused by SNPs missense mutations or high impact Indels in a specific gene, we decided to compare the list of genes generated for each list of exclusive SNPs and Indels. So, based on the venn diagrams of SNPs and Indels, we selected 4 final lists of genes (SNPs and Indels, separately). For the SNPs, the final lists were: (A) H400L-exclusive genes containing 122 genes in comparison to H11648, (B) H11648-exclusive genes containing 76 genes in comparison to H400L, (C) A. sisalana-exclusive genes containing 862 genes in comparison to hybrids (H400L and H11648) and (D) Hybrids-exclusive genes containing 118 genes in comparison to *A. sisalana*. For the Indels, we have: (A) H400L-exclusive genes containing 12 genes in comparison to H11648, (B) H11648-exclusive genes containing 8 genes in comparison to H400L, (C) A. sisalana-exclusive genes containing 580 genes in comparison to hybrids (H400L and H11648) and (D) hybrids-exclusive genes containing 14 genes in comparison to *A. sisalana*.

3.4.5 GO Enrichment of exclusive gene lists

If a biological pathway plays a role in the expression of a given trait, there's a probability that the candidate SNPs will be overrepresented within the genes

constituting the pathway. Therefore, in order to achieve a more comprehensive understanding of the candidate SNPs and Indels, we performed a biological process analysis to uncover the genes and mechanisms implicated in the manifestation of our traits of interest (David et al, 2014). We performed a GO enrichment analysis for the final lists of genes: (A) H400L-exclusive genes (in comparison to H11648), (B) H11648-exclusive genes in comparison to H400L, (C) A. sisalana-exclusive genes in comparison hybrids (H400L and H11648) and (D) hybrids-exclusive genes in comparison to *A. sisalana.*

For the H400L (in comparison to H11648) list, we can highlight cellular response to ammonium ion (GO:0071242, p-value 0.006) for the SNPs list as seen in Figure 13 and supplementary table 1. In plants, ammonium up-regulates genes associated with plant defense and immunity, and responses to jasmonic acid. In addition, ammonium can affect genes associated with reactive oxygen species and external stress responses (Liu et al 2017). We should also highlight the glutathione catabolic process (GO:0006751, p-value 0.0274) because glutathione is a plant metabolite responsible for the control of reactive oxygen species (ROS) and the detoxification of methylglyoxal (MG). Plants facing environmental constraints such as salinity, drought, cold, heavy metals, pathogens have an increased level of ROS and MG, which have deleterious effects on multiple classes of biomolecules (Dorion et al, 2021).

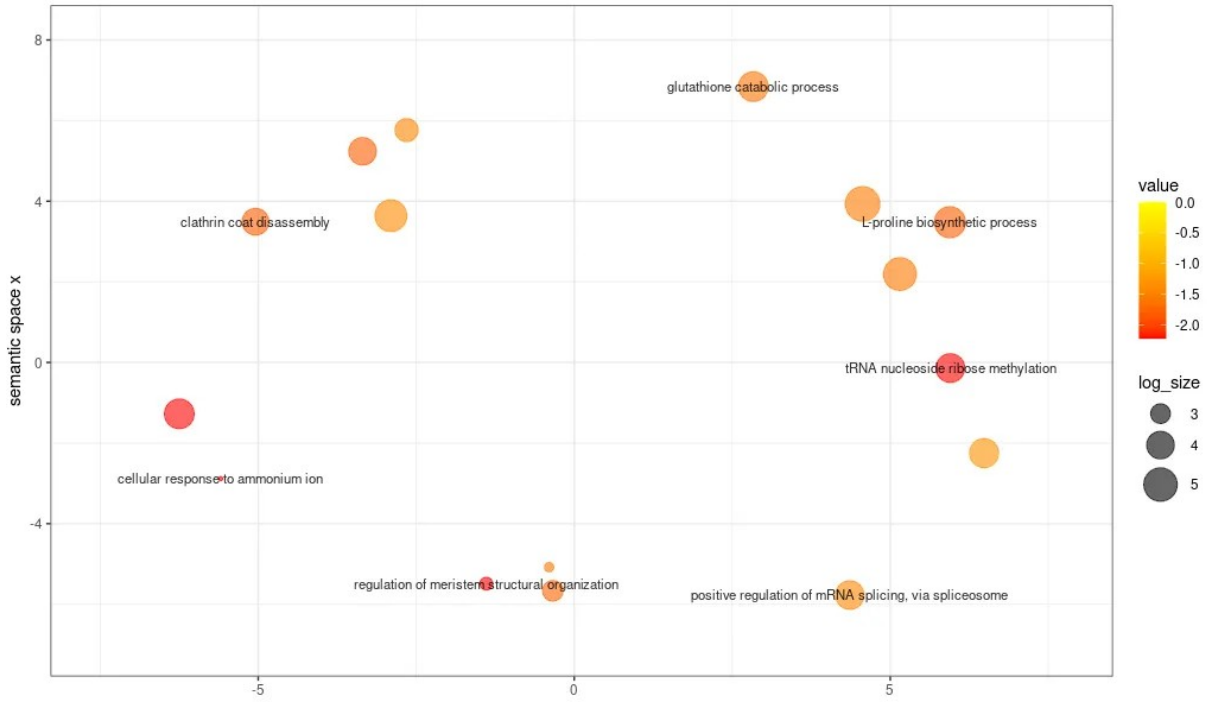

Figure 13 - GOs of the missense SNPs H400L-exclusive genes (in comparison to H11648)

For the H400L list of exclusive high impact indels (Table 4), we can highlight the biological process of phosphate ion transport (GO:0006817, p-value 0.0202). The gene associated with this process is AgveH2v21079969m, which codifies a protein with a Nodulin-like domain. Nodulin-like proteins are increasingly recognized for their roles in transporting nutrients, solutes, amino acids, or hormones, as well as their significance in various aspects of plant development. Notably, these proteins play a crucial role at the plant–microbe interface, where they contribute to nutrient exchange and influence interactions with symbiotic partners or pathogens (Denancé et al, 2024). High-impact indels can cause significant disruptions in proteins, most likely impairing their biological function in the cell. This suggests that the nodulin-like proteins could be compromised in H400L individuals. However, since plants usually have multiple copies of the same gene, we would need to verify if there are other functional copies of this gene in the H400L genome before we could state that the biological function of these proteins (defense mechanisms) are in fact being compromised.

GO ID	Term	Adjusted p-value	Genes
GO:0000492	box C/D snoRNP assembly	0.0029	AgveH2v21004292m
GO:0006817	phosphate ion transport	0.0202	AgveH2v21079969m
GO:0009820	alkaloid metabolic process	0.0409	AgveH2v21091935m

Table 4 - GOs of the high impact indels H400L-exclusive genes (in comparison to H11648)

For the H11648 missense SNPs list (Figure 14; supplementary table 3), we can highlight the following GO: cellular response to iron ion starvation (GO:0010106, p-value 0.0036). Since the detection of iron depletion is a mechanism by which plants recognize a pathogen threat (Herlihy et al 2020), this enriched biological process

suggests that there is a difference in the mechanism that H11648 and H400L react in the presence of a phytopathogen. Additionally, we can highlight positive regulation of flavonoid biosynthesis (GO:0009963, p-value 0.0035). In Agave species, flavonoids production and accumulation are associated with their adaptation to drought stress, UV radiation exposure, high temperature, nutrient deficiency and in their defense system against pathogens (Morreeuw et al 2021).

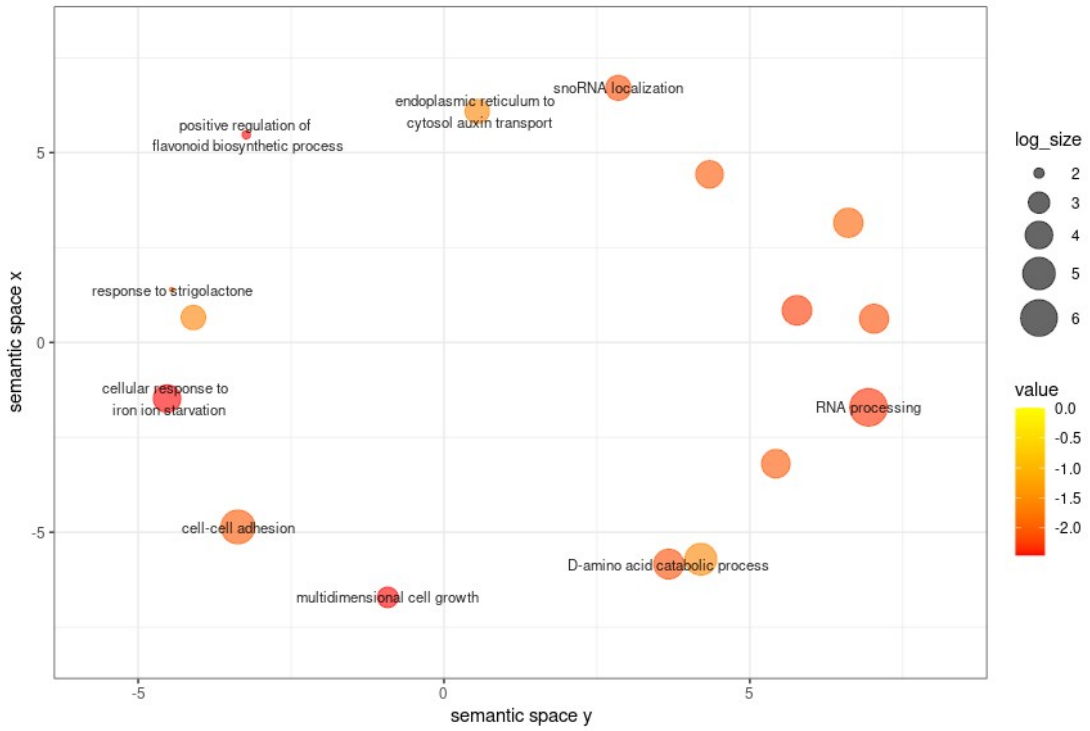

Figure 14 - GOs of the missense SNPs H11648-exclusive genes (in comparison to H400L)

In the H11648 high-impact indels gene list (Table 5), the acyl-CoA metabolic process (GO:0006637, p-value 0.0090) stands out as a significant biological process. The gene AgveH2v21045049m encodes a protein containing an acyl-CoA thioesterase domain. Acyl thioesterases are enzymes that hydrolyze fatty acyl thioesters, releasing free fatty acids. In plant metabolism, these fatty acids are integral to the synthesis of acyl lipids, which fulfill diverse cellular, physiological, and defensive functions. These include the formation of essential membrane, storage, and surface lipids, as well as the production of fatty acid-derived metabolites involved in signaling and defense mechanisms (Kalinger et al, 2020). These high impact indels in the AgveH2v21045049m gene suggests that the protein encoded by it is not functional, suggesting that the biological functions performed by them, signaling and

defense mechanisms, could be impaired or be done differently than H400L. However, like mentioned above, there may be other functioning copies of this gene in H11648, so a further investigation in the genome is necessary.

GO ID	Term	Adjusted p-value	Genes
GO:0018230	peptidyl-L-cysteine S- palmitoylation	0.0052	AgveH2v21005964m
GO:0006261	DNA-templated DNA replication	0.0080	AgveH2v21008473m
GO:0006637	acyl-CoA metabolic process	0.0090	AgveH2v21045049m

Table 5 - GOs of the high impact indels H11648-exclusive genes (in comparison to H400L)

For the *A. sisalana* missense SNPs exclusive gene list (Figure 15; supplementary table 5), we can highlight the indole glucosinolate biosynthetic process (GO:0009759, p-value 0.0241). Indole glucosinolates (IGs) are plant secondary metabolites derived from the amino acid tryptophan and have many roles in processes related to defense against pathogen threats (Pfalz ett al, 2016). This enriched biological process suggests a difference in the response to a phytopathogen between *A. sisalana* and the hybrids (H400L and H11648). Additionally, we can highlight the process glycine betaine biosynthetic process from choline (GO:0019285, p-value 0.04483). Choline is the precursor of glycine betaine, which has osmoprotectant properties and confers tolerance to salinity, drought, and other stresses to plants (McNeil et al 2001).

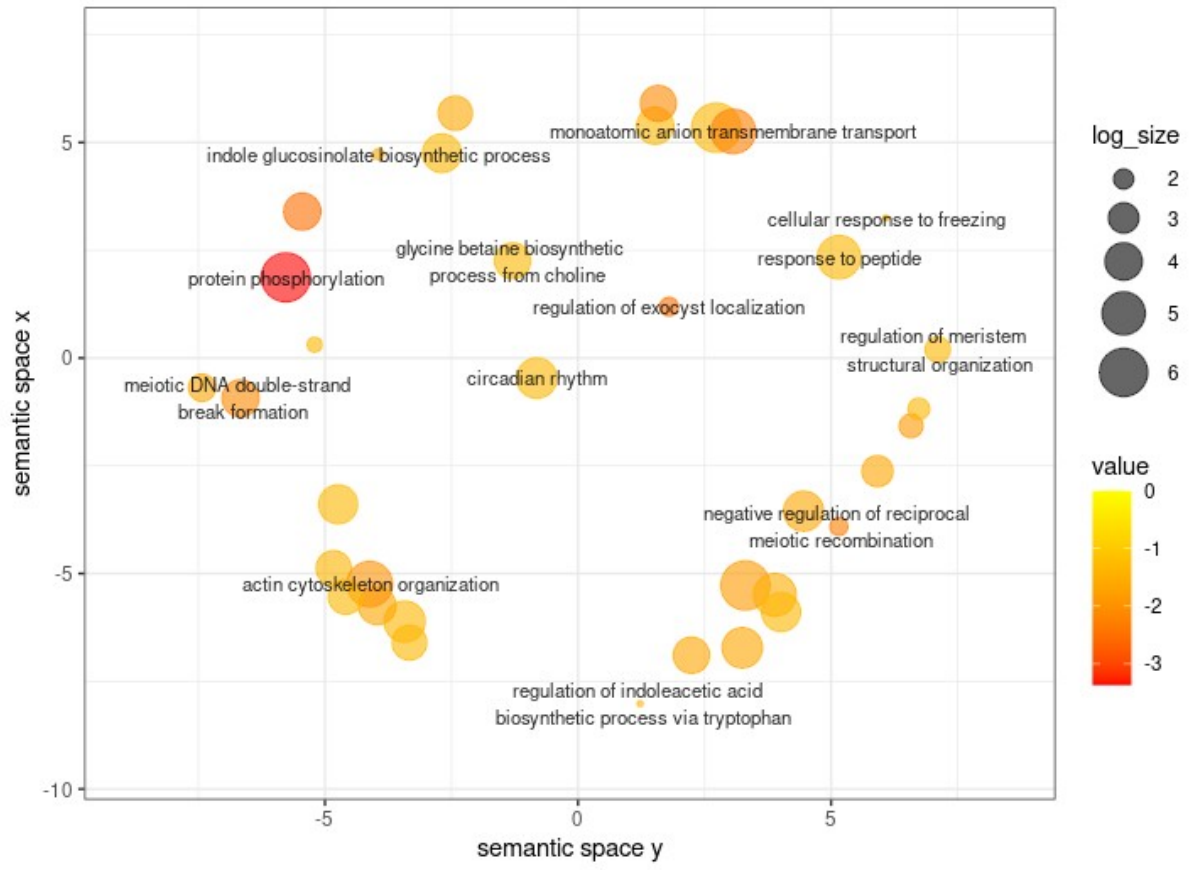

Figure 15 - GOs of the missense SNPs *A. sisalana*-exclusive genes in comparison hybrids (H400L and H11648)

For the high impact indels list (Figure 16; supplementary table 6), we can highlight the glucosylceramide catabolic process (GO:0006680, p-value 0.0037). Glucosylceramides levels decrease during cold acclimatization (Lynch et al, 2004). Data collected in our lab (not published yet) shows that H400L and H11648 are much more resistant than *A. sisalana* when exposed to cold. Since we are talking about a high impact indel, the proteins from this copy of the gene are most likely impaired in *A. sisalana* , which could help understand its different response cold when compared to the hybrids. However, we need to further investigate if there are the other copies of this gene in the *A. sisalana* genome.

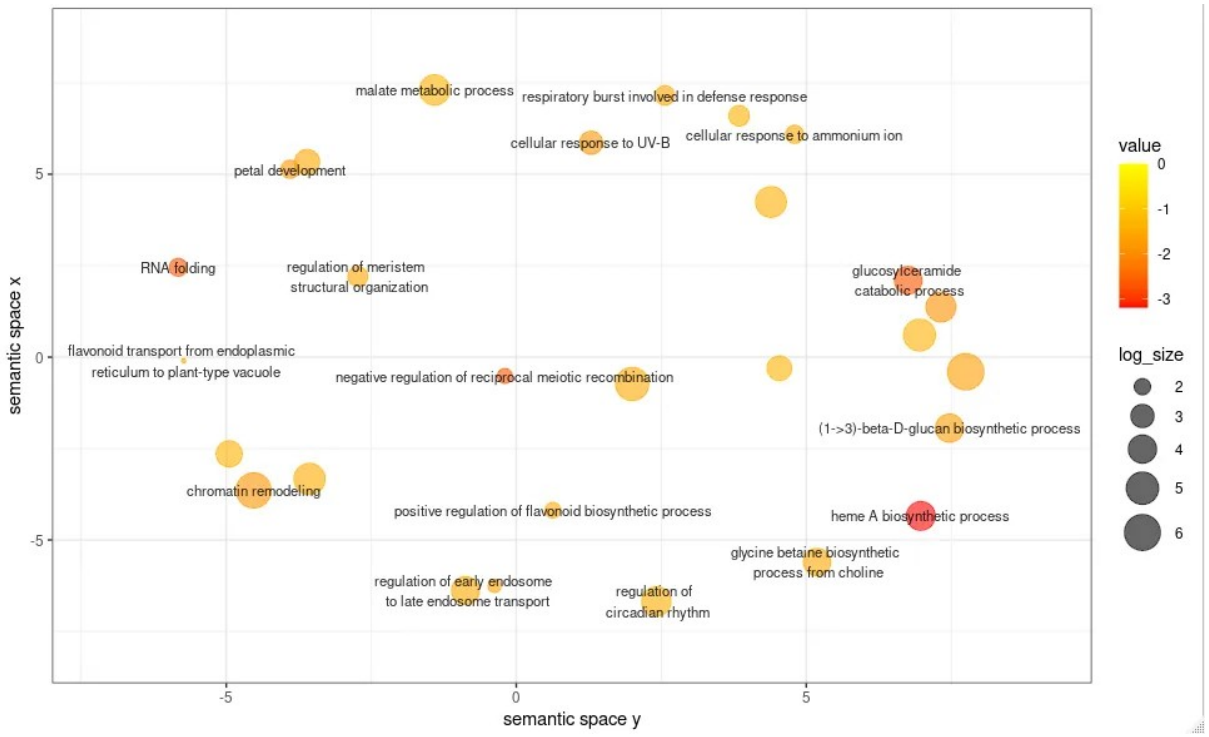

Figure 16 - GOs of the high impact indels *A. sisalana*-exclusive genes in comparison hybrids (H400L and H11648)

For the hybrid missense SNPs exclusive gene list (Figure 17; supplementary table 7), we can highlight the biological process: response to strigolactone (GO:1902347, p-value 0.0109). Strigolactones, signaling compounds made by plants, function as endogenous hormones regulating plant development and growth (Smith et al, 2014). According to Shu et al, strigolactones may be unfavorable for bulbil development, as they act alongside auxin in inhibiting bulbil initiation (Shu et al 2024).From data collected in our lab (not published yet), we know that H11648, H400L and *A. sisalana* grow a different quantity of bulbils, with different weights and sizes throughout the year. The data showed that while H11648 grew 20 bulbils with 1 kg, H400L grew 37 bulbils with 7,4kg and A. sisalana 58 bulbils with 1,6kg. It is interesting to note that having a SNP in a gene that is associated with response to strigolactone may interfere in the response to this hormone in the hybrid plants (H400L and H11648) and consequently make them differ from *A. sisalana* in the process of growth and development of bulbils.

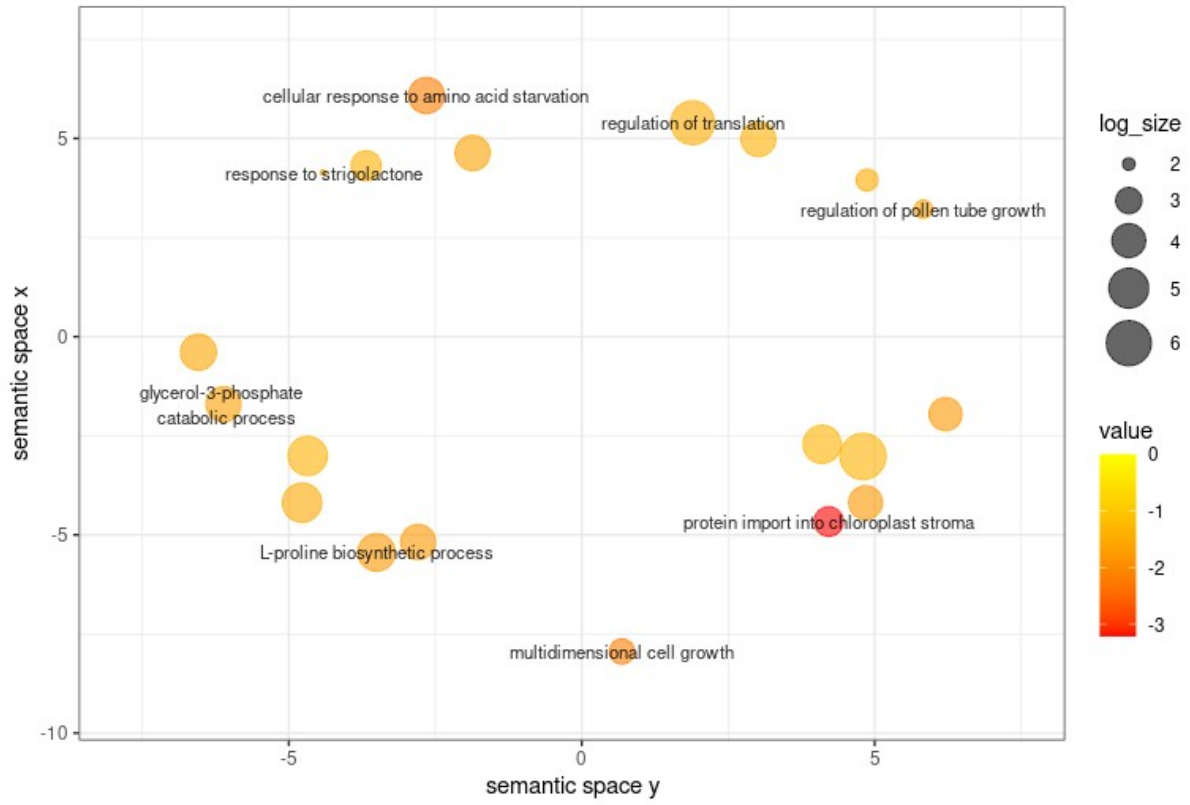

Figure 17 - GOs of the missense SNPs hybrids-exclusive genes in comparison to *A. sisalana*.

For the hybrids high impact exclusive gene list (Table 6), we can highlight the biological process defense response to fungus (GO:0050832, p-value 0.049) and alkaloid metabolic process (GO:0009820, p-value 0.04091). Both processes are linked to one of our phenotypes of interest: defense against pathogens. Alkaloids, as secondary metabolites, play essential roles in various plant processes, including pollinator attraction, seed dispersal, and protection against pathogens (Pereira et al., 2023). Since, there is a high impact indel in this gene, their proteins are most likely impaired in *A. sisalana* individuals. Analyzing if there are other copies of this gene in the genome, can help us understand if this Indel is affecting *A. sisalana's* defense against pathogens when compared to H400L and H11648.

GO ID	Term	Adjusted p-value	Genes
GO:0000493	box H/ACA snoRNP assembly	0.00066	AgveH2v21102813m

GO:0006261	DNA-templated DNA replication	0.01790	AgveH2v21008473m
GO:0006637	acyl-CoA metabolic process	0.02016	AgveH2v21045049m
GO:0009820	alkaloid metabolic process	0.04091	AgveH2v21091935m
GO:0050832	defense response to fungus	0.04972	AgveH2v21110166m

Table 6 - GOs of the high impact indels Hybrids-exclusive genes in comparison to *A. sisalana*.

3.5. Conclusion

Population genomics and high impact mutation analysis gives us important insights into genetic diversity and potential targets for genetic engineering in Agave, especially focusing on improving biomass productivity and stress tolerance for biofuel production.

The population genomics analysis revealed that H11648 showed greater genetic diversity compared to H400L, suggesting that H400L may have suffered a more recent bottleneck or founder event. The significant Fst value of 0.25 between H11648 and H400L and the clear clustering in the PCA indicated a genetic divergence between these genotypes, despite their close genetic relationship. The Admixture analysis also demonstrated the genetic structure of these populations, with clear differentiation among H11648, H400L, and *A. angustifolia*. Additionally, the phylogenomics analysis provided a better understanding of the relationships between these populations.

In parallel, the population genetics analysis of *A. sisalana* individuals from the three production regions showed significant genetic differentiation and moderate genetic diversity. The CF and VA populations had higher genetic variability (He = 0.399, 0.391; Ho = 0.156, 0.191) compared to LB, which showed the lowest diversity (He = 0.270, Ho = 0.104). The pairwise Fst values also showed this genetic divergence, with a high differentiation between LB and VA (Fst = 0.2698) and moderate differentiation between CF and VA (Fst = 0.1747). The PCA and admixture analysis indicated the genetic structure and separation between populations, which is consistent with propagation characteristics for these cultivated species. These are indications that the gene flow patterns could be caused by asexual reproduction, sterile conditions, and restricted pollinator migration. These results give us an insight

in the impact of geographic and reproductive factors on genetic variation, and provide ideas for future crop productivity improvement.

Through the analysis of non-synonymous SNP and Indel variants, we identified genetic mutations exclusive to each genotype, H400L, H11648 and *A. sisalana* and improved our understanding of molecular mechanisms associated with desirable traits: number of leaves (biomass productivity), drought tolerance and fungal diseases tolerance.

Furthermore, our enrichment analysis revealed biological processes associated with the identified genetic variants, such as glutathione catabolic processes, defense response to fungus, cellular responses to iron starvation, acyl-CoA metabolic process, phosphate ion transport, alkaloid metabolic process, and response to strigolactone. These findings provide potential ideas for further research and genetic engineering for improving crop productivity in Agave species.

4. CHAPTER II: COMPARATIVE GENOMICS BETWEEN HIGH FIBER CONTENT AGAVES (H11648 AND *A. SISALANA*) AND HIGH SUGAR CONTENT AGAVES (*A. TEQUILANA*)

In the previous chapter, we used population genomics and SNP/Indels analysis applied to ddRADseq data from 3 populations: H11648, H400L and H11648 to analyze the structure and diversity between the populations and the similarities and differences that can be associated with the different phenotypes important in fiber production in Brazil: biomass productivity, disease resistance and drought tolerance. In this chapter, we aim at expanding our comprehension of the genotypes important to fiber production: H11648 and *A. sisalana*, using comparative genomics with whole genome data. Additionally, we expand our analysis to include *A. tequilana*, enabling a comparison between genotypes with high fiber content (H11648 and *A. sisalana*) and those with high sugar content (*A. tequilana*), a key trait for biofuel production.

4.1 Introduction

4.1.1 Plant Genomes

Understanding the genome of an organism is the first step to comprehend what mechanisms underlie phenotypes of interest. However, it is not an easy task. Sequencing and assembling plant genomes is challenging due to their large size, high repeat content, significant heterozygosity, and polyploidy (Xie et al., 2024; Li et al., 2017). Advances in next-generation sequencing (NGS) provided us with a greater volume of data at a lower cost, which makes the process easier. High-quality reference genome assemblies are the key to advancing plant breeding programs, since this helps in the identification and selection of favorable genes associated with desirable traits such as increased biomass yield, tolerance to environmental stresses, and resistance to diseases caused by pathogens (Li et al., 2017).

Genome assembly is the process of reconstructing the complete genomic sequence of an organism from short DNA fragments generated during the sequencing process with the goal of obtaining a complete and accurate representation of the whole genome (Sohn et al., 2018). However, despite advances made in next-generation sequencing technologies, de novo genome assembly remains a challenge, especially for complex plant genomes. These complexities

normally result in very fragmented assemblies at even high coverage when using only short-read sequencing (Li et al., 2017; Sohn et al., 2018). Therefore, the designing of new bioinformatics methods, using comparative genomic information and reference-guided algorithms, is important for improving genome quality. One of the pipelines that can be followed is the Polyploid Gene Assembler (PGA), which is dedicated to gene assembly, including exons, introns, UTRs, and promoters, by integrating software for read mapping, de novo assembling, and scaffolding (Nascimento et al, 2016).

Following the assembly of the genome, the annotation process is used to explain the gene structure and assign gene functions to a genome sequence, which is crucial to understanding the genetic makeup of an organism (Bolger et al., 2018). The process consists of two major components which includes: structural annotation and functional annotation. First, we have to recognize and mask noncoding regions (repeats). Then, we make predictions for the spatial position of genetic features, including protein-coding regions, promoters, and exon-intron boundaries, among others. This step relies on ab initio methods, which predict genes based on patterns in the DNA sequence, and on evidence-based approaches, including aligning the genome to known sequences from other organisms or using transcriptomic data (Vuruputoor et al., 2023). The following step is functional annotation, which involves defining the biological functions for the identified genes. It includes assigning potential functions to genes by comparing them to previously known genes, protein domains, and other functional elements documented in databases (Bolger et al, 2018; Vuruputoor et al, 2023).

Despite the complexity of plant genomes, involving polyploidy and widespread structural variations, advances in long-read sequencing technologies and bioinformatics tools are making it possible to overcome these challenges, thus pushing forward research in plant genomics and its application in sustainable agriculture.

4.1.2 Comparative Genomics

Comparative genomics is the process of comparing two or more genomes to explain their similarities and differences (Wei et al., 2002). It allows for the discovery of both conserved and divergent elements, the identification of DNA regions under purifying selection, it helps distinguish functional from non-functional sequences, and

eventually it can clarify the genetic basis of phenotypic variation (Hardison et al, 2003). Comparative genomics is based on the presumption that sequences conserved across multiple or distantly related species are likely under evolutionary constraint, suggesting they serve a biological function (Alföldi et al, 2013).

An essential part of comparative genomics includes the analysis of orthologs and the study of expanded and contracted gene families. Orthologs are genes in different species that have evolved from a common ancestral gene by the action of speciation, and their study is fundamental in understanding both functional conservation and divergence among different lineages (Conte et al., 2008). Gene families are groups of genes showing high degrees of sequence homology. These families can either be expanded or contracted through events of gene acquisitions or deletions, respectively, often as a consequence of diversification and adaptation allowing us to infer the importance of their biological functions in the evolution and adaptation of the species being studied (Casola et al., 2019).

By using comparative genomics, we can obtain a deeper level of understanding of evolutionary mechanisms, such as gene retention, functional specialization, and the acquisition of new traits. These analyses help explain how genomes change and adapt to environmental changes, and are fundamental for improving crop productivity.

In the context of Agaves, we can use comparative genomics to understand the similarities and differences underlying traits important for commercial purposes: fiber and biofuel production. Using whole genome data and comparative genomics, we can search for genes and gene families that are exclusive to each genotype and help us understand the mechanisms responsible for the phenotypes: biomass productivity, disease resistance, drought tolerance and fiber/sugar content.

4.2 Objectives

In the context of this chapter, we use the potential of Whole Genome Sequencing and comparative genomics to examine the differences between the genomes from genotypes relevant to bioenergy production (*A. sisalana,* H11648 and *A. tequilana*). These analyses can help us clarify the biological processes associated with specific phenotypes biomass productivity, sugar content, drought tolerance and disease resistance, and possibly offer potential gene targets for genetic engineering and future advancements in crop productivity.

4.2.1 Specific Objectives:

Use comparative genomics to analyze the difference between genotypes relevant to biomass production (*A. sisalana*, H11648 and *A. tequilana*)

- Perform the draft genome assembly of *A. sisalana* and genome annotation
- Comparative genomics analysis between (*A. sisalana*, H11648 and *A. tequilana*) based on exclusive and expanded gene families.

4.3 Material and Methods

4.3.1 Data and Publicly available genome data

For the *A. sisalana* genome assembly, we used DNA samples from an individual collected in the Bahia semiarid field trip for Marina Püpke Marone PhD thesis mentioned in section 3.3.1. The sample was sequenced and we obtained Illumina HiSeq paired-end reads of 150 bp. The *A. sisalana* RNA-seq transcripts used for the robustness of the genome assembly, was retrieved from the work of Raya et al 2021.

For the other species used in the comparative genomics analysis, we retrieve the genome and proteome information from public databases: H11648 from Yang et al 2024; *Agave tequilana and Asparagus officinalis* from JGI repository (Nordberg et al 2014); and the *Phalaenopsis equestris* from NCBI (Cai et al 2015).

4.3.2 Genome Assembly

We performed the assembly of the A. sisalana genome using the PGA v1.2 pipeline (Nascimento et al., 2019). This pipeline consists of assembling only the genic regions, using DNA-seq and RNA-seq data, and the genome from a reference species. For our analysis, we used the genome from *A. tequilana* as reference. This pipeline was constructed and validated for assembling the *Saccharum spontaneum* genome, but it has been previously used to assemble other complex plant genomes as well.

4.3.3 Genome Structural Annotation

Before performing genome annotation, RepeatModeler v2.0.5 (Flynn et al, 2020) and Repeatmasker v4.1.7 (Smit et al, 2015) were used to identify and mask repetitive elements respectively. Then, we used Braker v3.0.8 (Stanke et al, 2008) with RNA-Seq evidence for the genome annotation. Finally, we used BUSCO v5.6.1 (Manni et al, 2021) to verify the completeness of the genome assembly and

annotation. A BUSCO script was used to create the graphical representation of the results.

4.3.4 Comparative and evolutionary analysis

In order to perform a robust comparative genomics analysis of *A. sisalana* (5n), H11648 (2n) and *A. tequilana* (2n), we chose other two publicly available genome assemblies, *Asparagus officinalis* (2n) *and Phalaenopsis equestris* (2n), which are species relatively close to agaves. The proteome file for *A. sisalana* was obtained from Braker v3.0.8, while the proteomes from the other species were downloaded from the public databases. We set *Phalaenopsis equestris* as an outgroup.

The proteins were clustered and orthogroups were identified using OrthoFinder v2.5.5 (Emms et al, 2019). Single-copy orthologous genes were selected and aligned using MAFFT v7.20 (Katoh et al, 2013) with parameter “--globalpair --maxiterate 1000”. Then, with AMAS (Borowiec et al, 2016) we concatenated all alignments in a supertree, which was used as input to IQTREE v. v2.0.3 (Minh et al, 2020) with parameter “-b 1000 -m TEST”. An in-house python script was created to identify exclusive orthogroups for each genotype.

4.3.5 Expanded and contracted gene families

For the analysis of gene families, we used CAFE v.5.0.0 (De Bie et al, 2006), with the gene count file generated by Orthofinder as input. We performed the analysis with different λ values (parameter “-k”) and chose the 0.4 model for downstream analysis. We considered only families with p-value < 0.01 to have undergone expansion and contraction.

4.3.6 Functional Annotation and Gene Ontology (GO)

The *A. sisalana*, H11648 and *A. tequilana* proteins were annotated with PANNZER2 (Törönen et al., 2018) and the Biological Processes (BP) GO terms were selected to be used as background for the enrichment analysis. The exclusive orthogroups and extended and contracted families of H11648, *A. sisalana* and *A. tequilana* were submitted to GO (Gene Ontology) enrichment analysis using topGO package v2.48.0 (Alexa et al., 2006).

4.4 Results

4.4.1 Genome Assembly

The assembly of the *A. sisalana* genome was performed with the PGA pipeline, using DNA Illumina HiSeq paired-end reads of 150 bp and RNA-seq as input, and the *A. tequilana* genome as reference. We obtained a draft genome with a total length of 535,217,891 pb, 511,436 contigs and N50 of 1,087 pb. It is important to note that this genome was assembled using the PGA pipeline, which consists of assembling only the genic regions, which justifies the fragmented pieces shown in the assembling statistics above. However, we can see in the next section, in the BUSCO results, that the assembly was able to contemplate a significant number of genes.

4.4.2 Genome Annotation

In order to perform a better-quality genome annotation, we need to first identify and mask repetitive elements. We identified 31% of retroelements and 1,5% of DNA transposons. For comparison, we can analyze the H11648 genome statistics from Yang et al 2024, which identified 67% of retroelements and 9% of DNA transposons.Then, we performed genome annotation with Braker v3.0.8 using RNA-seq as evidence. We examined the quality of genome annotation using BUSCO v5.6.1. The analysis resulted in 31.3% (133) complete and single-copy BUSCOs, 51.7% (220) fragmented BUSCOs, and 16,9% (72) missing BUSCOs (Figure 18).

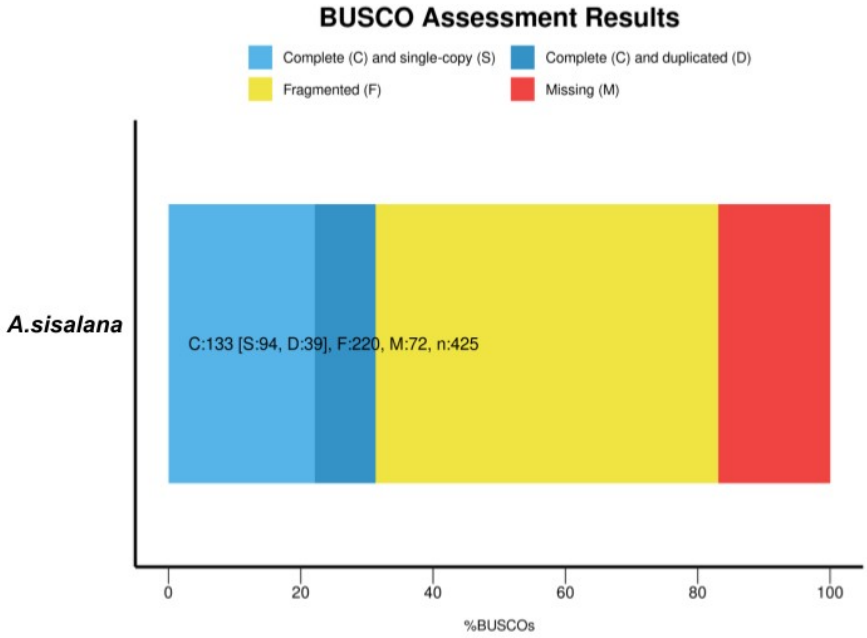

Figure 18 - BUSCO result for the *A. sisalana* genome assembly

The missing BUSCO results of 16,9% show us that although the genome is still in a draft form and fragmented, it still allows us to perform comparative genomics with the other species of interest like H11648 and *A. tequilana*.

4.4.3 Comparative Genomics

4.4.3.1 Orthofinder Results

Comparative genomics consists of comparing two or more genomes to understand the differences and similarities between them. One method to perform this analysis is through identifying orthologs. Our focus was to compare the *A. sisalana*, H11648 and *A. tequilana* genotypes to obtain a better understanding on the differences and similarities of their genome and try to identify the markers behind the phenotypes: biomass productivity, disease and drought resistance. Based on the evolutionary position of the species, we included two more genomes in the analysis: *Asparagus officinalis* (2n) and *Phalaenopsis equestris* (2n) as an outgroup. We clustered the protein-coding genes of these 5 species using Orthofinder v2.5.5. The Orthofinder analysis identified 28,719 orthogroups, of which 6,181 are species-specific. and 9,460 orthogroups have all species present. The overall results per species from the Orthofinder analysis can be seen in table 7.

	H11648	***A officinalis***	***A sisalana***	***A tequilana***	***P equestris***
Number of genes*	58,605	27,395	71,310	42,199	29,894
Number of genes in orthogroups (%)	49,585 (84.6%)	23,628 (86.2%)	58,131 (81.5%)	38,517 (91.3%)	28,341 (94.8%)
Number of unassigned genes (%)	9,020 (15.4%)	3,767 (13.8%)	13,179 (18.5%)	3,682 (8.7%)	1,553 (5.2%)
Number of orthogroups containing species (%)	20,325 (70.8%)	14,663 (51.1%)	23,275 (81.0%)	18,296 (63.7%)	13,253 (46.1%)
Number of species-specific orthogroups	909	663	4,051	279	916
Number of genes in species-specific orthogroups (%)	2,138 (3,6%)	3,481 (12,7%)	16,387 (23,0%)	1,138 (2,7%)	3,309 (11,1%)

* Number of Protein-coding Genes present in the assembled genomes

Table 7 - Overall Orthofinder results per species. The assembled genomes used in this analysis were retrieved from: H11648 from Yang et al 2024; *Agave tequilana* and *Asparagus officinalis* from JGI repository (Nordberg et al 2014); and the *Phalaenopsis equestris* from NCBI (Cai et al 2015).

Using the single-copy genes information, we performed a phylogenetic analysis (Figure 19). The results are compliant with our previous knowledge of the evolutionary relationship between these genotypes with H11648 and *A. sisalana* showing more proximity in comparison with *A. tequilana*, and *P. equestris* behaving as an outgroup.

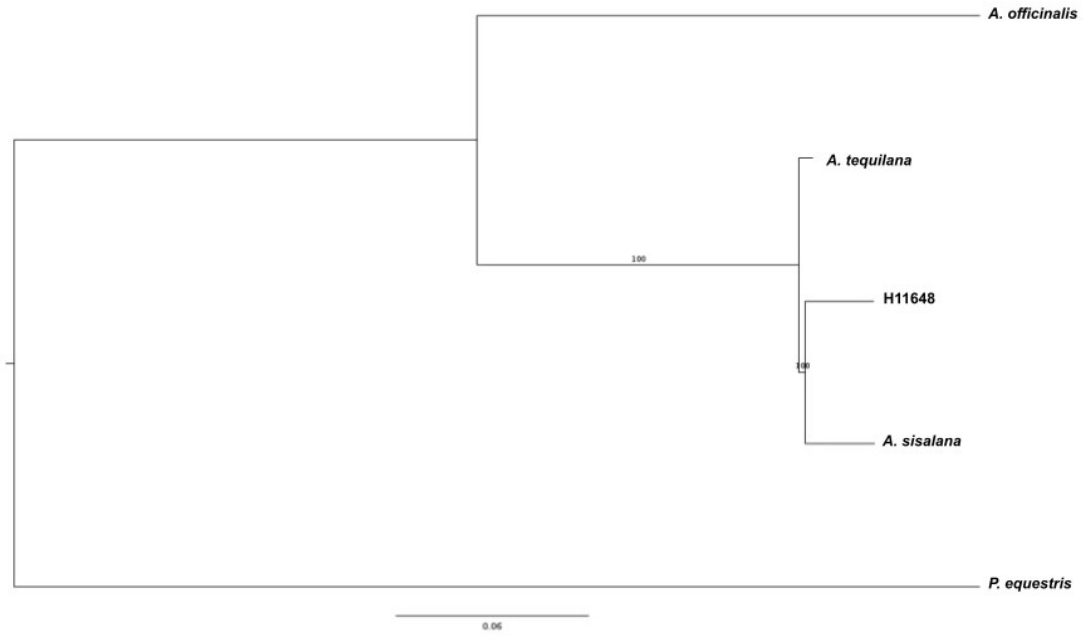

Figure 19 - Phylogenomic tree constructed using 356 single-copy genes identified by Orthofinder. Gene alignments were performed with MAFFT using the parameters '--globalpair --maxiterate 1000'. The individual alignments were concatenated into a supermatrix using AMAS and subsequently used as input for tree inference in IQTREE.

4.4.3.2 Exclusive Orthologs

Then, we proceeded to investigate which orthogroups were exclusive for our genotypes of interest. As we can see in the graph from Figure 20, the number of exclusive orthogroups were 4,051 for *A. sisalana*, 909 for H11648 and 279 for *A. tequilana*.

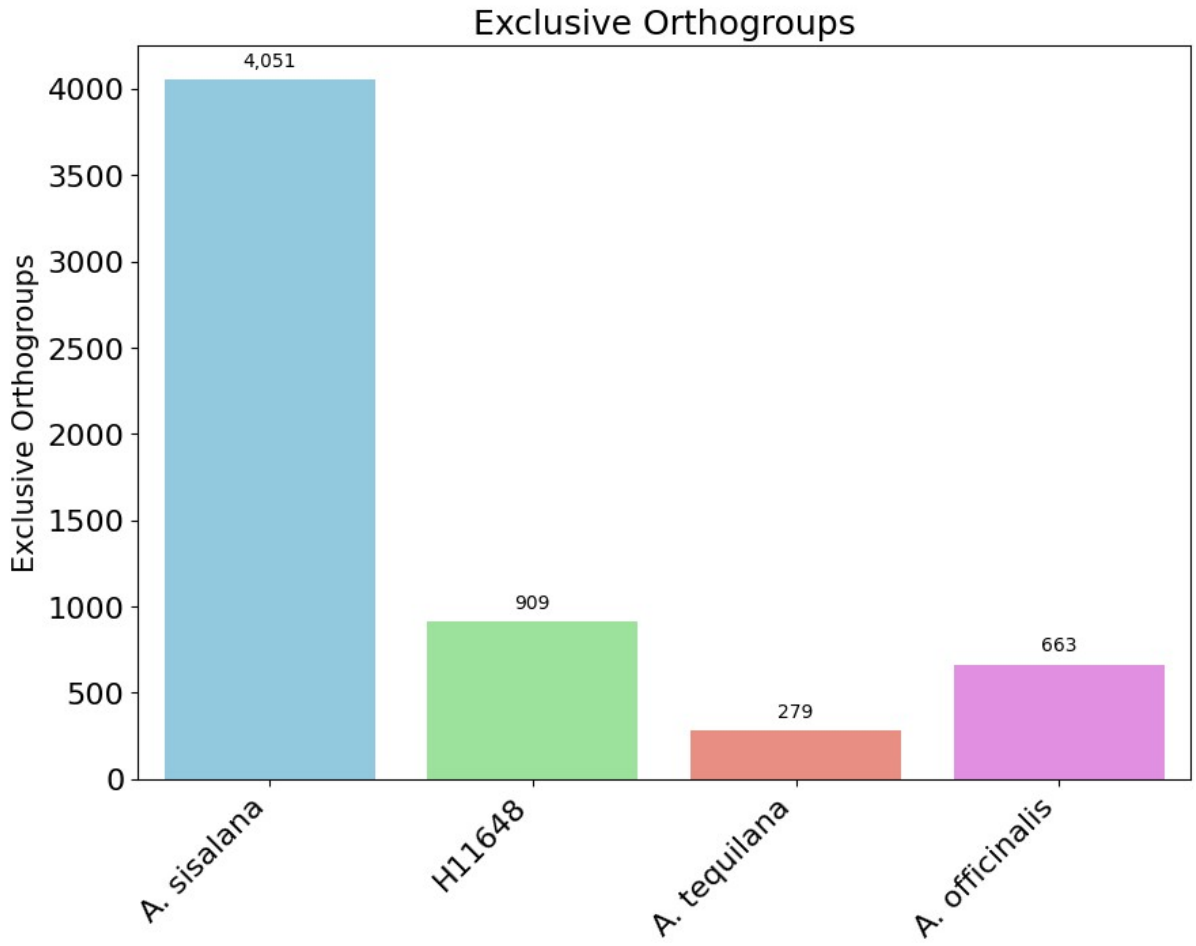

Figure 20 - Exclusive orthogroups

This analysis allows us to identify what groups of genes are exclusive to each genotype and investigate if they are key genes in molecular mechanisms related to the phenotypes we want to compare. So, we used Pannzer software to annotate the biological processes linked to the genes in each exclusive orthogroup and TopGO to do an enrichment analysis and verify what biological processes are enriched in each genotype. We divided the analysis in two main parts: (1) comparison of H11648 and *A. sisalana* focusing on biological processes associated with biomass productivity, disease resistance and drought tolerance; and (2) comparison of *A. tequilana* x (H11648 and *A. sisalana*), focusing on biological processes related to sugar and fiber content.

As shown in Table 1 in section 1.4, H11648 exhibits superior phenotypes compared to *A. sisalana* when focusing on the productivity traits: it has a higher number of leaves (higher biomass productivity), a higher tolerance to drought and a higher resistance to the bole rot disease. So, for the first analysis, we searched for exclusive

orthogroups in H11648 with biological processes that can justify their superior productivity in these traits.

From the enrichment analysis of the exclusive orthogroups of H11648 (Figure 21; Table 8, Supplementary Material), we can highlight response to cold (GO:0009409, p-value 0.01678). As mentioned in the previous chapter, according to experiments conducted in our lab, H11648 is much more resistant to cold than *A. sisalana*, so having an exclusive orthogroup with genes associated with response to cold is a good clue to why this happens. We can also highlight the diterpenoid biosynthetic process (GO:0016102, p-value 0.04073). Studies show that "abiotic stresses such as drought, high salt, high humidity and UV exposure can disturb the biosynthesis of diterpenoids in plants", specifically diterpenoids were accumulated in response to UV irradiation and drought in many monocots (Junze et al 2022).

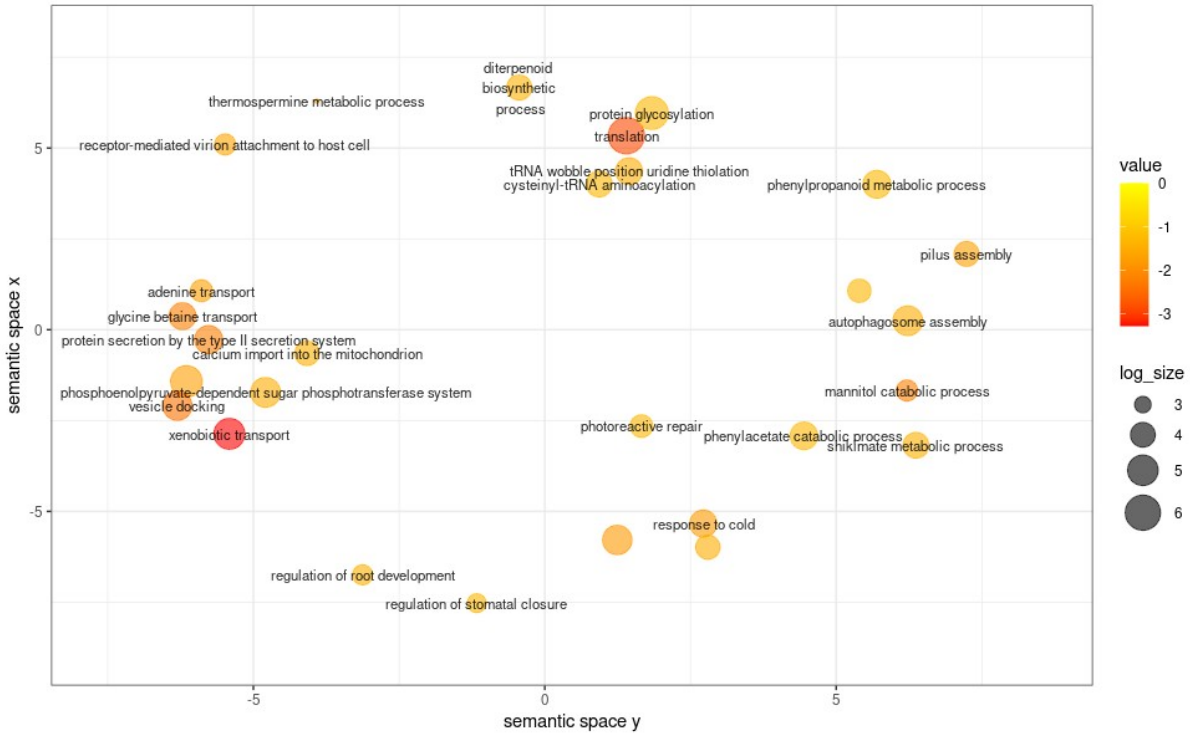

Figure 21 - GOs of *H11648* exclusive orthogroups

Another important biological process enriched in the H11648 exclusive orthogroups, is regulation of stomatal closure (GO:0090333, p-value 0.04998). Stomata are pores on the surface of leaves and stems that regulate gas exchange and water balance in plants, that is, it regulates transpiration and photosynthesis (Lee, J 2010; Lawson, T et al, 2009). In CAM plants, the regulation of stomatal

opening is a crucial mechanism for its water use efficiency because opening the stomata at night minimizes water loss (Davis et al., 2010; Yin et al., 2018). In addition to its association with biomass productivity (photosynthesis) and drought tolerance (transpiration regulation), stomata can be linked to plant immune defense against pathogens, as well. A number of plant pathogens use stomatal pores as entry points of invasion, and the plant can activate mechanisms that closes these pores, stomatal immunity, for defense (Hou, S et al, 2024).

In addition to the enrichment analysis, we searched for other exclusive orthogroups of H11648 related to the phenotypes of interest. For the biological processes linked to biomass productivity, we found exclusive orthogroups in H11648 related to carbon utilization (GO:0015976), plant-type cell wall cellulose biosynthetic process (GO:0052324), photorespiration (GO:0009853), response to auxin (GO:0009733). For the processes related to drought tolerance in the H11648 exclusive orthogroups, we found calcium-mediated signaling (GO:0019722), jasmonic acid-mediated signaling pathway (GO:0009867), response to salt stress (GO:0009651), response to cold (GO:0009409), and trehalose-phosphatase activity (GO:0004805). And for the processes associated with disease resistance in the H11648 exclusive orthogroups, we found phenylpropanoid metabolic process (GO:0009698), glutamate-cysteine ligase activity (GO:0004357), jasmonic acid-mediated signaling pathway (GO:0009867), xenobiotic detoxification by transmembrane export (GO:1990961), peroxidase activity (GO:0004601), and COP9 signalosome (GO:0008180).

From the enrichment analysis of the exclusive orthogroups of *A. sisalana* (Figure 22; Table 9, Supplementary Material), we can highlight cellular response to heat (GO:0034605, p-value 0.0085), which is a process associated with drought resistance and can help us understand why *A. sisalana* is less resistant to drought than H11648. We also found enriched processes associated with biomass productivity such as response to glucose (GO:0009749, p-value 0.0061) and photosynthesis, light harvesting in photosystem I (GO:0009768, p-value 0.0483).

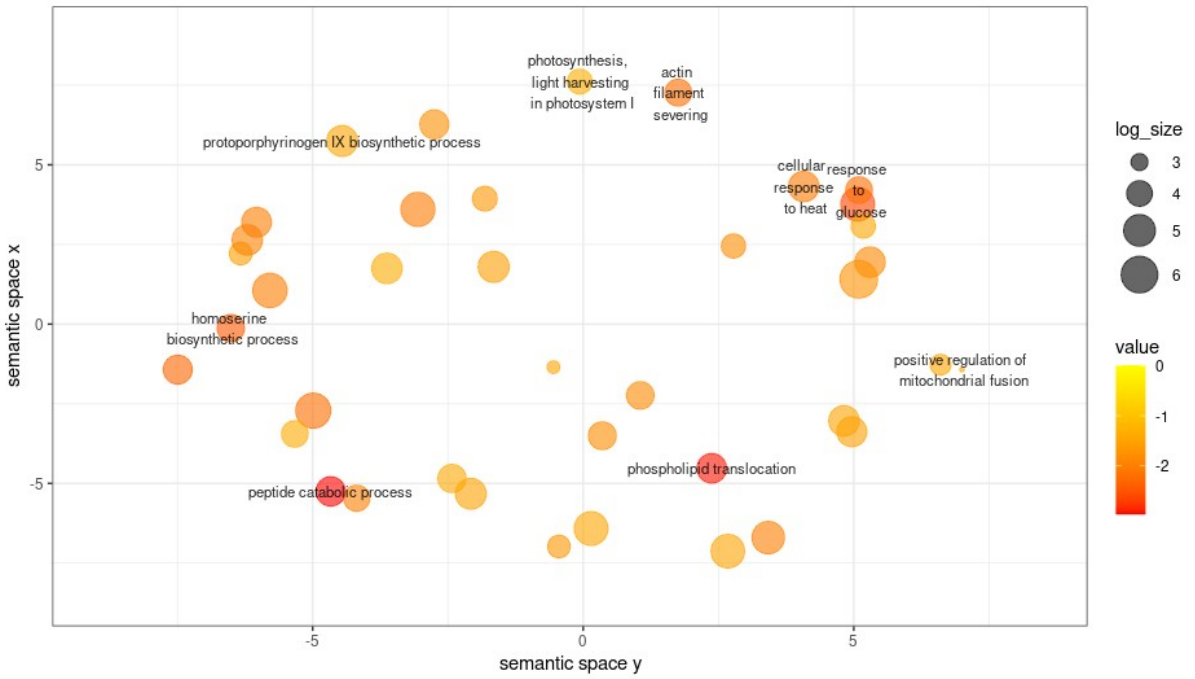

Figure 22 - GOs of *A. sisalana* exclusive orthogroups

Additionally, we searched for these biological processes in the other exclusive orthogroups of *A. sisalana* as well. We found the following processes related to biomass: sucrose synthase activity (GO:0016157), fructokinase activity (GO:0008865), photosynthesis (GO:0015979), cellulose synthase activity (GO:0016759), cellulose synthase (UDP-forming) activity (GO:0016760) lignin biosynthetic process (GO:0009809). Related to drought resistance we found water channel activity (GO:0015250) and abscisic acid-activated signaling pathway (GO:0009738). And finally, related to disease resistance, we identified the biological processes: defense response (GO:0006952) and MAP kinase activity (GO:0004707).

For the second part, we wanted to understand the differences between the genotypes with high sugar content (*A. tequilana*) and the ones with low sugar content but high fiber content (H11648 and A. sisalana). So, we searched for biological processes related to these phenotypes.

The fibers present in Agave species are composed of cellulose (Gebretsadik, T et al, 2023). So, we identified biological processes in the exclusive orthogroups of H11648 and *A. sisalana* that can explain why they have higher fiber content than *A. tequilana*. In H11648 we found: plant-type cell wall cellulose biosynthetic process (GO:0052324) and cellulose microfibril organization (GO:0010215). And in *A. sisalana* we found

orthogroups with genes related to cellulose synthase activity (GO:0016759), cellulose synthase (UDP-forming) activity (GO:0016760).

In *A. tequilana* we found exclusive orthogroups with the following biological processes enriched (Figure 23) (Table 10, Supplementary Material): photosynthesis light reaction (GO:0019684, p-value 0.0010), which is a process related to photosynthesis and might be associated with why this genotype a higher sugar content than H11648 and *A. sisalana*. However, we need to look more deeply into the genes in the orthogroup and the metabolic pathways they participate in to understand this relationship better.

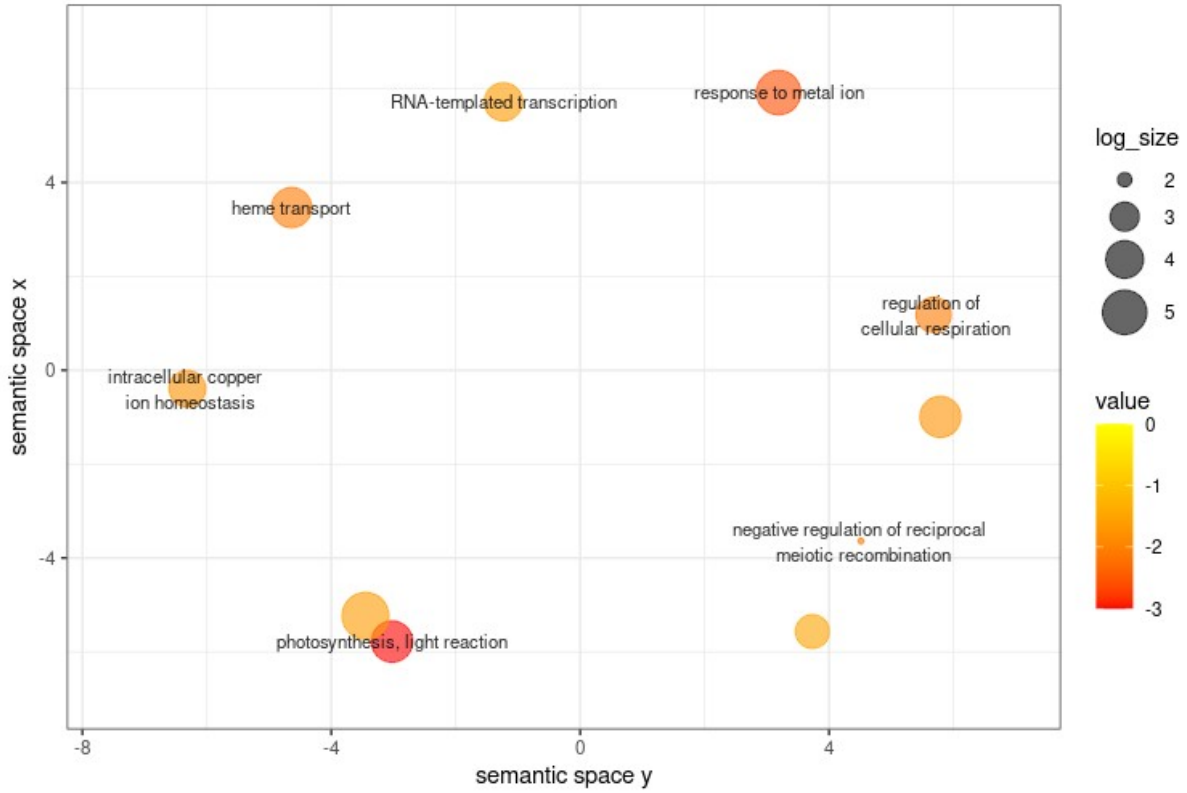

Figure 23 - GOs of *A. tequilana* exclusive orthogroups

4.4.3.3 Expanded and contracted families

Gene families are groups of genes that share a high level of sequence homology. They can expand or contract depending on gene gains or losses caused by diversification and adaptation of the species over time. We analyzed the expanded and contracted gene families with CAFE5 v5.0.0 using the results extracted from Orthofinder v2.5.5. We considered only families with p-value < 0.01 to have undergone expansion and contraction. The results for the expanded and contracted families of each species are shown in Table 8.

Species	Number of expanded families	Number of contracted families
A. sisalana	1,545	2,788
H11648	3,515	751
A. tequilana	1,678	856
A. officinalis	188	3,499
P. equestris	1,505	1,211

Table 8 - Expanded and contracted families for each species

As we can observe in figure 24, H11648 has the highest number of expanded families. One hypothesis for this phenomenon is that since H11648 is a hybrid, the combination of two species likely introduced a significant number of species-specific genes that may have played a key role in the observed expansion. Analyzing the genomes of its parentals *A. angustifolia* and *A. ameniensis* would allow us to better investigate the contribution of each of them in the H11648 genome and the expanded families. Additionally, we note in figure 25, that H11648 and *A. tequilana* have the higher number of families in common. This can be associated with the fact that *A. tequilana* is a domesticated species from *A. angustifolia*, one of H11648 parentals.

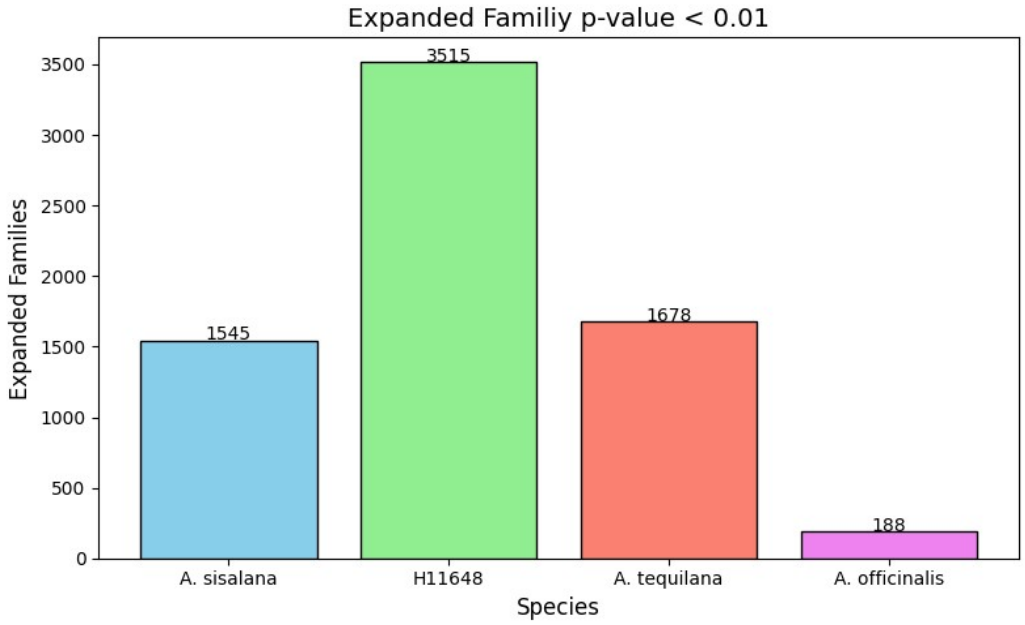

Figure 24 - Expanded families

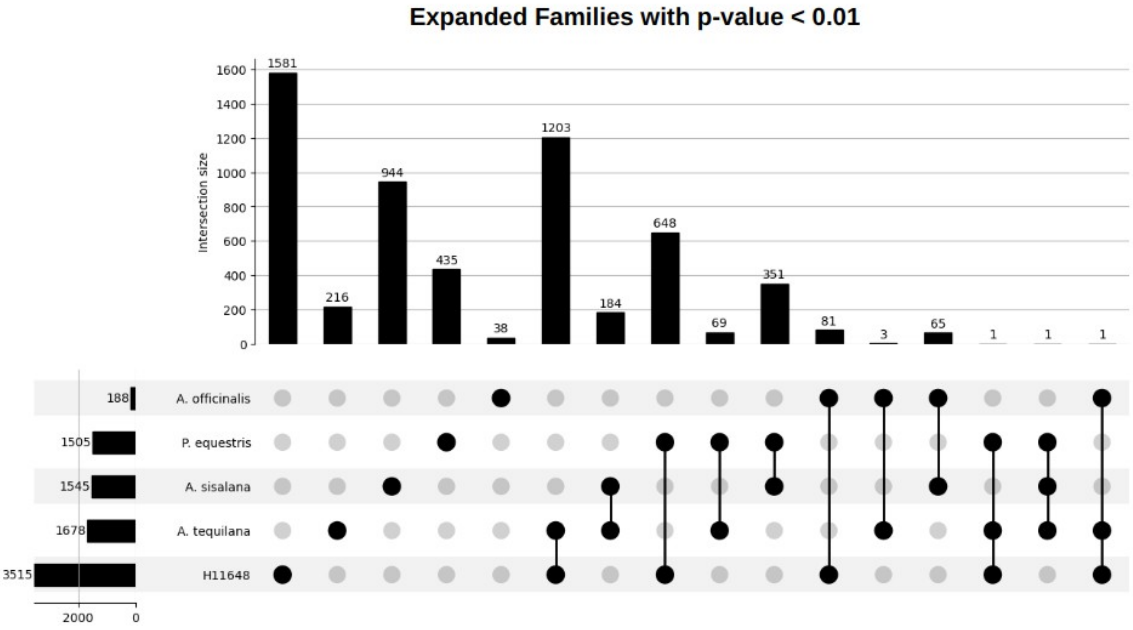

Figure 25 - Comparison between the expanded families (with p-value < 0.01)

In Figure 26, we see that the species with the highest number of contracted families are *A. officinalis* and *A. sisalana*. And we can see in Figure 27, that they have the highest number of contracted families in common too.

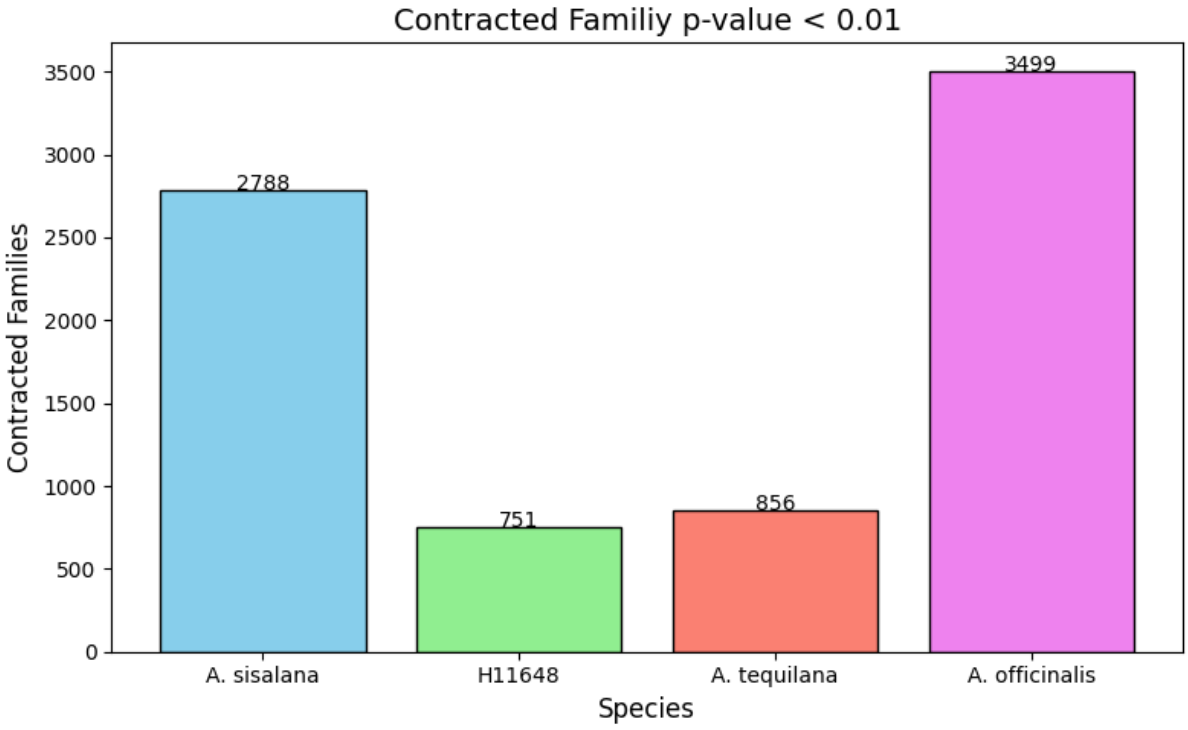

Figure 26 - Contracted families

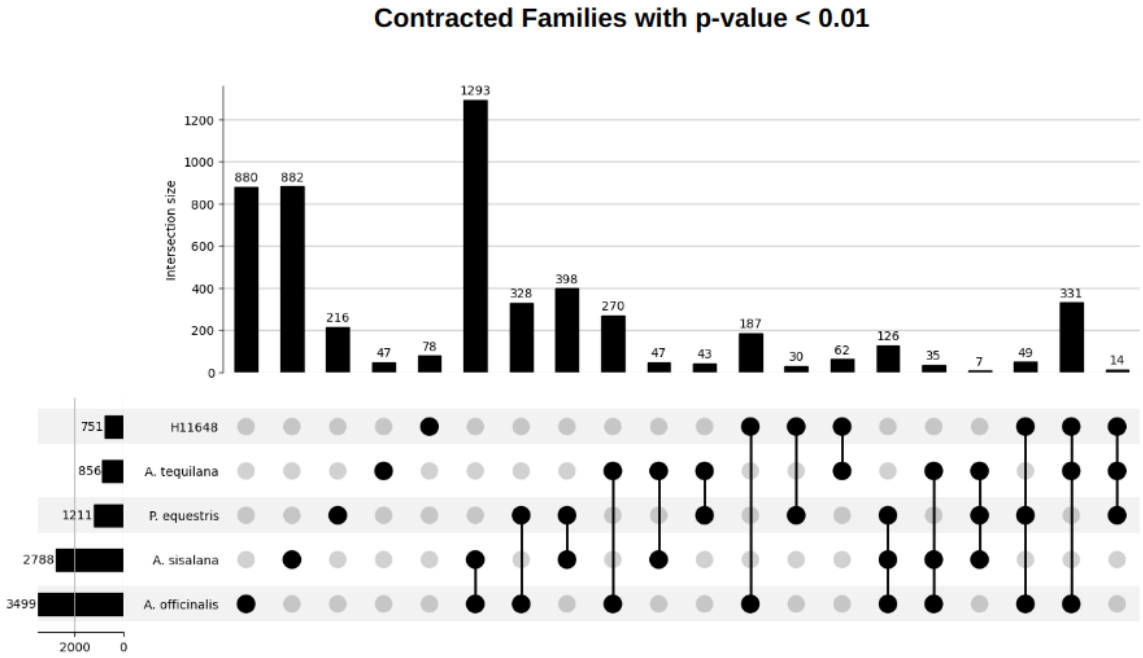

Figure 27 - Comparison between the contracted families (with p-value < 0.01)

The process of expansion and contraction can be associated with diversification and adaptation of the species. By looking at the biological processes related to these families, we can try to identify what genes are related to the adaptation of specific phenotypic traits that differ between our genotypes. For this analysis, we used Pannzer to annotate the biological processes related to the expanded families of H11648, *A. sisalana* and *A. tequilana*. Additionally, we did an enrichment analysis with TopGO to verify what biological processes are enriched in each genotype. We followed the previous section and separated the analysis in two parts: (1) comparison of H11648 and *A. sisalana* focusing on biological processes related to biomass productivity, disease resistance and drought tolerance; and (2) comparison of *A. tequilana* x (H11648 and *A. sisalana*), focusing on biological processes related to sugar and fiber content.

For the first part, in the enrichment analysis of the H11648 expanded families (Figure 28; Table 11 - Supplementary Material), we can highlight the gibberellin catabolic process (GO:0045487, p-value 0.01155). Gibberellin is a hormone essential for many development processes in plants, like shoot growth and xylogenesis, which

are processes directly related to yield and biomass productivity (Yamaguchi 2008, Castro-Camba et al 2022).

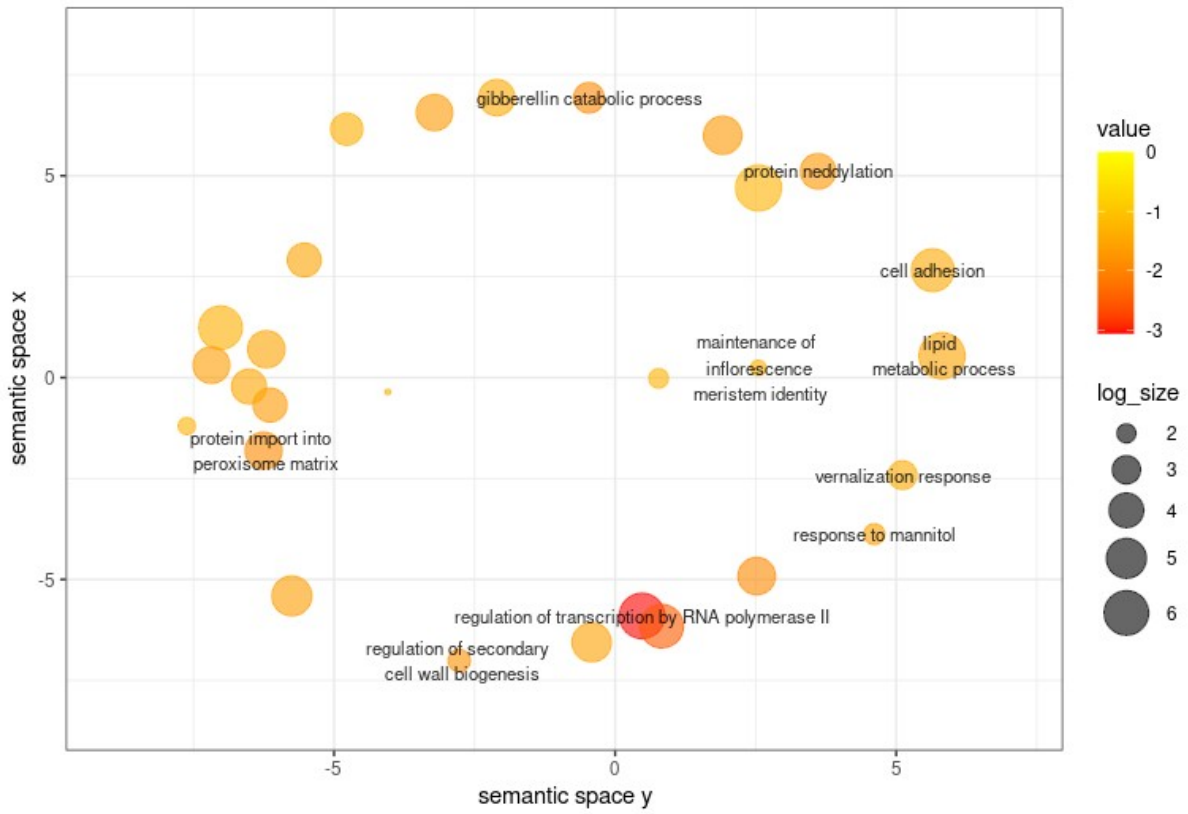

Figure 28 - GOs of the H11648 expanded families

Additionally, we identified biological processes related to biomass productivity found in the other H11648 expanded families: positive regulation of growth (GO:0045927), response to blue light (GO:0009637), glucose metabolic process (GO:0006006), regulation of secondary shoot formation (GO:2000032), regulation of stomatal opening (GO:1902456), plant organ development (GO:0099402), regulation of developmental process (GO:0050793). For the biological processes related to drought tolerance found in the H11648 expanded families we can highlight: regulation of stomatal opening (GO:1902456), stomatal complex development (GO:0010374), cold acclimation (GO:0009631), response to heat (GO:0009408), response to high light intensity (GO:0009644), abscisic acid-activated signaling pathway (GO:0009738). For the biological processes related to disease resistance found in the H11648 expanded families we can highlight: induced systemic resistance, jasmonic acid mediated signaling pathway (GO:0009864); response to other

organism (GO:0051707); response to fungus (GO:0009620), cellular response to phosphate starvation (GO:0016036), response to salicylic acid (GO:0009751).

From the H11648 expanded families, we can highlight the one associated with the jasmonic acid mediated signaling pathway (GO:0009864). The jasmonic acid is an endogenous signaling molecule that mediates diverse responses in the plant's metabolism that can be associated with the phenotypes we want to understand. It can induce stomatal opening, inhibit Rubisco biosynthesis, affect the transport of glucose, which are processes related to biomass productivity (Ruan, J et al, 2019). Additionally, it can induce gene expression prompting plant responses to external damage (mechanical, herbivore, and insect damage) and pathogen infection, which are processes related to disease resistance (Ruan, J et al, 2019). And the biosynthesis of the hormone jasmonic acid is associated with stress response, and numerous studies have substantiated its role in enhancing the stress tolerance of drought-resistant cultivars (Sawar et al., 2019).

For the second part of the GO analysis, we focused on biological processes related to fiber and sugar content. In H11648, we found cellulose synthase (UDP-forming) activity (GO:0016760) and cellulose biosynthetic process (GO:0030244). For the enriched biological processes found in the *A. sisalana* expanded families (Figure 29; Table 12 - Supplementary Material) associated with fiber content we can highlight the cellulose biosynthetic process (GO:0030244, p-value 0.0022), which can help explain their higher fiber content when compared to *A. tequilana*.

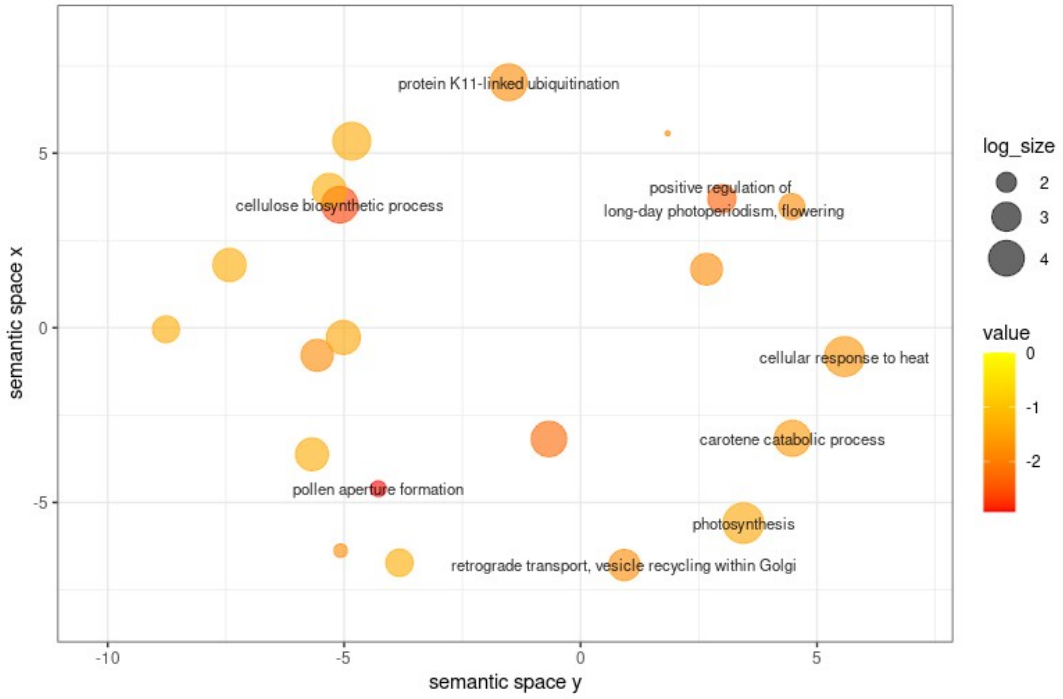

Figure 29 - GOs of the *A. sisalana* expanded families

Other interesting processes found in the *A. sisalana* expanded families are: photorespiration (GO:0009853) and leaf formation (GO:0010338) related to biomass productivity. Response to water deprivation (GO:0009414), cellular response to heat (GO:0034605), trehalose biosynthetic process (GO:0005992) and cellular response to oxidative stress (GO:0034599) related to drought tolerance. And finally, response to aluminum ions (GO:0010044) and systemic acquired resistance (GO:0009627) related to disease resistance.

Finally, we searched for expanded gene families with biological processes that could explain the *A. tequilana* higher sugar content. We found enriched biological processes (Figure 30; Table 13 - Supplementary Material) associated with sugar content: photosynthesis (GO:0015979, p-value 1.9e-05) and photosynthetic electron transport chain (GO:0009767 , p-value 0.02914). Additionally, we found other *A. tequilana* expanded families with genes related to: chloroplast thylakoid (GO:0009534), UDP-glucose transmembrane transporter activity (GO:0005460), intracellular glucose homeostasis (GO:0001678), glucose-6-phosphate 1-epimerase activity (GO:0047938), TRAPP complex (GO:0030008) and photosystem I (GO:0009522), which are also biological processes that can help explain its higher sugar content.

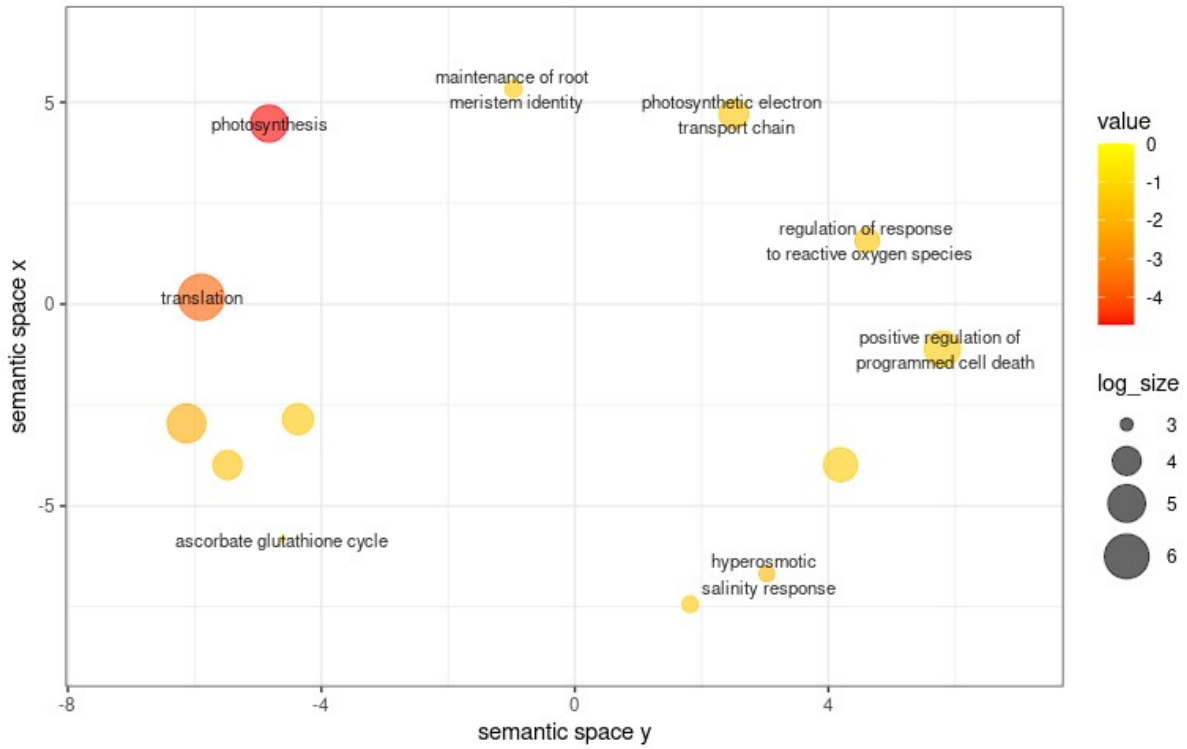

Figure 30 - GOs of the *A. tequilana* expanded families

By analyzing the biological processes of the expanded families of H11648, *A. sisalana* and *A. tequilana* closely, we can identify genes exclusive of each genotype that performs functions associated with production traits: biomass, drought tolerance, disease resistance and sugar/fiber content.

4.5 Conclusion

Comparative genomics is a powerful tool to compare the differences and similarities present in the DNA of the species we want to study. By analyzing orthologs and expanded and contracted families, we can have a better understanding of their genomes and try to identify what differences are responsible for phenotypes of economic interest, specifically biomass productivity, disease resistance and drought tolerance.

In the present chapter, we focused on comparing the *A. sisalana*, H11648 and *A. tequilana* genomes, adding *A. officinalis* and *P. equestris* for a more robust and complete analysis. The results show that A. sisalana has a significantly higher number of exclusive orthogroups compared to H11648 and *A. tequilana*. Regarding the expanded and contracted families, we see that H11648 has the highest number

of expanded families, which could be explained by its hybrid nature. H11648 and A. tequilana have the highest number of families in common, which is consistent with the fact that *A. tequilana* is a domesticated species derived from *A. angustifolia*, one of H11648's parental species.For the contracted families, we have A. sisalana and *A. officinalis* with the highest total number and highest number of families in common.

Finally, we used GO analysis to identify exclusive orthogroups and expanded gene families that have biological processes related to specific production traits: biomass productivity, drought tolerance, disease resistance and sugar and fiber content.

Comparing the two genotypes relevant for fiber production, we found that H11648 have enriched exclusive orthogroups with genes related to stomatal activity. Since stomata is directly related to the process of transpiration, photosynthesis, and defense, we can hypothesize that this orthogroup has great potential for further analysis in understanding the mechanisms that make H11648 have a higher biomass productivity (photosynthesis), drought tolerance (transpiration regulation), immune defense against pathogens.Additionally, we found enriched expanded families with genes associated with the jasmonic acid pathway, which is a molecule that mediates diverse responses in the plant's metabolism associated with the phenotypes above.

Comparing the genotypes focusing on fiber and sugar content, we found enriched exclusive and expanded orthogroups in *A. tequilana* associated with with photosynthesis and respiration, which shows that *A. tequilana* has a number of exclusive genes that participate in the production of glucose and can explain why this genotype has a higher sugar content than H11648 and *A. sisalana*. While in H11648 and *A. sisalana,* we found enriched exclusive and expanded orthogroups associated with cellulose and fiber, which can help us understand why they exhibit phenotypes with high fiber content.

5. GENERAL DISCUSSION AND CONCLUSION

Biofuels are an important strategy used in the attempt of mitigating emissions of greenhouse gases in the atmosphere. Biomass plays a key role in this production chain, but there are limitations with the use of arable land for energy crop plantations. In this context, Agave is a great solution due to their remarkable drought resistance

and capability of thriving in high temperatures, high insolation, and low water availability.

In this master's project, we used population genomics, SNPs/Indels analysis and comparative genomics to understand the biological processes underlying mechanisms related to phenotypes of interest in agave, such as biomass productivity, disease resistance, drought tolerance and sugar/fiber content, which can be targets for further genetic engineering and crop productivity improvement.

Population genomics analyses showed a clear genetic diversity between H11648 and H400L and indicated that H400L may have suffered a more recent founder event. Additionally, we obtained an insight into how geographic and propagation methods can affect the genetic structure of cultivated crops, as shown in the population metrics between the subpopulations of agave sisalana from 3 different production regions.

Through the analysis of non-synonymous SNP and Indel variants, we identified genetic mutations exclusive to each genotype, H400L, H11648 and *A. sisalana* and improved our understanding of molecular mechanisms associated with desirable traits: number of leaves (biomass productivity), drought tolerance and fungal diseases tolerance. GO enrichment identified biological processes related these traits, such as glutathione catabolic processes, defense response to fungus, cellular responses to iron starvation, acyl-CoA metabolic process, phosphate ion transport, alkaloid metabolic process, and response to strigolactone.

In the comparative genomics analysis between *A. sisalana*, H11648 and *A. tequilana*, we identified 28,719 orthogroups, of which 6,818 species-specific and 9,460 orthogroups with all species present. For the gene family analysis, we obtained 1545 expanded families and 2788 contracted families for *A. sisalana*; 3515 expanded families and 751 contracted families for H11648; 1678 expanded families and 856 contracted families for *A. tequilana*. The GO analysis showed H11648 exclusive orthogroups associated with the jasmonic acid-mediated signaling pathway and expanded families with genes associated with regulation of stomatal opening and stomatal complex development. These results help us analyze the H11648 superior productivity phenotypes when compared to *A. sisalana*.

Overall, this study gives a comprehensive understanding of the genetic factors driving the important traits in Agave, allowing for insights that can be used for future genetic engineering and crop improvement that aims to improve the potential for biofuel production. With this, we can contribute to the development of Agave cultivars

that are better suited for biofuel production in semiarid regions, and support the transition to more sustainable energy systems and reduce dependence on fossil fuels.

6. REFERENCES

Alexander, David H.; Novembre, John; Lange, Kenneth. Fast Model-Based Estimation Of Ancestry In Unrelated Individuals. Genome Research, V. 19, N. 9, P. 1655-1664, 2009.

Alföldi, J.; Lindblad-Toh, K. Comparative Genomics As A Tool To Understand Evolution And Disease. Genome Research, V. 23, N. 7, P. 1063-1068, 2013.

Alves-Pereira, A. Et Al. A Population Genomics Appraisal Suggests Independent Dispersals For Bitter And Sweet Manioc In Brazilian Amazonia. Evol Appl 13: 342–361. 2020.

Amiteye, Samuel. Basic Concepts And Methodologies Of Dna Marker Systems In Plant Molecular Breeding. Heliyon, V. 7, N. 10, 2021.

Andrews, Kimberly R. Et Al. Harnessing The Power Of Radseq For Ecological And Evolutionary Genomics. Nature Reviews Genetics, V. 17, N. 2, P. 81-92, 2016.

Antunez-Sanchez, Javier Et Al. A New Role For Histone Demethylases In The Maintenance Of Plant Genome Integrity. Elife, V. 9, P. E58533, 2020.

Armisén, D.; Lecharny, A.; Aubourg, S. Unique Genes In Plants: Specificities And Conserved Features Throughout Evolution. Bmc Evolutionary Biology, V. 8, P. 1-20, 2008.

Asai, Tsuneaki Et Al. Map Kinase Signalling Cascade In Arabidopsis Innate Immunity. Nature, V. 415, N. 6875, P. 977-983, 2002.

Assenov, Yassen Et Al. Computing Topological Parameters Of Biological Networks. Bioinformatics, V. 24, N. 2, P. 282-284, 2008.

Bader, Gary D.; Hogue, Christopher Wv. An Automated Method For Finding Molecular Complexes In Large Protein Interaction Networks. Bmc Bioinformatics, V. 4, P. 1-27, 2003.

Boatwright, J. L. Et Al. Sorghum Association Panel Whole-Genome Sequencing Establishes Cornerstone Resource For Dissecting Genomic Diversity. The Plant Journal, V. 111, N. 3, P. 888-904, 2022.

Bolger, M. E.; Arsova, B.; Usadel, B. Plant Genome And Transcriptome Annotations: From Misconceptions To Simple Solutions. Briefings In Bioinformatics, V. 19, N. 3, P. 437-449, 2018.

Borland, A. M.; Griffiths, H.; Hartwell, J.; Smith, J. A. C. Exploiting The Potential Of Plants With Crassulacean Acid Metabolism For Bioenergy Production On Marginal Lands. Journal Of Experimental Botany, V. 60, P. 2879–2896, 2009.

Borowiec, M. L. Amas: A Fast Tool For Alignment Manipulation And Computing Of Summary Statistics. Peerj, V. 4, P. E1660, 2016. Disponível Em: <Https://Doi.Org/10.7717/Peerj.1660>.

Broad Institute. Picard Toolkit. 2019. Github Repository. Available At: Https://Broadinstitute.Github.Io/Picard/. Accessed On: 20 July 2024.

Cabrera-Toledo, Dánae Et Al. Genomic And Morphological Differentiation Of Spirit

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W. C., Liu, K. W., ... & Liu, Z. J. (2015). The genome sequence of the orchid Phalaenopsis equestris. Nature genetics, 47(1), 65-72.

Carvalho, Lucas M. Et Al. Analysis Of Protein-Protein Interaction And Weighted Co-Expression Networks Revealed Key Modules And Genes In Multiple Organs Of Agave Sisalana. Frontiers In Chemical Engineering, V. 5, P. 1175235, 2023.

Casola, C., & Lawing, A. M. (2019). The nonrandom evolution of gene families. American Journal of Botany, 106(1).

Castro-Camba, R., Sánchez, C., Vidal, N., & Vielba, J. M. (2022). Plant development and crop yield: The role of gibberellins. Plants, 11(19), 2650.

Chaney, L.; Sharp, A. R.; Evans, C. R.; Udall, J. A. Genome Mapping In Plant Comparative Genomics. Trends In Plant Science, V. 21, N. 9, P. 770-780, 2016.

Chin, Chia-Hao Et Al. Cytohubba: Identifying Hub Objects And Sub-Networks From Complex Interactome. Bmc Systems Biology, V. 8, P. 1-7, 2014.

Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., & Jackson, S. A. (2015). Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. Molecular plant, 8(6), 831-846.

Conte, M. G., Gaillard, S., Droc, G., & Perin, C. (2008). Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. BMC genomics, 9, 1-16.

Coudert, E. Et Al. Annotation Of Biologically Relevant Ligands In Uniprotkb Using Chebi. Bioinformatics, V. 39, P. Btac793, 2023.

Coudert, E.; Gehant, S.; De Castro, E.; Pozzato, M.; Baratin, D.; Neto, T.; Sigrist, C. J. A.; Redaschi, N.; Bsarwarridge, A.;Uniprot Consortium. Annotation Of Biologically Relevant Ligands In Uniprotkb Using Chebi. Bioinformatics, Oxford University Press, V. 39, P. Btac793, 2023

Curien, Gilles Et Al. Identification Of Six Novel Allosteric Effectors Of Arabidopsis Thaliana Aspartate Kinase-Homoserine Dehydrogenase Isoforms: Physiological Context Sets The Specificity. Journal Of Biological Chemistry, V. 280, N. 50, P. 41178-41183, 2005.

Dacosta, J. M.; Sorenson, M. D. Ddrad-Seq Phylogenetics Based On Nucleotide, Indel, And Presence–Absence Polymorphisms: Analyses Of Two Avian Genera With Contrasting Histories. Molecular Phylogenetics And Evolution, V. 94, P. 122–135, Jan. 2016.

Davière, J. M., & Achard, P. (2013). Gibberellin signaling in plants. Development, 140(6), 1147-1151.

Davis, Sarah C.; Dohleman, Frank G.; Long, Stephen P. The Global Potential For Agave As A Biofuel Feedstock. Gcb Bioenergy, V. 3, N. 1, P. 68-78, 2011.

Davis, S. C.; Long, S. P. Sisal/Agave. In: Cruz, V. M. V.; Dierig, D. A. (Eds.). Industrial Crops. Handbook Of Plant Breeding, V. 9. New York, Ny: Springer, 2015. Available At: Https://Doi.Org/10.1007/978-1-4939-1447-0_15. Accessed On: 20 Aug. 2024.

Davis, S. C., Simpson, J., Gil-Vega, K. D. C., Niechayev, N. A., Tongerlo, E. V., Castano, N. H., ... & Búrquez, A. (2019). Undervalued potential of crassulacean acid metabolism for current and future agricultural production. Journal of Experimental Botany, 70(22), 6521-6537.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics, 22(10), 1269-1271.

De Souza, Silmara Chaves Et Al. Genetic Divergence In Agave Accessions Through ISSR Markers And Phenotypic Traits. 2018.

Denancé, N.; Szurek, B.; Noël, L. D. Emerging Functions Of Nodulin-Like Proteins In Non-Nodulating Plant Species. Plant And Cell Physiology, V. 55, N. 3, P. 469-474, 2014.

Do, Bich Hang Et Al. Loss-Of-Function Mutation Of Actin-Related Protein 6 (Arp6) Impairs Root Growth In Response To Salinity Stress. Molecular Biotechnology, V. 65, N. 9, P. 1414-1420, 2023.

Dorion, Sonia; Ouellet, Jasmine C.; Rivoal, Jean. Glutathione Metabolism In Plants Under Stress: Beyond Reactive Oxygen Species Detoxification. Metabolites, V. 11, N. 9, P. 641, 2021.

Eguiarte, L. E. Et Al. Evolutionary Ecology Of Agave: Distribution Patterns, Phylogeny, And Coevolution (An Homage To Howard S. Gentry). American Journal Of Botany, V. 108, P. 216–235, 2021. Doi:10.1002/Ajb2.1609.

Emms, David M.; Kelly, Steven. Orthofinder: Phylogenetic Orthology Inference For Comparative Genomics. Genome Biology, V. 20, P. 1-14, 2019.

Epa. Economics Of Biofuels. Available At: Https://Www.Epa.Gov/Environmental-Economics/Economics-Biofuels. Accessed On: July 24, 2024

Fao Agriculture [Www Document] Food And Agriculture Organization (2020) Statistical Database (2020)

Foll, Matthieu; Gaggiotti, Oscar. A Genome-Scan Method To Identify Selected Loci Appropriate For Both Dominant And Codominant Markers: A Bayesian Perspective. Genetics, V. 180, N. 2, P. 977-993, 2008.

Fleming, T. H.; Geiselman, C.; Kress, W. J. The Evolution Of Bat Pollination: A Phylogenetic Perspective. Annals Of Botany, V. 104, N. 6, P. 1017-1043, 2009.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences, 117(17), 9451-9457.

Frankham, R.; Ballou, J. D.; Briscoe, D. A. Introduction To Conservation Genetics. Cambridge: Cambridge University Press, 2002. Doi: 10.1017/Cbo9780511808999.

Fujishige, Naoko Et Al. A Novel Arabidopsis Gene Required For Ethanol Tolerance Is Conserved Among Plants And Archaea. Plant And Cell Physiology, V. 45, N. 6, P. 659-666, 2004.

Gillmor, C. Stewart Et Al. A-Glucosidase I Is Required For Cellulose Biosynthesis And Morphogenesis In Arabidopsis. The Journal Of Cell Biology, V. 156, N. 6, P. 1003-1013, 2002.

Gebretsadik, T. T., Tesfay, A. H., Gebru, A. G., Assayehegn, E., Desta, Y. H., Gebremedhin, K. H., ... & Teklemedhin, T. B. (2023). Characterization and Comparative Insights on Agave Americana and Agave Sisalana Leaf Fibers for High-Performance Applications. Journal of Natural Fibers, 20(2), 2246648.

Granick, E. B. A Karyosystematic Study Of The Genus Agave. American Journal Of Botany, V. 31, N. 5, P. 283-298, 1944.

Grünwald, Niklaus J. Et Al. Best Practices For Population Genetic Analyses. Phytopathology, V. 107, N. 9, P. 1000-1010, 2017.

Hardison, R. C. Comparative Genomics. Plos Biology, V. 1, N. 2, E58, 2003.

He, Weiming Et Al. Vcf2pcacluster: A Simple, Fast And Memory-Efficient Tool For Principal Component Analysis Of Tens Of Millions Of Snps. Bmc Bioinformatics, V. 25, N. 1, P. 173, 2024.

Herlihy, John H.; Long, Terri A.; Mcdowell, John M. Iron Homeostasis And Plant Immune Responses: Recent Insights And Translational Implications. Journal Of Biological Chemistry, V. 295, N. 39, P. 13444-13457, 2020.

Hou, S., Rodrigues, O., Liu, Z., Shan, L., & He, P. (2024). Small holes, big impact: stomata in plant-pathogen-climate epic trifecta. Molecular Plant.

Hua, Lei; Yang, Zhong; Shao, Jiyou. Impact Of Network Density On The Efficiency Of Innovation Networks: An Agent-Based Simulation Study. Plos One, V. 17, N. 6, P. E0270087, 2022.

J.A. Pérez‐pimienta, M.G. López‐ortega, A. Sanchez Recent Developments In Agave Performance As A Drought‐tolerant Biofuel Feedstock: Agronomics, Characterization, And Biorefining Biofuels Bioprod. Biorefin., 11 (2017), Pp. 732-748, 10.1002/Bbb.1776

Jombart, T.; Ahmed, I. Adegenet 1.3-1: New Tools For The Analysis Of Genome-Wide Snp Data. Bioinformatics, Oxford University Press, V. 27, N. 21, P. 3070-3071, 2011. Doi: 10.1093/Bioinformatics/Btr521.

Junze, R. E. N., Yu, W. U., Zhanpin, Z. H. U., Ruibing, C. H. E. N., & Zhang, L. (2022). Biosynthesis and regulation of diterpenoids in medicinal plants. Chinese Journal of Natural Medicines, 20(10), 761-772.

Katoh, K. & Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Krichevsky, Alexander Et Al. C2h2 Zinc Finger-Set Histone Methyltransferase Is A Plant-Specific Chromatin Modifier. Developmental Biology, V. 303, N. 1, P. 259-269, 2007.

Kalinger, R. S.; Pulsifer, I. P.; Hepworth, S. R.; Rowland, O. Fatty Acyl Synthetases And Thioesterases In Plant Lipid Metabolism: Diverse Functions And Biotechnological Applications. Lipids, V. 55, N. 5, P. 435-455, 2020.

Lawson, T., & Morison, J. I. (2004). Stomatal function and physiology. In The evolution of plant physiology (pp. 217-242). Academic Press.

Lee, J. S. (2010). Stomatal opening mechanism of CAM plants. Journal of Plant Biology, 53, 19-23.

Li, C., Lin, F., An, D., Wang, W., & Huang, R. (2017). Genome sequencing and assembly by long reads in plants. Genes, 9(1), 6.

Li, H. A Statistical Framework For Snp Calling, Mutation Discovery, Association Mapping And Population Genetical Parameter Estimation From Sequencing Data. Bioinformatics, Oxford University Press, V. 27, N. 21, P. 2987-2993, 2011.

Li, Heng Et Al. The Sequence Alignment/Map Format And Samtools. Bioinformatics, V. 25, N. 16, P. 2078-2079, 2009.

Li, Heng; Durbin, Richard. Fast And Accurate Long-Read Alignment With Burrows–Wheeler Transform. Bioinformatics, V. 26, N. 5, P. 589-595, 2010.

Li, Li Et Al. The Versatile Gaba In Plants. Plant Signaling & Behavior, V. 16, N. 3, P. 1862565, 2021.

Liu, Y., & von Wirén, N. (2017). Ammonium as a signal for physiological and morphological responses in plants. Journal of Experimental Botany, 68(10), 2581-2592.

Lu, Xiangyi; Ruden, Douglas M. A Program For Annotating And Predicting The Efects Of Single Nucleotide Polymorphisms, Snpef: Snps In The Genome Of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3. 2012.

Luikart G, England Pr, Tallmon D, Jordan S, Taberlet P. The Power And Promise Of Population Genomics: From Genotyping To Genome Typing. Nat Rev Genet. 2003;4(12):981–94. Pmid:14631358

Lynch, D. V.; Dunn, T. M. An Introduction To Plant Sphingolipids And A Review Of Recent Advances In Understanding Their Metabolism And Function. New Phytologist, V. 161, N. 3, P. 667-702, 2004.

Macho, Alberto P. Et Al. Aspartate Oxidase Plays An Important Role In Arabidopsis Stomatal Immunity. Plant Physiology, V. 159, N. 4, P. 1845-1856, 2012.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular biology and evolution, 38(10), 4647-4654.

Manuel L. Robert, K. Yoong Lim, Lynda Hanson, Filipe Sanchez-Teyer, Michael D. Bennett, Andrew R. Leitch, Ilia J. Leitch, Wild And Agronomically Important Agave Species (Asparagaceae) Show Proportional Increases In Chromosome Number, Genome Size, And Genetic Markers With Increasing Ploidy, Botanical Journal Of The Linnean Society, Volume 158, Issue 2, October 2008, Pages 215–222, Https://Doi.Org/10.1111/J.1095-8339.2008.00831.X

Marwal, Avinash; Gaur, Rajarshi Kumar. Molecular Markers: Tool For Genetic Analysis. In: Animal Biotechnology. Academic Press, 2020. P. 353-372.

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics, 36(22-23), 5516-5518.

McNeil, S. D., Nuccio, M. L., Ziemak, M. J., & Hanson, A. D. (2001). Enhanced synthesis of choline and glycine betaine in transgenic tobacco plants that overexpress phosphoethanolamine N-methyltransferase. Proceedings of the National Academy of Sciences, 98(17), 10001-10005.

Minh, Bui Quang Et Al. Iq-Tree 2: New Models And Efficient Methods For Phylogenetic Inference In The Genomic Era. Molecular Biology And Evolution, V. 37, N. 5, P. 1530-1534, 2020.

Morreeuw, Z. P., Escobedo-Fregoso, C., Ríos-González, L. J., Castillo-Quiroz, D., & Reyes, A. G. (2021). Transcriptome-based metabolic profiling of flavonoids in Agave lechuguilla waste biomass. Plant Science, 305, 110748.

Nascimento, L. C., Yanagui, K., Jose, J., Camargo, E. L., Grassi, M. C. B., Cunha, C. P., ... & Carazzolle, M. F. (2019). Unraveling the complex genome of Saccharum spontaneum using Polyploid Gene Assembler. DNA Research, 26(3), 205-216.

Nei, M. Analysis Of Gene Diversity In Subdivided Populations. Proceedings Of The National Academy Of Sciences Of The United States Of America, V. 70, P. 3321–3323, 1973. Doi: 10.1073/Pnas.70.12.3321.

Newman, M. E. J. A Measure Of Betweenness Centrality Based On Random Walks. Arxiv, 2003. Available At: Https://Arxiv.Org/Abs/Cond-Mat/0309045. Accessed On: 24 Jul. 2024

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., ... & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic acids research, 42(D1), D26-D31.

Novogene. The Application Of Whole Genome Sequencing (Wgs) In Agricultural Breeding. Disponível Em: Https://Www.Novogene.Com/Us-En/Resources/Blog/The-Application-Of-Whole-Genome-Sequencing-Wgs-In-Agricultural-Breeding/. Acesso Em: 24 Jul. 2024.

Palomino, G.; Dolezel, J.; Rubluo, A. Nuclear Genome Size Analysis Of Agave Tequilana Weber. Caryologia, V. 56, N. 1, P. 37-46, 2003.

Papenbrock, Jutta; Schmidt, Ahlert. Characterization Of A Sulfurtransferase From Arabidopsis Thaliana. European Journal Of Biochemistry, V. 267, N. 1, P. 145-154, 2000.

Pereira, A. G. Et Al. Plant Alkaloids: Production, Extraction, And Potential Therapeutic Properties. In: Natural Secondary Metabolites: From Nature, Through Science, To Industry. Cham: Springer International Publishing, 2023. P. 157-200.

Perini, S., Johannesson, K., Butlin, R. K., & Westram, A. M. (2025). Short INDELs and SNPs as markers of evolutionary processes in hybrid zones. Journal of Evolutionary Biology, voaf002.

Peterson, B. K.; Weber, J. N.; Kay, E. H.; Fisher, H. S.; Hoekstra, H. E.Double Digest Radseq: An Inexpensive Method For De Novo Snp Discovery And Genotyping In Model And Non-Model Species. Plos One, V. 7, N. 5, P. E37135, 31 Maio 2012.

Pfalz, Marina Et Al. Methyl Transfer In Glucosinolate Biosynthesis Mediated By Indole Glucosinolate O-Methyltransferase 5. Plant Physiology, V. 172, N. 4, P. 2190-2203, 2016.

Pontes, Olga Et Al. Rna Polymerase V Functions In Arabidopsis Interphase Heterochromatin Organization Independently Of The 24-Nt Sirna-Directed Dna Methylation Pathway. Molecular Plant, V. 2, N. 4, P. 700-710, 2009.
Producing Agave Angustifolia Traditional Landraces Cultivated In Jalisco, Mexico. Plants, V. 11, N. 17, P. 2274, 2022.

Projeto Mapbiomas: Coleção 3.0 Da Série Anual De Mapas De Cobertura E Uso De Solo Do Brasil [Www Document]. Projeto Mapbiomas Coleção 3.0 Da Série Anual De Mapas De Cobertura E Uso De Solo Do Brasil, 2019.

Queiroga, Paula Et Al. Sisal (Agave Sisalana, Perrine): Tecnologias De Plantio E Utilização. 2021.

Queiroz, Sandra Regina De Oliveira Domingos Et Al. Análise Cromossômica Em Bulbilhos De Sisal (Agave Spp.) Cultivados Em Diferentes Municípios Baianos, Brasil. Acta Botanica Brasilica, V. 26, P. 842-848, 2012.

Rambaut, A. Figtree 2010. Institute Of Evolutionary Biology, University Of Edinburgh, Edinburgh. Available At: Http://Tree.Bio.Ed.Ac.Uk/Software/Figtree/. Accessed On: 24 Jul. 2024.

Raya, F. T. Et Al. New Feedstocks For Bioethanol Production: Energy Cane And Agave. In: Soccol, C. R.; Pereira, G. A. G.; Dussap, C.; Vandenberghe, L. P. S. (Eds.). Liquid Biofuels: Bioethanol. Germany: Springer, 2022. P. 431-455. Doi:10.1007/978-3-031-01241-9_18

Raya, Fabio Trigo Et Al. Extreme Physiology: Biomass And Transcriptional Profiling Of Three Abandoned Agave Cultivars. Industrial Crops And Products, V. 172, P. 114043, 2021.

Raya, Fábio Trigo Et Al. Molecular Epidemiology Of Sisal Bole Rot Disease Suggests A Potential Phytosanitary Crisis In Brazilian Production Areas. Frontiers In Chemical Engineering, V. 5, P. 1174689, 2023.

Raya, Fabio Trigo Et Al. Rescuing The Brazilian Agave Breeding Program: Morphophysiological And Molecular Characterization Of A New Germplasm. Frontiers In Chemical Engineering, V. 5, P. 1218668, 2023.

Rodriguez-Murillo, Laura; Salem, Rany M. Insertion/Deletion Polymorphism. In: Encyclopedia Of Behavioral Medicine. Cham: Springer International Publishing, 2020. P. 1192-1193.

Ruan, J., Zhou, Y., Zhou, M., Yan, J., Khurshid, M., Weng, W., ... & Zhang, K. (2019). Jasmonic acid signaling pathway in plants. International journal of molecular sciences, 20(10), 2479.

Sarwar, M. B. Et Al. De Novo Assembly Of Agave Sisalana Transcriptome In Response To Drought Stress Provides Insight Into The Tolerance Mechanisms. Scientific Reports, V. 9, P. 396, 2019. Doi:10.1038/S41598-018-35891-6.

Sauge‐merle, Sandrine; Falconet, Denis; Fontecave, Marc. An Active Ribonucleotide Reductase From Arabidopsis Thaliana: Cloning, Expression And Characterization Of The Large Subunit. European Journal Of Biochemistry, V. 266, N. 1, P. 62-69, 1999.

Sehn, J. K. Insertions And Deletions (Indels). In: Clinical Genomics. Academic Press, 2015. P. 129-150.

Shannon, Paul Et Al. Cytoscape: A Software Environment For Integrated Models Of Biomolecular Interaction Networks. Genome Research, V. 13, N. 11, P. 2498-2504, 2003.

Shu, F., Wang, D., Sarsaiya, S., Jin, L., Liu, K., Zhao, M., ... & Chen, J. (2024). Bulbil initiation: A comprehensive review on resources, development, and utilisation, with emphasis on molecular mechanisms, advanced technologies, and future prospects. Frontiers in Plant Science, 15, 1343222.

Skuse, G. R., & Du, C. (2008). Bioinformatics tools for plant genomics. International journal of plant genomics, 2008.

Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

Smith, S. M. Q&A: What Are Strigolactones And Why Are They Important To Plants And Soil Microbes?. Bmc Biology, V. 12, P. 19, 2014. Available At: Https://Doi.Org/10.1186/1741-7007-12-19. Accessed On: 20 Aug. 2024.

Sohn, J. I., & Nam, J. W. (2018). The present and future of de novo whole-genome assembly. Briefings in bioinformatics, 19(1), 23-40.

Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, doi: 10.1093/bioinformatics/btn013.

Stanke. M., Schöffmann, O., Morgenstern, B. and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62.

Szkiba, David Et Al. Snp2go: Functional Analysis Of Genome-Wide Association Studies. Genetics, V. 197, N. 1, P. 285-289, 2014.

Szklarczyk, D.; Kirsch, R.; Koutrouli, M. Et Al. The String Database In 2023: Protein–Protein Association Networks And Functional Enrichment Analyses For Any Sequenced Genome Of Interest. Nucleic Acids Research, V. 51, N. D1, P. D638–D646, 6 Jan. 2023.

Valenzuela, A. N. A. A New Agenda For Blue Agave Landraces: Food, Energy And Tequila. Gcb Bioenergy, V. 3, N. 1, P. 15-24, 2011.

Van Der Auwera, G. A.; O'connor, B. D. Genomics In The Cloud: Using Docker, Gatk, And Wdl In Terra. 1. Ed. Sebastopol: O'reilly Media, 2020.

Vargas-Ponce, O., Zizumbo-Villarreal, D., Martínez-Castillo, J., & Coello-Coello, J. Colunga-GarcíaMarín P. 2009. Diversity and structure of landraces of Agave grown for spirits under traditional agriculture: a comparison with wild populations of A. angustifolia (Agavaceae) and commercial plantations of A. tequilana. American Journal of Botany, 96, 448-457.

Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., ... & Wegrzyn, J. L. (2023). Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. Applications in Plant Sciences, 11(4), e11533.

Wang, Qing-Long Et Al. Metabolic Reprogramming In Chloroplasts Under Heat Stress In Plants. International Journal Of Molecular Sciences, V. 19, N. 3, P. 849, 2018.

Wei, L., Liu, Y., Dubchak, I., Shon, J., & Park, J. (2002). Comparative genomics approaches to study organism similarities and differences. Journal of biomedical informatics, 35(2), 142-150.

Weir, B.S. (1996) Genetic Data Analysis Ii: Methods For Discrete Population Genetic Data. Sinauer Associates, Inc., Sunderland.

Xie, L., Gong, X., Yang, K., Huang, Y., Zhang, S., Shen, L., ... & Fan, L. (2024). Technology-enabled great leap in deciphering plant genomes. Nature Plants, 10(4), 551-566.

Yang, Li, Et Al. "Novel Insights Into The Gene Regulatory Networks Of Cam And C4 Photosynthesis In Agave." Nature Communications, Vol. 12, No. 1, 2021, P. 1-13.

Yang, Xiaohan High Quality Genome Sequencing Of Agave Tequilana, A Bioenergy Crop With High Drought Tolerance And Low Biomass Recalcitranceagave Tequilana Var. Weber's Blue V2.1 Doe-Jgi, Https://Phytozome-Next.Jgi.Doe.Gov/Info/Atequilanavar_webersbluehap2_v2_1, 2024

Yang, Ziping Et Al. A Chromosome-Level Genome Assembly Of Agave Hybrid No. 11648 Provides Insights Into The Cam Photosynthesis. Horticulture Research, V. 11, N. 2, P.269, 2024.

Yin, H. Et Al. Diel Rewiring And Positive Selection Of Ancient Plant Proteins Enabled Evolution Of Cam Photosynthesis In Agave. Bmc Genomics, V. 19, P. 588–616, 2018. Doi:10.1186/S12864-018-4964-7.

Yu, Su-May. Cellular And Genetic Responses Of Plants To Sugar Starvation. Plant Physiology, V. 121, N. 3, P. 687-693, 1999.

Zhang, Y. M., Li, X., Chen, Z., Li, J. F., Lu, J. Y., & Zhou, W. Z. (2013). Shoot organogenesis and plant regeneration in Agave hybrid, No. 11648. Scientia horticulturae, 161, 30-34.

Zheng, Xiuwen Et Al. A High-Performance Computing Toolset For Relatedness And Principal Component Analysis Of Snp Data. Bioinformatics, V. 28, N. 24, P. 3326-3328, 2012.

7. SUPPLEMENTARY MATERIAL

Table 1 - Table of the GOs of the missense SNPs H400L-exclusive genes (in comparison to H11648)

GO ID	Term	Adjusted p-value	Genes
GO:0002128	tRNA nucleoside ribose methylation	0.006	AgveH2v21094690m
GO:0009934	regulation of meristem structural organi...	0.006	AgveH2v21023674m
GO:0034198	cellular response to amino acid starvati...	0.006	AgveH2v21072733m
GO:0071242	cellular response to ammonium ion	0.006	AgveH2v21015618m
GO:0016560	protein import into peroxisome matrix, d...	0.015	AgveH2v21058081m
GO:0055129	L-proline biosynthetic process	0.015	AgveH2v21094562m
GO:0072318	clathrin coat disassembly	0.015	AgveH2v21047250m
GO:0009910	negative regulation of flower developmen...	0.018	AgveH2v21038516m
GO:0006751	glutathione catabolic process	0.024	AgveH2v21015483m
GO:0080092	regulation of pollen tube growth	0.024	AgveH2v21096636m
GO:0009263	deoxyribonucleotide biosynthetic process	0.027	AgveH2v21021287m
GO:0006096	glycolytic process	0.029	AgveH2v21057609m, AgveH2v21063169m
GO:0010030	positive regulation of seed germination	0.029	AgveH2v21004243m

GO:0045037	protein import into chloroplast stroma	0.035	AgveH2v21094013m
GO:0006189	'de novo' IMP biosynthetic process	0.038	AgveH2v21053936m
GO:0048026	positive regulation of mRNA splicing, vi...	0.038	AgveH2v21086812m
GO:0071577	zinc ion transmembrane transport	0.041	AgveH2v21034090m
GO:0018216	peptidyl-arginine methylation	0.050	AgveH2v21000346m

Table 2 - Table of the GOs of the missense SNPs H11648-exclusive genes (in comparison to H400L)

GO ID	**Term**	**p-value**	**Genes**
GO:0009825	multidimensional cell growth	0.0035	AgveH2v21065424m
GO:0009963	positive regulation of flavonoid biosynt...	0.0035	AgveH2v21015253m
GO:0010106	cellular response to iron ion starvation	0.0035	AgveH2v21091624m
GO:0006396	RNA processing	0.0047	AgveH2v21021276m,AgveH2v21033660m,AgveH2v21057844m,AgveH2v21096862m
GO:0072344	rescue of stalled ribosome	0.0052	AgveH2v21016159m
GO:0000467	exonucleolytic trimming to generate matu...	0.0070	AgveH2v21057844m
GO:0019478	D-amino acid catabolic process	0.0070	AgveH2v21110728m
GO:0048254	snoRNA localization	0.0070	AgveH2v21069278m

GO:0006434	seryl-tRNA aminoacylation	0.0087	AgveH2v21112978m
GO:0016560	protein import into peroxisome matrix, d...	0.0087	AgveH2v21106202m
GO:0098609	cell-cell adhesion	0.0087	AgveH2v21114207m
GO:0042256	cytosolic ribosome assembly	0.0105	AgveH2v21023970m
GO:1902347	response to strigolactone	0.0105	AgveH2v21102479m
GO:0035308	negative regulation of protein dephospho...	0.0139	AgveH2v21054921m
GO:0010167	response to nitrate	0.0225	AgveH2v21113080m
GO:0009083	branched-chain amino acid catabolic proc...	0.0242	AgveH2v21076014m
GO:0080162	endoplasmic reticulum to cytosol auxin t...	0.0242	AgveH2v21041956m
GO:0009860	pollen tube growth	0.0327	AgveH2v21072574m
GO:0043631	RNA polyadenylation	0.0445	AgveH2v21015128m

Table 3 - Table of the GOs of the missense SNPs *A. sisalana*-exclusive genes in comparison hybrids (H400L and H11648)

GO ID	**Term**	**Adjusted p-value**	**Genes**
GO:0006468	protein phosphorylation	0.00043	AgveH2v21000801m,AgveH2v21001614m,AgveH2v21006711m,AgveH2v21012555m,AgveH2v21012705m,AgveH2v21017534m,AgveH2v21017860m,AgveH2v21017948m,AgveH2v21018057m,AgveH2v21020298m,AgveH2v21022495m,AgveH2v21024941m,AgveH2v21025936m,AgveH2v21029840m,AgveH2v21034318m,AgveH2v21036914m,AgveH2v21038855m,AgveH2v

			21042129m,AgveH2v21042478m,AgveH2v21046921m,AgveH2v21047158m,AgveH2v21048780m,AgveH2v21050272m,AgveH2v21050500m,AgveH2v21054258m,AgveH2v21055099m,AgveH2v21058984m,AgveH2v21059453m,AgveH2v21062710m,AgveH2v21062992m,AgveH2v21063780m,AgveH2v21064852m,AgveH2v21065362m,AgveH2v21068336m,AgveH2v21069693m,AgveH2v21075331m,AgveH2v21079070m,AgveH2v21080398m,AgveH2v21081228m,AgveH2v21082081m,AgveH2v21082738m,AgveH2v21088028m,AgveH2v21088032m,AgveH2v21088891m,AgveH2v21093967m,AgveH2v21094310m,AgveH2v21095385m,AgveH2v21098199m,AgveH2v21100735m,AgveH2v21101425m,AgveH2v21102372m,AgveH2v21105038m,AgveH2v21108512m,AgveH2v21108891m,AgveH2v21115037m
GO:0052863	1-deoxy-D-xylulose 5-phosphate metabolic...	0.00299	AgveH2v21053202m,AgveH2v21080774m
GO:0060178	regulation of exocyst localization	0.00299	AgveH2v21021228m,AgveH2v21076285m
GO:0045128	negative regulation of reciprocal meioti...	0.00490	AgveH2v21033502m,AgveH2v21041523m
GO:0098656	monoatomic anion transmembrane transport	0.00520	AgveH2v21004952m,AgveH2v21042663m,AgveH2v21089702m,AgveH2v21103372m
GO:0042138	meiotic DNA double-strand break formatio...	0.00725	AgveH2v21055780m,AgveH2v21056420m
GO:0048207	vesicle targeting, rough ER to cis-Golgi	0.00725	AgveH2v21070317m,AgveH2v21088606m

GO:0030036	actin cytoskeleton organization	0.01001	AgveH2v21000628m,AgveH2v21002469m,AgveH2v21003451m,AgveH2v21006607m
GO:0080092	regulation of pollen tube growth	0.01313	AgveH2v21079987m,AgveH2v21098971m
GO:0006357	regulation of transcription by RNA polym...	0.01490	AgveH2v21002924m,AgveH2v21007501m,AgveH2v21030192m,AgveH2v21043140m,AgveH2v21045369m,AgveH2v21049120m,AgveH2v21051329m,AgveH2v21062587m,AgveH2v21064995m,AgveH2v21071197m,AgveH2v21090796m,AgveH2v21097387m,AgveH2v21103712m,AgveH2v21107207m,AgveH2v21107643m,AgveH2v21113537m
GO:0051091	positive regulation of DNA-binding trans...	0.02047	AgveH2v21017436m,AgveH2v21017441m
GO:0000076	DNA replication checkpoint signaling	0.02264	AgveH2v21069062m,AgveH2v21087189m
GO:0007019	microtubule depolymerization	0.02267	AgveH2v21065272m
GO:0009556	microsporogenesis	0.02267	AgveH2v21096864m
GO:0009759	indole glucosinolate biosynthetic proces...	0.02267	AgveH2v21085854m
GO:0060147	regulation of post-transcriptional gene ...	0.02267	AgveH2v21106377m
GO:0090173	regulation of synaptonemal complex assem...	0.02267	AgveH2v21035376m
GO:0010597	green leaf volatile biosynthetic process	0.02379	AgveH2v21017790m,AgveH2v21039732m,AgveH2v21058508m,AgveH2v21083814m,AgveH2v21115188m
GO:0017148	negative regulation of translation	0.02910	AgveH2v21020855m,AgveH2v21069529m,AgveH2v21071708m
GO:0010020	chloroplast fission	0.03900	AgveH2v21028675m,AgveH2v21062559m

GO:0006334	nucleosome assembly	0.04247	AgveH2v21053589m,AgveH2v21081208m,AgveH2v21098344m,AgveH2v21114550m
GO:0007623	circadian rhythm	0.04428	AgveH2v21019937m,AgveH2v21098561m,AgveH2v21109892m
GO:0034220	monoatomic ion transmembrane transport	0.04467	AgveH2v21003375m,AgveH2v21004952m,AgveH2v21013836m,AgveH2v21018037m,AgveH2v21027257m,AgveH2v21042663m,AgveH2v21060162m,AgveH2v21062530m,AgveH2v21080033m,AgveH2v21089702m,AgveH2v21100918m,AgveH2v21103372m
GO:0045910	negative regulation of DNA recombination	0.04470	AgveH2v21033502m,AgveH2v21041523m,AgveH2v21084084m
GO:0000493	box H/ACA snoRNP assembly	0.04483	AgveH2v21102813m
GO:0007004	telomere maintenance via telomerase	0.04483	AgveH2v21102241m
GO:0009934	regulation of meristem structural organi...	0.04483	AgveH2v21044932m
GO:0019285	glycine betaine biosynthetic process fro...	0.04483	AgveH2v21013897m
GO:0019805	quinolinate biosynthetic process	0.04483	AgveH2v21113089m
GO:0030951	establishment or maintenance of microtub...	0.04483	AgveH2v21006993m
GO:0045053	protein retention in Golgi apparatus	0.04483	AgveH2v21112966m
GO:0071497	cellular response to freezing	0.04483	AgveH2v21109903m
GO:0080178	5-carbamoylmethyl uridine residue modifi...	0.04483	AgveH2v21026940m
GO:1901652	response to peptide	0.04483	AgveH2v21066557m

GO:1901996	regulation of indoleacetic acid biosynth...	0.04483	AgveH2v21045800m
GO:2000014	regulation of endosperm development	0.04483	AgveH2v21058684m

Table 6 - Table of the GOs of the high impact indels *A. sisalana*-exclusive genes in comparison hybrids (H400L and H11648)

GO ID	Term	Adjusted p-value	Genes
GO:0006784	heme A biosynthetic process	0.00064	AgveH2v21038757m, AgveH2v21097170m
GO:0006680	glucosylceramide catabolic process	0.00210	AgveH2v21027037m, AgveH2v21087846m
GO:0034337	RNA folding	0.00210	AgveH2v21008177m, AgveH2v21073066m
GO:0045128	negative regulation of reciprocal meioti...	0.00210	AgveH2v21033502m, AgveH2v21041523m
GO:0006488	dolichol-linked oligosaccharide biosynth...	0.01295	AgveH2v21016812m, AgveH2v21057234m
GO:0006338	chromatin remodeling	0.01442	AgveH2v21024995m, AgveH2v21042015m, AgveH2v21046273m, AgveH2v21051742m, AgveH2v21098344m, AgveH2v21103622m, AgveH2v21107963m, AgveH2v21114550m
GO:0048441	petal development	0.01473	AgveH2v21027760m
GO:0071493	cellular response to UV-B	0.01473	AgveH2v21005053m
GO:0006075	(1->3)-beta-D-glucan biosynthetic proces...	0.01751	AgveH2v21053015m, AgveH2v21107552m

GO:0008610	lipid biosynthetic process	0.01967	AgveH2v21004967m, AgveH2v21009328m, AgveH2v21011081m, AgveH2v21029836m, AgveH2v21032082m, AgveH2v21059617m, AgveH2v21072974m, AgveH2v21073726m, AgveH2v21081874m, AgveH2v21083296m, AgveH2v21092373m, AgveH2v21096948m, AgveH2v21098231m, AgveH2v21110088m
GO:0002679	respiratory burst involved in defense re...	0.02924	AgveH2v21076647m
GO:0009934	regulation of meristem structural organi...	0.02924	AgveH2v21044932m
GO:0009963	positive regulation of flavonoid biosynt...	0.02924	AgveH2v21015253m
GO:0019285	glycine betaine biosynthetic process fro...	0.02924	AgveH2v21013897m
GO:0071242	cellular response to ammonium ion	0.02924	AgveH2v21015618m
GO:1903415	flavonoid transport from endoplasmic ret...	0.02924	AgveH2v21049246m
GO:1904216	positive regulation of protein import in...	0.02924	AgveH2v21081713m
GO:2000641	regulation of early endosome to late end...	0.02924	AgveH2v21031311m
GO:0009836	fruit ripening, climacteric	0.03018	AgveH2v21001896m, AgveH2v21016219m, AgveH2v21044564m

GO:0006402	mRNA catabolic process	0.03114	AgveH2v21043178m, AgveH2v21052028m, AgveH2v21064366m, AgveH2v21069529m, AgveH2v21077369m
GO:0034204	lipid translocation	0.03451	AgveH2v21063124m, AgveH2v21098551m
GO:0030433	ubiquitin-dependent ERAD pathway	0.03778	AgveH2v21001086m, AgveH2v21025776m
GO:0006506	GPI anchor biosynthetic process	0.03926	AgveH2v21004967m, AgveH2v21059617m, AgveH2v21092373m
GO:0042752	regulation of circadian rhythm	0.04116	AgveH2v21098561m, AgveH2v21114716m
GO:0009904	chloroplast accumulation movement	0.04353	AgveH2v21089260m
GO:0010193	response to ozone	0.04353	AgveH2v21104274m
GO:0010452	histone H3-K36 methylation	0.04353	AgveH2v21069041m
GO:0070143	mitochondrial alanyl-tRNA aminoacylation	0.04353	AgveH2v21044839m
GO:0006108	malate metabolic process	0.04465	AgveH2v21061481m, AgveH2v21072425m

Tabela 7 - Table of the GOs of the missense SNPs Hybrids-exclusive genes in comparison to *A. sisalana*.

GO ID	**Term**	**Adjusted p-value**	**Genes**
GO:0045037	protein import into chloroplast stroma	0.00061	AgveH2v21020225m, AgveH2v21094013m
GO:0009825	multidimensional cell growth	0.00619	AgveH2v21065424m

GO:0034198	cellular response to amino acid starvati...	0.00619	AgveH2v21072733m
GO:0006434	seryl-tRNA aminoacylation	0.01540	AgveH2v21112978m
GO:0016560	protein import into peroxisome matrix, d...	0.01540	AgveH2v21106202m
GO:0055129	L-proline biosynthetic process	0.01540	AgveH2v21094562m
GO:0072318	clathrin coat disassembly	0.01540	AgveH2v21047250m
GO:1902347	response to strigolactone	0.01845	AgveH2v21102479m
GO:0046168	glycerol-3-phosphate catabolic process	0.02149	AgveH2v21093744m
GO:0009873	ethylene-activated signaling pathway	0.02219	AgveH2v21069324m,PAVagav21.02509m
GO:0006751	glutathione catabolic process	0.02452	AgveH2v21015483m
GO:0080092	regulation of pollen tube growth	0.02452	AgveH2v21096636m
GO:0009263	deoxyribonucleotide biosynthetic process	0.02755	AgveH2v21021287m
GO:0010030	positive regulation of seed germination	0.03056	AgveH2v21004243m
GO:0006096	glycolytic process	0.03058	AgveH2v21057609m,AgveH2v21063169m
GO:0006417	regulation of translation	0.03318	AgveH2v21064713m,AgveH2v21094999m
GO:0006189	'de novo' IMP biosynthetic process	0.03955	AgveH2v21053936m
GO:0010167	response to nitrate	0.03955	AgveH2v21113080m
GO:0048026	positive regulation of mRNA splicing, vi...	0.03955	AgveH2v21086812m

GO:0034220	monoatomic ion transmembrane transport	0.04136	AgveH2v21034090m,AgveH2v21082528m,AgveH2v21100928m
GO:0009083	branched-chain amino acid catabolic proc...	0.04253	AgveH2v21076014m
GO:0071577	zinc ion transmembrane transport	0.04253	AgveH2v21034090m

Table 8 - GOs of H11648 Exclusive Orthogroups

GO.ID	Term	Adjusted p-value	Genes
GO:0042908	xenobiotic transport	0.00053	EVM0010457.1,EVM0011037.1, EVM0033425.1,EVM0034335.1
GO:0006412	translation	0.00130	EVM0000075.1,EVM0000487.1, EVM0000849.1,EVM0005115.1, EVM0007170.1,EVM0008001.1, EVM0011241.1,EVM0011715.1, EVM0012857.1,EVM0013120.1, EVM0018583.1,EVM0023110.1, EVM0031851.1,EVM0040367.1, EVM0040492.1,EVM0049250.1
GO:0000160	phosphorelay signal transduction system	0.00157	EVM0000531.1,EVM0005665.1, EVM0008072.1,EVM0015752.1
GO:0048278	vesicle docking	0.00404	EVM0014514.1,EVM0028402.1, EVM0033736.2
GO:0015628	protein secretion by the type II secreti...	0.00639	EVM0008691.1
GO:0019404	galactitol catabolic process	0.00639	EVM0011430.1
GO:0019592	mannitol catabolic process	0.00639	EVM0028530.1
GO:0031460	glycine betaine transport	0.00639	EVM0021696.1
GO:0019722	calcium-mediated signaling	0.01473	EVM0011108.1,EVM0027604.1

GO:0009409	response to cold	0.01678	EVM0028599.1,EVM0031519.1,EVM0051933.1
GO:0006888	endoplasmic reticulum to Golgi vesicle-m...	0.01786	EVM0001605.1,EVM0003713.1,EVM0023871.1
GO:0009297	pilus assembly	0.01904	EVM0032202.1
GO:0015853	adenine transport	0.01904	EVM0012290.1
GO:1903601	thermospermine metabolic process	0.01904	EVM0025901.1
GO:0046813	receptor-mediated virion attachment to h...	0.02531	EVM0000950.1
GO:0002143	tRNA wobble position uridine thiolation	0.03154	EVM0006091.1
GO:0009401	phosphoenolpyruvate-dependent sugar phos...	0.03154	EVM0050745.1
GO:0043153	entrainment of circadian clock by photop...	0.03154	EVM0046192.1
GO:0000045	autophagosome assembly	0.03360	EVM0012988.1,EVM0021128.1
GO:0046854	phosphatidylinositol phosphate biosynthe...	0.03360	EVM0004248.1,EVM0015760.1
GO:0019632	shikimate metabolic process	0.03773	EVM0009301.1
GO:0036444	calcium import into the mitochondrion	0.03773	EVM0023155.1
GO:2000280	regulation of root development	0.03773	EVM0007985.1
GO:0016102	diterpenoid biosynthetic process	0.04073	EVM0008714.1,EVM0026610.1
GO:0046274	lignin catabolic process	0.04373	EVM0030678.1,EVM0031264.1
GO:0010124	phenylacetate catabolic process	0.04388	EVM0017966.1
GO:0009698	phenylpropanoid metabolic process	0.04748	EVM0011717.1,EVM0030678.1,EVM0031264.1,EVM0036905.1,EVM0040968.1
GO:0006486	protein glycosylation	0.04976	EVM0001330.1,EVM0002450.1,EVM0006325.1,EVM0024543.1

GO:0000719	photoreactive repair	0.04998	EVM0017445.1
GO:0006423	cysteinyl-tRNA aminoacylation	0.04998	EVM0000487.1
GO:0090333	regulation of stomatal closure	0.04998	EVM0000115.1

Table 9 - GOs of *A. sisalana* Exclusive Orthogroups

GO.ID	Term	Adjusted p-value	Genes
GO:0043171	peptide catabolic process	0.0011	g54968.t1,g56338.t1,g71450.t1
GO:0045332	phospholipid translocation	0.0012	g106213.t1,g117817.t1,g1649.t1,g38587.t1
GO:0009725	response to hormone	0.0023	g108343.t1,g108895.t1,g117089.t1,g120367.t1,g121569.t1,g123269.t1,g123455.t1,g127350.t1,g128091.t1,g143324.t1,g143843.t1,g145315.t2,g15588.t1,g17988.t1,g28797.t1,g48216.t1,g60833.t1,g65561.t1,g7346.t1,g78861.t1,g92173.t1
GO:0009090	homoserine biosynthetic process	0.0035	g119182.t1,g136534.t1,g147205.t1
GO:0006108	malate metabolic process	0.0047	g133773.t1,g143045.t1,g19464.t1,g34762.t1,g60852.t1
GO:0016042	lipid catabolic process	0.0061	g100930.t1,g101124.t1,g119352.t1,g126729.t1,g126811.t1,g135510.t1,g145073.t1,g147132.t2,g18233.t1,g20329.t1,g55091.t1,g60952.t1,g61962.t1
GO:0009749	response to glucose	0.0061	g123765.t1,g129216.t1
GO:0051014	actin filament severing	0.0061	g128943.t1,g135772.t2,g14719.t1
GO:0034605	cellular response to heat	0.0085	g108100.t1,g12278.t1,g146697.t1,g32710.t1,g36703.t1
GO:0006486	protein glycosylation	0.0087	g10020.t1,g10628.t1,g111300.t1,g115085.t1,g119057.t2,g123749.t1,g131432.t1,g132265.t1,g133982.t1,g135600.t1,g138434.t1,g14628.t1,g33585.t1,g34302.t1,g43852.t1
GO:0006633	fatty acid biosynthetic process	0.0094	g102088.t1,g105505.t1,g106312.t1,g118019.t2,g123802.t1,g126096.t1,g129663.t1,g137914.t1,g140349.t1,g143244.t1,g5814.t1,g87261.t1,g89854.t1
GO:0046856	phosphatidylinositol	0.0096	g128274.t1,g131651.t1,g3293.t1,g43199.t1

	dephosphorylation		
GO:0051168	nuclear export	0.0099	g122446.t1,g123948.t2,g126795.t1,g129550.t1,g133416.t1,g145412.t1,g27754.t1
GO:0031222	arabinan catabolic process	0.0100	g106052.t1,g132602.t1
GO:0016126	sterol biosynthetic process	0.0109	g108373.t1,g120016.t1,g14622.t1,g34615.t2
GO:0000165	MAPK cascade	0.0127	g105828.t1,g12346.t1,g126716.t1,g135652.t1,g140191.t1
GO:0000289	nuclear-transcribed mRNA poly(A) tail sh...	0.0142	g122226.t1,g26715.t1,g81059.t1
GO:0000712	resolution of meiotic recombination inte...	0.0147	g103579.t1,g2328.t1
GO:0006287	base-excision repair, gap-filling	0.0147	g102517.t1,g8746.t1
GO:0030388	fructose 1,6-bisphosphate metabolic proc...	0.0168	g116224.t2,g14655.t1,g21281.t1
GO:0035556	intracellular signal transduction	0.0179	g101620.t1,g103481.t1,g1037.t1,g105828.t1,g106054.t1,g109934.t1,g109996.t1,g111620.t1,g112705.t1,g116762.t1,g121849.t1,g122620.t1,g12346.t1,g126716.t1,g128091.t1,g130591.t1,g133537.t1,g134205.t1,g135652.t1,g136754.t1,g137570.t1,g137743.t1,g138452.t1,g140191.t1,g145038.t1,g147953.t1,g15588.t1,g23375.t1,g3632.t1,g46653.t1
GO:0002188	translation reinitiation	0.0201	g13084.t1,g140805.t1
GO:0009228	thiamine biosynthetic process	0.0201	g125883.t1,g55726.t1
GO:0010265	SCF complex assembly	0.0201	g101098.t1,g101945.t1
GO:0006367	transcription initiation at RNA polymera...	0.0262	g114041.t1,g140671.t1,g141700.t1,g145036.t1
GO:0031647	regulation of protein stability	0.0262	g106292.t1,g133094.t1,g138450.t1
GO:0016120	carotene biosynthetic process	0.0262	g121693.t1,g139192.t1
GO:0034440	lipid oxidation	0.0266	g126251.t1,g18701.t1,g35449.t1,g58085.t1,g61962.t1

GO:0000045	autophagosome assembly	0.0299	g10101.t1,g125809.t1,g22128.t1
GO:0051453	regulation of intracellular pH	0.0318	g110491.t1,g132039.t1,g142460.t1,g16194.t1,g16198.t1
GO:0006897	endocytosis	0.0321	g104154.t1,g112577.t1,g122915.t1,g133436.t1,g32901.t1,g90059.t1
GO:0000388	spliceosome conformational change to rel...	0.0327	g34951.t1
GO:0010636	positive regulation of mitochondrial fus...	0.0327	g123611.t1
GO:0032392	DNA geometric change	0.0327	g111790.t1,g98679.t1
GO:0006782	protoporphyrinogen IX biosynthetic proce...	0.0330	g111017.t1,g125628.t1
GO:0045836	positive regulation of meiotic nuclear d...	0.0330	g11022.t1,g128103.t1
GO:0009742	brassinosteroid mediated signaling pathw...	0.0358	g108343.t1,g120367.t1,g123455.t1,g143843.t1,g65561.t1
GO:0043328	protein transport to vacuole involved in...	0.0379	g119462.t1,g140210.t1,g59756.t1
GO:0042853	L-alanine catabolic process	0.0404	g53015.t1,g63072.t1
GO:0006085	acetyl-CoA biosynthetic process	0.0420	g118301.t1,g118555.t1,g122916.t1,g150480.t1,g50995.t1,g89103.t1
GO:0006265	DNA topological change	0.0423	g112022.t1,g148326.t1,g2029.t1
GO:0009768	photosynthesis, light harvesting in phot...	0.0483	g17522.t1,g23716.t1

Table 10 - GOs of *A. tequilana* Exclusive Orthogroups

GO.ID	Term	Adjusted p-value	Genes
GO:0019684	photosynthesis, light reaction	0.0010	AgveH2v21022307m,AgveH2v21024085m, AgveH2v21024446m,AgveH2v21099440m

GO:0010038	response to metal ion	0.0025	AgveH2v21023001m,AgveH2v21077036m
GO:0015886	heme transport	0.0083	AgveH2v21018952m
GO:0043457	regulation of cellular respiration	0.0083	AgveH2v21040515m
GO:0045128	negative regulation of reciprocal meioti...	0.0083	AgveH2v21113749m
GO:0006878	intracellular copper ion homeostasis	0.0166	AgveH2v21053991m
GO:0051091	positive regulation of DNA-binding trans...	0.0166	AgveH2v21005350m
GO:0022904	respiratory electron transport chain	0.0228	AgveH2v21005639m,AgveH2v21021409m
GO:0001172	RNA-templated transcription	0.0232	AgveH2v21082010m
GO:0009788	negative regulation of abscisic acid-act...	0.0297	AgveH2v21075215m
GO:0022900	electron transport chain	0.0445	AgveH2v21005639m,AgveH2v21021409m, AgveH2v21035048m,AgveH2v21035049m

Table 11 - GOs of H11648 Expanded Families

GO.ID	Term	Adjusted p-value	Genes
GO:0006357	regulation of transcription by RNA polym...	0.00087	EVM0000287.1,EVM0001233.1,EVM0001282.1,EVM0001748.1,EVM0002165.1,EVM0002448.1,EVM0002500.1,EVM0003090.1,EVM0003838.1,EVM0003909.1,EVM0004598.1,EVM0005799.1,EVM0007712.1,EVM0008768.1,EVM0009331.1,EVM0010414.1,EVM0012094.1,EVM0017239.1,EVM0017654.1,EVM0018708.1,EVM0019615.1,EVM0021441.1,EVM0022055.1,EVM0029719.1,EVM0029885.1,EVM0029982.1,EVM0032442.1,EVM0032685.1
GO:0006355	regulation of DNA-templated transcriptio...	0.00258	EVM0000046.1,EVM0000049.1,EVM0000055.1,EVM0000116.1,EVM0000223.1,EVM0000287.1,EVM0000680.1,EVM0000809.1,EVM0001162.1,EVM0001233.1,EVM0001282.1,EVM0001687.1,EVM0001748.1,EVM0001929.1,EVM0002165.1,EVM0002253.1,EVM0002448.1,EVM0002500.1,EVM0002737.1,EVM0002934.1,EVM0002949.1,EVM0003090.1,EVM

			0003300.1,EVM0003317.1,EVM0003520.1,EVM0003637.1,EVM0003641.1,EVM0003646.1,EVM0003838.1,EVM0003878.1,EVM0003909.1,EVM0003918.2,EVM0004212.1,EVM0004538.1,EVM0004598.1,EVM0004676.1,EVM0005345.1,EVM0005799.1,EVM0006401.1,EVM0006495.1,EVM0006863.1,EVM0007320.1,EVM0007712.1,EVM0008169.1,EVM0008536.1,EVM0008768.1,EVM0009331.1,EVM0009344.1,EVM0009438.1,EVM0009928.1,EVM0010170.1,EVM0010179.1,EVM0010246.1,EVM0010414.1,EVM0010978.1,EVM0011026.1,EVM0011271.1,EVM0011549.1,EVM0011712.3,EVM0012094.1,EVM0012308.1,EVM0012564.1,EVM0013923.1,EVM0014234.1,EVM0014373.1,EVM0015087.1,EVM0016333.1,EVM0016586.1,EVM0017239.1,EVM0017490.1,EVM0017654.1,EVM0017843.1,EVM0018708.1,EVM0018748.1,EVM0019454.1,EVM0019615.1,EVM0020301.1,EVM0021441.1,EVM0022055.1,EVM0023338.1,EVM0024538.1,EVM0025993.1,EVM0028288.1,EVM0029719.1,EVM0029885.1,EVM0029982.1,EVM0030610.1,EVM0032442.1,EVM0032685.1,EVM0032766.1
GO:0045893	positive regulation of DNA-templated tra...	0.00293	EVM0000046.1,EVM0001162.1,EVM0003300.1,EVM0004538.1,EVM0011712.3,EVM0013923.1,EVM0015087.1,EVM0017654.1,EVM0018708.1,EVM0021441.1,EVM0023338.1,EVM0028288.1,EVM0029982.1,EVM0032685.1
GO:0045487	gibberellin catabolic process	0.01155	EVM0002266.1,EVM0002678.1
GO:0048015	phosphatidylinositol-mediated signaling	0.01155	EVM0005389.1,EVM0045844.1
GO:0016558	protein import into peroxisome matrix	0.01214	EVM0008774.1,EVM0009034.1,EVM0018331.1,EVM0018505.1
GO:2000652	regulation of secondary cell wall biogen...	0.01515	EVM0004884.1,EVM0013118.1
GO:0045116	protein neddylation	0.01791	EVM0004154.1,EVM0007086.1,EVM0010214.1
GO:0031146	SCF-dependent proteasomal ubiquitin-depe...	0.01892	EVM0001532.1,EVM0003131.1,EVM0007678.1,EVM0012105.1,EVM0012254.1,EVM0038627.1
GO:0035493	SNARE complex assembly	0.01917	EVM0016876.1,EVM0025165.1

GO:2001295	malonyl-CoA biosynthetic process	0.01917	EVM0008723.1,EVM0030205.1
GO:0010468	regulation of gene expression	0.02035	EVM0000046.1,EVM0000049.1,EVM0000055.1,EVM0000116.1,EVM0000223.1,EVM0000287.1,EVM0000465.1,EVM0000680.1,EVM0000809.1,EVM0001162.1,EVM0001233.1,EVM0001282.1,EVM0001687.1,EVM0001748.1,EVM0001776.1,EVM0001929.1,EVM0002165.1,EVM0002253.1,EVM0002448.1,EVM0002500.1,EVM0002737.1,EVM0002934.1,EVM0002949.1,EVM0003090.1,EVM0003300.1,EVM0003317.1,EVM0003520.1,EVM0003637.1,EVM0003641.1,EVM0003646.1,EVM0003838.1,EVM0003878.1,EVM0003909.1,EVM0003918.2,EVM0004212.1,EVM0004356.1,EVM0004538.1,EVM0004598.1,EVM0004676.1,EVM0004873.1,EVM0005345.1,EVM0005418.1,EVM0005799.1,EVM0006401.1,EVM0006495.1,EVM0006503.1,EVM0006863.1,EVM0007320.1,EVM0007712.1,EVM0008169.1,EVM0008536.1,EVM0008768.1,EVM0009331.1,EVM0009344.1,EVM0009438.1,EVM0009928.1,EVM0010170.1,EVM0010179.1,EVM0010246.1,EVM0010414.1,EVM0010978.1,EVM0011026.1,EVM0011271.1,EVM0011432.1,EVM0011549.1,EVM0011712.3,EVM0012094.1,EVM0012308.1,EVM0012564.1,EVM0013159.1,EVM0013163.1,EVM0013713.1,EVM0013923.1,EVM0014234.1,EVM0014373.1,EVM0015087.1,EVM0015187.1,EVM0016235.1,EVM0016333.1,EVM0016586.1,EVM0016825.1,EVM0016953.1,EVM0017192.1,EVM0017239.1,EVM0017464.1,EVM0017490.1,EVM0017654.1,EVM0017843.1,EVM0018708.1,EVM0018748.1,EVM0018881.1,EVM0019454.1,EVM0019615.1,EVM0020301.1,EVM0020708.1,EVM0021441.1,EVM0022055.1,EVM0023338.1,EVM0024538.1,EVM0025399.1,EVM0025993.1,EVM0028288.1,EVM0029719.1,EVM0029885.1,EVM0029982.1,EVM0030610.1,EVM0032442.1,EVM0032685.1,EVM0032766.1,EVM0053340.1
GO:0051028	mRNA transport	0.02243	EVM0001754.1,EVM0003377.1,EVM0025352.1,EVM0032175.1
GO:0034314	Arp2/3 complex-mediated actin nucleation	0.02243	EVM0000252.1,EVM0005729.1,EVM0034844.1
GO:0010555	response to mannitol	0.02444	EVM0004594.1
GO:0030155	regulation of cell adhesion	0.02444	EVM0020076.1
GO:0006629	lipid metabolic process	0.02806	EVM0000086.1,EVM0000615.1,EVM0000911.1,EVM0001127.1,EVM0001579.1,EVM0001674.1,EVM0002027.1,EVM0002148.1,EVM0002266.1,EVM000

			2418.1,EVM0002678.1,EVM0002721.1,EVM00029 16.1,EVM0003169.1,EVM0003222.1,EVM0003302. 1,EVM0003318.1,EVM0003928.1,EVM0004724.1,E VM0005095.1,EVM0005450.1,EVM0006647.1,EVM 0007310.1,EVM0007988.1,EVM0009299.1,EVM00 09770.1,EVM0011062.1,EVM0011651.1,EVM0012 912.1,EVM0013585.1,EVM0013711.1,EVM001550 1.1,EVM0016347.1,EVM0016973.1,EVM0020535.1 ,EVM0021080.1,EVM0023663.1,EVM0024005.1,E VM0024203.1,EVM0024553.1,EVM0024798.1,EVM 0025951.1,EVM0027508.1,EVM0027724.1,EVM00 36803.1,EVM0037674.4,EVM0043125.1
GO:0000395	mRNA 5'-splice site recognition	0.02836	EVM0002223.1,EVM0003374.1
GO:0033617	mitochondrial cytochrome c oxidase assem...	0.02836	EVM0003425.1,EVM0004323.1
GO:0090158	endoplasmic reticulum membrane organizat...	0.03028	EVM0001610.1,EVM0012052.1,EVM0042086.1
GO:0006656	phosphatidylcholine biosynthetic process	0.03338	EVM0002418.1,EVM0003318.1,EVM0004724.1,EV M0016973.1
GO:0007155	cell adhesion	0.03342	EVM0001557.1,EVM0009360.1,EVM0020076.1
GO:0010048	vernalization response	0.03894	EVM0008517.1,EVM0014877.1
GO:0000373	Group II intron splicing	0.04270	EVM0001274.1,EVM0002762.1,EVM0014399.2
GO:0006338	chromatin remodeling	0.04375	EVM0000290.1,EVM0000545.1,EVM0002104.1,EV M0002353.1,EVM0003128.1,EVM0003311.1,EVM0 003487.1,EVM0004103.1,EVM0004988.1,EVM000 6309.1,EVM0006949.1,EVM0009628.1,EVM00096 67.1,EVM0011713.1,EVM0012195.1,EVM0012883. 1,EVM0013123.1,EVM0013163.1,EVM0014060.1,E VM0014570.1,EVM0016092.1,EVM0017464.1,EVM 0018558.2,EVM0022781.1,EVM0023286.1,EVM00 24888.1,EVM0025399.1,EVM0025890.1,EVM0029 074.1
GO:0010077	maintenance of inflorescence meristem id...	0.04829	EVM0005114.1
GO:0048439	flower morphogenesis	0.04829	EVM0005486.1
GO:0055047	generative cell mitosis	0.04829	EVM0007825.1
GO:0080119	ER body organization	0.04829	EVM0005707.1
GO:0006508	proteolysis	0.04973	EVM0000084.3,EVM0000564.1,EVM0001532.1,EV

			M0001740.1,EVM0002103.1,EVM0002186.1,EVM0003131.1,EVM0003170.1,EVM0003187.1,EVM0003234.1,EVM0003238.1,EVM0003557.1,EVM0003715.1,EVM0003961.1,EVM0004102.1,EVM0004180.1,EVM0005579.1,EVM0005761.1,EVM0005921.1,EVM0006517.1,EVM0006800.1,EVM0007678.1,EVM0007735.1,EVM0009691.1,EVM0009900.1,EVM0009931.1,EVM0010826.1,EVM0011303.1,EVM0012105.1,EVM0012254.1,EVM0012670.1,EVM0013133.1,EVM0017478.1,EVM0018289.1,EVM0021681.1,EVM0022080.1,EVM0023096.1,EVM0023549.1,EVM0023930.1,EVM0025464.1,EVM0027066.1,EVM0033293.1,EVM0038627.1

Table 12 - GOs of *A. sisalana* Expanded Families

GO.ID	Term	Adjusted p-value	Genes
GO:0062075	pollen aperture formation	0.0012	g126427.t1,g127598.t1
GO:0030244	cellulose biosynthetic process	0.0022	g104353.t1,g118998.t1,g119145.t1,g121555.t1,g2502.t1
GO:0048578	positive regulation of long-day photoper...	0.0040	g106817.t1,g41509.t1
GO:0016560	protein import into peroxisome matrix, d...	0.0053	g145609.t1,g149052.t1
GO:0000301	retrograde transport, vesicle recycling ...	0.0142	g120577.t1
GO:0000389	mRNA 3'-splice site recognition	0.0142	g137030.t1
GO:0061137	bud dilation	0.0142	g141069.t1
GO:0070979	protein K11-linked ubiquitination	0.0142	g129624.t1
GO:0090213	regulation of radial pattern formation	0.0142	g115849.t1
GO:1901002	positive regulation of response to salt ...	0.0142	g13673.t1
GO:2000280	regulation of root development	0.0142	g11844.t1

GO:0034605	cellular response to heat	0.0178	g10412.t1,g119623.t1,g34649.t1
GO:0016121	carotene catabolic process	0.0212	g102347.t1,g117891.t1
GO:0045487	gibberellin catabolic process	0.0238	g100381.t1,g128239.t1
GO:1901259	chloroplast rRNA processing	0.0323	g129461.t1,g54852.t1
GO:0015979	photosynthesis	0.0331	g103292.t1,g105920.t1,g106012.t1,g113860.t1,g113911.t1,g118620.t1,g119211.t1,g128989.t1,g139072.t1,g15467.t1
GO:0005992	trehalose biosynthetic process	0.0395	g107021.t1,g12479.t1,g136468.t1
GO:0045489	pectin biosynthetic process	0.0414	g100234.t1,g140496.t1,g142810.t1
GO:0010143	cutin biosynthetic process	0.0418	g130568.t1,g132161.t1
GO:0010253	UDP-rhamnose biosynthetic process	0.0420	g104542.t1
GO:0015015	heparan sulfate proteoglycan biosyntheti...	0.0420	g145143.t1
GO:0090610	bundle sheath cell fate specification	0.0420	g126182.t1

Table 13 - GOs of *A. tequilana* Expanded Families

GO.ID	Term	Adjusted p-value	Genes
GO:0015979	photosynthesis	1.9e-05	AgveH2v21002187m,AgveH2v21006811m,AgveH2v21024312m,AgveH2v21024316m,AgveH2v21024317m,AgveH2v21024321m,AgveH2v21024411m,AgveH2v21024434m,AgveH2v21024469m,AgveH2v21024565m,AgveH2v21024589m,AgveH2v21077973m
GO:0006412	translation	0.00014	AgveH2v21005981m,AgveH2v21006815m,AgveH2

			v21009935m,AgveH2v21010796m,AgveH2v21016467m,AgveH2v21018957m,AgveH2v21019125m,AgveH2v21020046m,AgveH2v21021408m,AgveH2v21022004m,AgveH2v21022305m,AgveH2v21024334m,AgveH2v21024339m,AgveH2v21033675m,AgveH2v21052210m,AgveH2v21058839m,AgveH2v21061258m,AgveH2v21078470m,AgveH2v21092331m
GO:0015986	proton motive force-driven ATP synthesis	0.00665	AgveH2v21004325m,AgveH2v21005633m,AgveH2v21009885m
GO:0042538	hyperosmotic salinity response	0.01136	AgveH2v21058527m
GO:0019427	acetyl-CoA biosynthetic process from ace...	0.02259	AgveH2v21067296m
GO:1901031	regulation of response to reactive oxyge...	0.02259	AgveH2v21079634m
GO:0033355	ascorbate glutathione cycle	0.02816	AgveH2v21078485m
GO:0009767	photosynthetic electron transport chain	0.02914	AgveH2v21006811m,AgveH2v21024317m,AgveH2v21024434m
GO:0006571	tyrosine biosynthetic process	0.03369	AgveH2v21097577m
GO:0007229	integrin-mediated signaling pathway	0.04467	AgveH2v21059736m
GO:0009635	response to herbicide	0.04467	AgveH2v21046224m
GO:0010078	maintenance of root meristem identity	0.04467	AgveH2v21054868m
GO:0043068	positive regulation of programmed cell d...	0.04467	AgveH2v21067878m

ANEXOS

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação de Mestrado, intitulada Análise Bioinformática De Espécies De Agave: Explorando A Diversidade Genética E As Características Adaptáveis Para A Produção Sustentável De Biocombustíveis, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 11 de abril de 2025

Assinatura : Alexandra Russolo Cardelli

Nome do(a) autor(a): Alexandra Russolo Cardelli
RG n.° 36.962.416.-6

Assinatura : _____

Nome do(a) orientador(a): Marcelo Falsarella Carazzolle
RG n.°

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15** referente a Bioética e Biossegurança, declaro que o conteúdo de minha , intitulada " ***Análise Bioinformática de Espécies De Agave: Explorando a Diversidade Genética e as Características Adaptáveis para a Produção Sustentável de Biocombustíveis***", desenvolvida no Programa de Pós-Graduação em do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: Alexandra Russolo Cardelli

Nome do(a) aluno(a): Alexandra Russolo Cardelli

Assinatura: _____

Nome do(a) orientador(a): Marcelo Falsarella Carazzolle

Data: 11 de abril de 2025

UNICAMP

Of. CIBio/IB 35/2010

Cidade Universitária "Zeferino Vaz", 16 de agosto de 2011.

Prof. Dr. MARCELO BROCCHI
Chefe do Departamento Genética, Evolução e Bioagantes
Instituto de Biologia - UNICAMP

Prezado Professor:

Informamos que o projeto abaixo relacionado, envolvendo OGM do tipo I, sob responsabilidade do **Prof. Dr. GONÇALO A. G. PEREIRA**, protocolado sob o número **2011/03**, foi aprovado pela **CIBio-IB/UNICAMP**, em reunião sua 55ª. ordinária (15/08/2011) para ser desenvolvido nas dependências do Departamento Genética, Evolução e Bioagantes, Laboratório de Genômica e Expressão:

No. Projeto (data da aprovação)	Data de recepção	Nome do Projeto	Prazo para envio de relatório à CIBio
2011/03 (15/08/2011)	19/07/2011	Genômica e Biotecnologia, sub-projetos: 1) 2010/01 - Projeto Gene Discovery em Eucalipto; 2) Rotas Verdes para o Propeno, 3) Modificação de linhagens industriais de Saccharomyces cerevisiae para o aumento da produtividade e floculação condicional., 4) Projeto genoma de Crinipellis perniciosa, fungo causador da doença vassoura-de-bruxa do cacau, 5) Cultivo de microalgas para produção de cadeias carbônicas lipídicas, e 6) Transformação genética de cana-de-açúcar com o gene do inibidor de ripsina de inga laurina e análise da toxidade das plantas transgênicas sobre o desenvolvimento biológico de Diatraea saccharalis	Fevereiro/2012

Recomendamos que sejam observadas as instruções normativas referentes transporte e contenção da OGMs, disponíveis na webpage da CTNBio <www.ctnbio.gov.br>.

Atenciosamente,

Marcelo Lancellotti
Presidente da CIBio/IB-UNICAMP

C/C.: Prof. Dr. Gonçalo A. G. Pereira