De novo assembly and analysis of energy cane transcripts for sugarcane smut disease studies

Gabriela Romêro Campos

Dissertation presented to obtain the degree of Master in Science. Area: Genetics and Plant Breeding

Piracicaba

2024
Gabriela Romêro Campos
Agronomic Engineering

De novo assembly and analysis of energy cane transcripts for sugarcane smut disease studies

Advisor:
Dr. CLAUDIA BARROS MONTEIRO-VITORELLO

Dissertation presented to obtain the degree of Master in Science. Area: Genetics and Plant Breeding

Piracicaba
2024
Campos, Gabriela Roméro

De novo assembly and analysis of energy cane transcripts for sugarcane smut disease studies / Gabriela Roméro Campos. - - Piracicaba, 2024.

63 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

With all my love,

To Thamires and Claudio, who illuminate my path with love and support every day.
ACKNOWLEDGMENTS

I would like to express my sincere gratitude to everybody and all institutions that contributed to the success of this work.

I am thankful to the “Luiz de Queiroz” College of Agriculture at the University of São Paulo for the opportunity to participate in the Plant Genetics and Breeding program as a master's student and for their academic support.

I am grateful to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) for the financial support, and I extend my appreciation to BioVertis/GranBio and the Agronomic Institute of Campinas for providing the necessary plant material for this study.

I am immensely grateful to my advisor, Dr. Claudia Barros Monteiro Vitorello, who welcomed and supported me throughout my master's journey. I am extremely thankful for her trust and teachings. It was an honor to learn from a professional who is completely passionate about her work and who, with enthusiasm and dedication, cares to pass on this passion.

I also want to thank all my colleagues from the Genomics Group: Enrico Diniz, Thaís Dal’Sasso, Renato Bombardelli, Pedro Vilanova, Ana Julia Franceschi, Jonathan Macedo, Joyce Ferreti, Jéssica Mendes, Gustavo Crestana, João Vitor Ferro, Gustavo Husein, Gustavo Lima, Marcella Ferreira, Tiarla Souto, Maurício de Souza, and Elaine Vidotto. I deeply appreciate all the support they offered during my master's, taught me a lot, and whenever I needed help, there was someone willing to assist. Thank you for making the work environment pleasant and fun and, above all, for helping me grow professionally and personally.

I thank my husband, Claudio Dias da Silva Jr., for always being by my side and supporting me every day, giving me strength and courage to face each new challenge. I also express my gratitude to my sister, Thamires Romêro Campos, and my friends, Laura Panzarin Nerastri, Maria Fernanda Verissimo de Oliveira, and Nathan Barreto. Thank you for making life lighter and being there whenever I need you. Your love gives me the strength to continue.

Finally, I thank my parents, Marilene Romêro Neves Campos and Alexandre Costa Campos, for everything they have done for me that has led me to where I am today. Without you, none of this would be possible.
“A vida em seu nível mais simples e mais fundamental demonstra um nível impressionante de complexidade.”
- Unknown
SUMMARY

RESUMO

ABSTRACT

1. LITERATURE REVIEW
   1.1. Energy cane
   1.2. Sugarcane smut
   1.3. Genomics and Transcriptomics of Energy Sugarcane
   1.4. De novo Assembly of Transcriptomes from Polyploid Organism
   1.5. Objective
      1.5.1. General objective
      1.5.2. Specific objectives

REFERENCES

2. De novo Assembly and Analysis of Energy Cane Transcripts for Sugarcane Smut Disease Studies
   2.1. Abstract
   2.2. Introduction
   2.3. Materials and Methods
      2.3.1. Biological materials
      2.3.2. RNA Extraction and mRNA-Seq Library Construction
      2.3.3. De novo assembly and assessment of assembly completeness
      2.3.4. Evaluation of orthologs
      2.3.5. Transcriptome functional annotation and gene ontology analysis
      2.3.6. Differential expression and functional enrichment analyses
   2.4. Results
      2.4.1. Sequencing, de novo assembly and assessment of assembled transcriptome
      2.4.2. Evaluation of orthologs
      2.4.3. Transcriptome annotation
      2.4.4. Differential expression and functional enrichment analyses
   2.5. Discussion

REFERENCES

SUPPLEMENTARY MATERIAL
Montagem de novo e análise de transcritos de cana-energia para estudos de doenças do carvão da cana-de-açúcar

A cana energia, ao contrário da cana-de-açúcar convencional, foi selecionada para ter um maior teor de fibra do que sacarose. Nas últimas décadas, o cultivo de cana-energia aumentou devido ao grande potencial da cultura para a produção de energia renovável, possibilitando a redução na dependência de combustíveis fósseis e nos impactos ambientais causados por esses. Até o momento, nenhum genoma ou transcriptoma desta cultura foi relatado, principalmente no que diz respeito à interação com o fungo *Sporisorium scitamineum* ou outros patógenos. O *S. scitamineum* é um dos principais patógenos que atacam a cana energia, causando impactos significativos na produtividade. Ha informações disponíveis na interação entre o patógeno e a cana-de-açúcar, mas muito é necessário para entender a interação do patógeno com a cana energia. A fim de lançar alguma luz nesta interação, foi realizado duas montagens de novo do transcriptoma de genótipos Vertix 1 suscetíveis e Vertix 2 resistentes de cana-energia com dados de RNA-seq obtidos de gemas de cana-energia 48 horas após a inoculação (48hpi). Análises comparativas foram conduzidas utilizando os transcriptomas de novo de ambos os genótipos, cultivares modernas de cana-de-açúcar e espécies relacionadas. Além disso, foi realizada uma análise de expressão diferencial. Identificamos que a cana energia apresenta semelhanças com cultivares modernas de cana-de-açúcar e com a espécie *S. officinarum* e as respostas de defesa são semelhantes às observadas nas interações entre cana-de-açúcar e *S. scitamineum*. No entanto, ela ainda possui características próprias. Esses resultados trazem informações importantes sobre a interação cana-energia x *S. scitamineum*.

Palavras-chave: *Sporisorium scitamineum*, RNAseq, Interação planta-patógeno, Complexo *Saccharum*, Melhoramento genético de cana energia
ABSTRACT

De novo assembly and analysis of energy cane transcripts for sugarcane smut disease studies

Energy sugarcane has emerged as a significant resource of biomass for several countries, offering promising potential for the production of biofuels and other industrial products. To date, no genomes or transcriptomes of this crop have been reported, especially regarding interaction with the fungus *Sporisorium scitamineum* or other pathogens. The fungus *S. scitamineum* is one of the main pathogens that attack energy cane, causing significant impacts on yield. A good amount of information is available in the interaction between the pathogen and sugarcane, but much is needed to understand the interaction of the pathogen with energy cane. In order to shed some light in this interaction, two de novo assembly of the transcriptome of susceptible Vertix 1, and resistant Vertix 2 genotypes of energy cane was assembled with RNA-seq data obtained from energy cane buds 48 hours post inoculation (48hpi). Comparative analyses were conducted using both de novo transcriptomes, modern sugarcane cultivars, and related species. Additionally, a differential expression analysis was performed. We identified that despite energy cane presenting similarities with modern sugarcane cultivars and the species *S. officinarum* and as the defense responses are similar to those observed in interactions between sugarcane and *S. scitamineum*, it still has its own characteristics. These results bring important insights on the energy cane - *S. scitamineum* interaction.

Keywords: *Sporisorium scitamineum*, RNAseq, Plant-pathogen interaction, *Saccharum* complex, Energy cane breeding
1. Literature review

1.1. Energy cane

Energy cane has been developed as a promising bioenergy crop due to its high fiber content and potential for bioethanol production (Duval et al., 2013). The initial emergence of this crop dates back to the second half of the 1970s in Puerto Rico (Matsuoka et al., 2014), and since then, there has been an increase in the annual production of energy cane. One of the key events that has shaped the current scenario was the 1970 petroleum embargo led by the member countries of the Organization of the Petroleum Exporting Countries (OPEC), which led to the urgent need for alternative energy sources (Matsuoka et al., 2014). Different countries adopted distinct strategies in response to this situation. For instance, the United States opted for ethanol production from corn crops (Mumm et al., 2014), while Brazil chose sugarcane as the primary crop for ethanol production (Coombs, 1984; Goldemberg, 2008; Xavier, 2007). Through the National Alcohol Program (ProÁlcool), implemented by the Brazilian government in 1975, the sugarcane industry was able to develop, becoming one of the most efficient systems in converting photosynthates into energy in the present day (Cursi et al., 2022; Matsuoka et al., 2014).

The ProÁlcool program was initiated to produce anhydrous ethanol from sugarcane (Saccharum officinarum) for blending with gasoline. Following the second oil crisis in 1979, the program was expanded with the goal of producing ethanol as a substitute fuel for gasoline (Michellon et al., 2008). In this phase, ethanol was produced through the fermentation of sugarcane sugar. However, the industry faced challenges due to high operational costs, limited productivity, and competition between energy generation and food production (Santos et al., 2016). The use of cellulose as a raw material has opened up exciting opportunities for ethanol production. To meet industry demands for second-generation ethanol, a variety of technologies for biomass processing have been developed. These include genetically modified microorganisms that can efficiently convert C5 and C6 sugars into ethanol (Demeke et al., 2013), more streamlined processes for cellulosic ethanol production, and the creation of new hybrids known as energy cane that have higher biomass yields (Carvalho-Netto et al., 2014; Cursi et al., 2022; Silveira et al., 2016; Santos et al., 2016; Matsuoka, 2017).

The first attempts to create these new hybrids were carried out in Louisiana and Puerto Rico with the aim of increasing biomass production, regardless of sucrose content (Santos et al., 2016). Energy cane cultivars are obtained by crossing modern cultivars of S. officinarum with genotypes of S. spontaneum (Tew & Cobill, 2008). Compared to traditional sugarcane cultivars, energy cane has unique characteristics such as narrower leaf blades, thinner stems, increased...
tillering, and high productivity with rapid and vigorous initial development (Abreu et al., 2020; Matsuoka et al., 2014; Surendra et al., 2018). Moreover, energy cane is possessing increased resistance to biotic and abiotic stresses, allowing cultivation in regions not traditionally reserved for food production (Carvalho-Netto et al., 2014). Therefore, energy cane is an environmentally friendly crop as it requires fewer agrochemicals and fertilizers, captures carbon efficiently, and generates renewable energy (Carvalho-Netto et al., 2014; Barbosa et al., 2020).

Breeding programs aimed at developing energy cane hybrids have four main objectives. Firstly, increasing the potential for biomass production enhances the capacity for converting atmospheric carbon into organic carbon. Secondly, enhancing energy density makes energy cane more efficient than food plants. Thirdly, enhancing resistance to pathogens and biotic stresses reduces the crop’s dependence on agrochemicals. Lastly, adapting the root system improves carbon fixation efficiency in the soil and controls erosion (Cursi et al., 2022). There are variations in the strategies defined by breeding programs in the development of different energy cane cultivars. However, most institutions establish their strategies based on the desired relative proportion of sugar and fiber (Barbosa et al., 2020). Fiber and sugar content are the primary factors used to classify genotypes into two types (Tew & Cobill, 2008).

Type I cultivars are varieties selected to maximize sugar and fiber content. Comparatively, energy cane Type I presents higher fiber content than the traditional sugarcane, while low variation in sugar content (Tew & Cobill, 2008). Therefore, Type I cultivars can be used for both sugar and ethanol production. In addition, due to their higher fiber content, they contribute more to the production of second-generation ethanol (E2G) and electricity (Barbosa et al., 2020). Type II cultivars are selected to maximize fiber content only (Tew & Cobill, 2008). The demand for this type of material primarily comes from industries generating energy through biomass (Barbosa et al., 2020). Table 1 provides a comparison between two energy cane cultivars produced in Brazil by the company GranBio, Vertix 1 (Type I) and Vertix 2 (Type II), and traditional sugarcane.

<table>
<thead>
<tr>
<th>Tabela 1</th>
<th>Table 1: Main differences between Type I, Type II energy cane and sugar cane. (Adapted from Cursi et al. (2022)).</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sugarcane</td>
</tr>
<tr>
<td>Productivity (x)</td>
<td>x</td>
</tr>
<tr>
<td>Sugars (kg/t)</td>
<td>150</td>
</tr>
<tr>
<td>Fiber (%)</td>
<td>15</td>
</tr>
</tbody>
</table>
A third type of energy cane was proposed due to the identification of intrinsic variations in sugarcane and energy cane genotypes' biomass characteristics (Santchurn et al., 2014). Categorizing energy cane into three types improves the understanding and utilization of genetic variability present in genetic materials, allowing new strategies for breeding and more effective selection (Santchurn et al., 2014). The classification is based on the proposal by Tew and Cobill in 2008, however, Type I is further divided into two categories: genotypes with a higher fiber concentration without compromising sugar content and genotypes with a significant reduction in sugar content and a fiber content of no more than 22% (Santchurn et al., 2014) (Figure 1).

There are different types of energy cane that play distinct roles in the sugarcane industry (Matsuoka et al., 2014; Cursi, 2022). The industry requires exploration of significant genetic diversity to meet the demands of various sectors and to develop different cultivars. This is especially important for the sugar and fiber composition of species within the Saccharum and Miscanthus complexes (Barbosa et al., 2020). Major sugarcane-producing countries use accessions from the *Saccharum* genus and modern cultivars to produce new hybrids. In Brazil, key germplasms are maintained by the Sugarcane Technology Center (CTC), BioVertis/GranBio, Agronomic Institute of Campinas (IAC), and the Inter-university Network for the Development...
of the Sugarcane Industry (RIDESA) (Cursi et al., 2022). Other significant centers for research and crop improvement include The Sugarcane Breeding Institute (SBI) in India, the United States Department of Agriculture (USDA) in the United States, Sugar Research Australia (SRA) in Australia, the South African Sugarcane Research Institute (SASRI) in South Africa, the Mauritius Sugar Industry Research Institute (MSIRI) in Mauritius, and the West Indies Central Sugarcane Breeding Station (WICSCBS) in Barbados (Barbosa et al., 2020).

Specific production data for energy cane are scarce, as it is a relatively new crop. Generally, the production of energy cane is reported as a part of sugarcane production, which was the most produced crop worldwide until 2022, with Brazil leading as the largest producer globally, followed by India, China, and Thailand (FAOSTAT, 2024). The production capacity of energy cane varies significantly depending on the environmental conditions of each country. However, many countries that cultivate sugarcane also produce E2G and have specific breeding programs for energy cane (Figure 2). This indicates that energy cane has great potential as a sustainable alternative to power the global energy matrix.

![World sugarcane production in 2022 in tons. Each red dot on the map indicates that the country has a sugarcane breeding program and countries with the biofuel symbol produce ethanol. (Adapted from Barbosa et al. (2020), Diniz et al. (2019), FAOSTAT (2024)).](image-url)
During the 2022/23 harvest in Brazil, approximately 54.1% of the total sugarcane and energy cane produced were used to ethanol production. Notably, 57% of this production consisted of E2G (UNICA, 2024). Additionally, bioelectricity production from sugarcane is the fourth most important source in the Brazil's electrical matrix. It generates 72% (18.4 thousand GWh) of all biomass electricity (UNICA, 2024). If the biomass available in sugarcane fields were fully used, bioelectricity could potentially reach 151 thousand GWh, representing over 30% of the National Integrated System (SIN)'s energy consumption (UNICA, 2024). This highlights the increasing trend of using biomass for bioenergy production. Energy cane has the potential to further increase bioenergy production and meet the national demand for it.

The potential of energy cane biomass to generate biofuels and bioelectricity can be increased by managing the factors that limit its growth. Diseases affecting sugarcane and energy cane are among the main factors responsible for productivity losses, with sugarcane smut, caused by the biotrophic fungus *Sporisorium scitamineum*, being particularly detrimental, capable of causing losses of up to 100% when highly susceptible varieties are cultivated (Tokeshi & Rago, 2016). Therefore, it is essential to investigate the interaction between smut disease and energy cane to develop new strategies that mitigate these losses and promote more sustainable and productive cultivation.

### 1.2. Sugarcane smut

The sugarcane smut disease is caused by fungi belonging to the Basidiomycota phylum, which can attack various species, including both cultivated and wild (Singh et al., 2004). *Sporisorium scitamineum* (Syd.) (Syn: *Ustilago scitaminea*) is the agent responsible for smut disease in sugarcane and energy cane, representing one of the main threats to the crop (Bhuiyan et al., 2021; Monteiro Vitorello et al., 2018; Rajput et al., 2021). Anecdotal evidence suggests that the fungus originated in India, with species *S. barberi* and *S. spontaneum* acting as hosts (Bhuiyan et al., 2021). Southeast Asia was predicted as the center of origin for sugarcane smut through study of genetic diversity of different isolates (Cortes et al., 2024) since showed greater genetic diversity compared to isolates collected from 15 other countries, particularly in African and American populations (Braithwaite et al., 2024; Que et al., 2014a; Raboin et al., 2007; Shen et al., 2016; Xu et al., 2004).

The first report of the disease was in 1877 in Natal, South Africa (Lee-Lovick, 1978). Since then, it has spread to other producing countries in Central, East, and West Africa, Indonesia, Central and South America, Australia (Croft et al., 2008; Thokoane & Rutherford, 2001) and, recently it has been reported in Papua New Guinea (Tom et al., 2017). In Brazil, the first report was in São Paulo in 1946 (Rago et al., 2009) and genetic evidence suggests that the
dispersion was mediated by humans within Brazilian territory and among neighboring countries (Benevenuto et al., 2016). Currently, Fiji is the only sugarcane-producing country that has not been affected by this fungus (Bhuiyan et al., 2021; Monteiro Vitorello et al., 2018; Rajput et al., 2021).

The damages caused by sugarcane smut disease vary depending on the pathogen races, environmental conditions, cultivar genotype, as well as the interaction between these factors (Rajput et al., 2021). When plants get infected, their contribution to crop yield is minimal, which can result in total losses at harvest (Cortes et al., 2024; Tokeshi & Rago, 2016). The infection affects plant physiology, including changes in photosynthetic rate and significant reductions in sucrose content, brix, purity, and other indicators of juice quality. Plant morphology is negatively affected by the fungi, resulting in reduced height, stem diameter, number of internodes, and increased fibrous content of the stems (Bhuiyan et al., 2021; Mansoor et al., 2016; Marchelo-d’Ragga, 2016; Rajput et al., 2021). Productivity losses are correlated with the number of whips per area (Rajput et al., 2021). The whips are structures formed due to modifications in the apical meristem of the stem and can be easily identified in the field (Ferreira & Comstock, 1989). They vary in length from a few centimeters to over 1.5 meters (Monteiro Vitorello et al., 2018; Tokeshi & Rago, 2016), and contain a central core of parenchyma and vascular elements, surrounded by a cylinder of teliospores (Ferreira & Comstock, 1989; Marques et al., 2018).

Teliospores are resistance spores that result from karyogamy and hyphae fragmentation (Benevenuto et al., 2016). They are responsible for primary infection when they are spread through soil or planting infected stubble, and secondary infection when they are carried by air and infect healthy crops (Rajput et al., 2021). Teliospore germination under favorable environmental conditions, leading to germination in the probasidium and meiosis, resulting in the production of haploid sporidia. These sporidia grow as yeast-like haploid cells, and host infection occurs when two haploid sporidia of different mating types undergo plasmogamy, giving rise to a dikaryotic infective hypha (Monteiro Vitorello et al., 2018; Taniguti et al., 2015; Tokeshi & Rago, 2016). The infective hypha is capable of infecting the host through the formation of an appressorium, then colonizing it inter- and intracellularly systemically until reaching the apical meristem, eventually undergoing karyogamy, initiating sporogenesis, and inducing the formation of a new whip (Monteiro Vitorello et al., 2018; Taniguti et al., 2015; Tokeshi & Rago, 2016).

During the interaction, the pathogen faces several challenges imposed by the plant, known as defense mechanisms. Such mechanisms can be classified as external and internal or a combination of both (McNeil et al., 2018). External resistance mechanisms are related to physical and chemical barriers, such as the number of scales and quantity of trichomes in the bud region


(Longatto et al., 2015; Monteiro Vitorello et al., 2018) in addition to the presence of flavonoid compounds and phenylpropanoids present in these same regions (Fontaniella et al., 2002; Millanes et al., 2005) that act to inhibit and/or delay the germination of teliospores (Monteiro Vitorello et al., 2018). Internal resistance mechanisms are those expressed when the plant recognizes the presence of the fungus in internal tissues and triggers various defense responses. Among them is an increase in lignification (Santiago et al., 2012), the production of glycoproteins, phytoalexins and polyamines and hormonal changes (McNeil et al., 2018; Monteiro Vitorello et al., 2018). However, the presence and composition of such mechanisms vary among different sugarcane genotypes, resulting in different levels of plant resistance to the pathogen. It has already been observed that smut-resistant sugarcane genotypes perceive the presence of the fungus earlier compared to susceptible genotypes, which leads to extensive transcriptional reprogramming, mainly a strong oxidative burst and modulations in the antioxidant system (Peters et al., 2017).

Different methods can be used to manage sugarcane smut disease, including physically treating the setts with hot water, thinning out diseased plants, or applying fungicides. However, these strategies are not very efficient (Sundar et al., 2012). The most effective way to minimize the disease's impact on the crop is by using resistant varieties (Cortes et al., 2024; Rajput et al., 2021). Sugarcane and energy cane varieties exhibit different levels of resistance, and the disease's progression is influenced by the variety's resistance level (Bhuiyan et al., 2021). The resistance level is determined by the number of whips emitted by the fungus in an artificially inoculated population, with the scale ranging from no whips emitted to 12.5% of plants with whips (resistant varieties) to 25.6 to 100% of plants with whips (susceptible varieties) (Lemma et al., 2015).

Several studies have been conducted to determine factors that make a sugarcane variety resistant or another susceptible to the smut. A complex network of defense responses to pathogen attack is being discovered. It has been identified that resistant cultivars exhibit higher levels of ethylene, salicylic acid, and jasmonic acid compared to susceptible cultivars (Chen et al., 2023). Moreover, a reduction in lipid peroxidation has been observed, along with a decrease in catalase activity, an increase in glutathione S-transferase activity (Peters et al., 2017), and a significant increase in hydrogen peroxide accumulation, coinciding with the germination of the fungus's teliospores and the formation of appressoria, essential structures for pathogen infection (Peters et al., 2017). Transcriptome analysis of the S. sipaimewum-sugarcane interaction has led to the identification of differentially expressed genes associated with resistance in infected plants (Agisha et al., 2022; Huang et al., 2018; McNeil et al., 2018; Que et al., 2014). This information,
along with other data on the genetic architecture of energy cane, can help in the development of more resistant genotypes against smut.

1.3. Genomics and Transcriptomics of Energy Sugarcane

The genome of sugarcane is considered to be the most intricate among cultivated species (Piperidis & D’Hont, 2020). This is because of its complicated evolutionary history. The term "Saccharum complex" was coined to describe the closely related genera that contribute to the genetic modern cultivars (Mukherjee, 1957). This complex includes the genera Erianthus, Miscanthus, Narenga, Saccharum, and Sclerostachya (Tew & Cobill, 2008). Among these, Saccharum and Miscanthus are more closely related to each other than to the other genera (Hodkinson et al., 2002). There are six species included in the Saccharum genus: S. officinarum (x = 10, 2n = 80); S. robustum (x = 10, 2n = 60, 80); S. edule (2n = 60–80); S. barberi (2n = 111–120); S. sinense (2n = 81–124); and S. spontaneum (x = 8, 2n = 40–128) (Tew & Cobill, 2008). However, recently study proposed that there are three founder genomes in the Saccharum genus (A, B, and C), with A and B unevenly distributed in S. officinarum and its wild ancestor S. robustum, while the third, C, is observed in wild S. spontaneum (Pompidor et al., 2021). On the other hand, S. barberi and S. sinense exhibit a lesser relationship with the main Saccharum germplasm pool, suggesting possible introgression from other genera (Tew & Cobill, 2008) (Figure 3).

![Figure 3](image-url)

Figure 3. The evolutionary lineage of Saccharum hybrids is depicted as follows: Wild species are represented by white boxes, while cultivated species are depicted in gray. Arrows indicate hybridization events, with dashed lines indicating minor contributions to modern sugarcane cultivars. (Adapted of Setta et al. (2012), D’Hont et al. (2008), Thirugnanasambandam et al. (2018).

Modern sugarcane cultivars are interspecific hybrids resulting from crosses between accessions of S. officinarum and S. spontaneum (D’Hont et al., 2008). These hybrids have a mixture of genetic contributions from each of these parental species, along with undergoing multiple
events of aneuploidy and recombination between their genomes. In terms of polyploidy, they typically possess between 10 to 13 sets of their 10 basic chromosomes and generally consist of 70-80% of the genetic material from *S. officinarum*, 10-20% from *S. spontaneum*, and approximately 10% with chromosomes recombined between these two ancestors (Piperidis & D’Hont, 2020). Energy cane cultivars originate from interspecific hybridization between modern sugarcane cultivars and accessions of closely related wild species, with the most commonly used being *S. spontaneum* (Silva, 2017). *S. spontaneum* possesses greater genetic diversity within the *Saccharum* complex, is highly polymorphic (Silva, 2017; Tew & Cobill, 2008), and is utilized as a source of adaptability to different environments and resistance to biotic stresses (Silva, 2017).

The genome of modern sugarcane cultivars is complex due to crosses and hybridization. This complexity is further increased by high ploidy and frequent recombination, which can cause molecular and epigenetic changes that affect gene expression and phenotype (Madlung & Wendel, 2013). As a result of this complexity, sugarcane genetics has been slower to advance compared to other economically important crops, and breeding still primarily depends on conventional methods (Piperidis & D’Hont, 2020). While some progress has been made, such as the production of genetic maps based on single-dose markers (Aitken et al., 2014), the lack of high-quality multiple-dose markers makes it difficult to precisely identify homologs and assign them to homology groups. Although sugarcane genetic maps have become denser and cover larger fractions of the genome (Barreto et al., 2019; You et al., 2019).

The genomes of various sugarcane cultivars have been studied to understand their structure and evolution. The assembly of monoploid sequences based on the R570 cultivar (Garsmeur et al., 2018), the construction of 32 pseudo-chromosomes from the haploid genome of the AP85-441 cultivar of *Saccharum spontaneum* (Zhang et al., 2018), along with the draft genome sequence of the SP80-3280 cultivar (Riaño-Pachón & Martiello, 2017), partial sequences of the same cultivar (Souza et al., 2019), and the genome of the Khon Kaen 3 cultivar of sugarcane (Shearman et al., 2022), have greatly contributed to unraveling the genome structure and its evolutionary aspects. However, understanding the origins of multiple copies of a gene and its different alleles still presents significant challenges due to the high level of polyploidy and aneuploidy, along with structural variations in chromosomes. Especially when specific allelic expression occurs, which has been observed in sugarcane, mainly in genes related to the defense response and cell wall biosynthesis (Correr et al., 2022; Margarido et al., 2022).

Transcriptomic analysis is a highly valuable tool in studying sugarcane and energy cane characteristics. It provides insights into gene expression, isoform diversity, and elucidating the genetic mechanisms underlying specific genotype characteristics (Hoang et al., 2017).
Comparative transcriptome analysis allows for the identification of conserved and divergent gene expression patterns, evolutionary relationships, functional adaptations, and responses to environmental stimuli. Research on sugarcane transcript studies began in South Africa, analyzing 7,242 expressed sequence tags (ESTs) (Carson & Botha, 2002). The largest EST collection was developed by the Brazilian project SUCEST and consists of approximately 238,000 ESTs. Additionally, three more sets of ESTs were generated by researchers in Australia (Bower et al., 2005; Casu et al., 2003) and the United States (Ma et al., 2004). Currently, there are over 280,000 ESTs described and incorporated into the Sugarcane Gene Index (Cardoso-Silva et al., 2014).

Several studies involving transcriptome analysis have been developed to explore the genome and understand the molecular mechanisms underlying different phenotypic characteristics, such as sucrose accumulation, resistance to biotic and abiotic stresses, and vegetative development, among others (Huang et al., 2018; Taniguti et al., 2015; Thirugnanasambandam et al., 2017; Yan et al., 2021). Sugarcane smut disease is a significant biotic stressor, and the first publicly available transcriptome sequencing study of sugarcane using next-generation sequencing (NGS) used six sugarcane genotypes for de novo transcriptome analysis and assembly, identifying putative genes related to disease and sucrose accumulation (Cardoso-Silva et al., 2014). Other transcriptome analyses of the sugarcane-\textit{S. scitamineum} interaction led to the identification of similarly differentially expressed genes induced by \textit{S. scitamineum} in resistant and susceptible genotypes (Huang et al., 2018; Que et al., 2014; Schaker et al., 2016), as well as the identification of different resistance mechanisms (McNeil et al., 2018). Various defense signaling pathways activated based on pathogen virulence, hormonal signaling, and defense-related metabolite synthesis involved in defense mechanisms have also been identified in this pathosystem through transcriptomics (Agisha et al., 2022). However, there is still much unknown information in the sugarcane and energy cane genomes, highlighting the need for ongoing exploration.

1.4. \textit{De novo} Assembly of transcriptomes from polyploid organism

RNA-Seq is a widely used methodology for identifying differentially expressed genes and discovering SNPs (Single Nucleotide Polymorphisms) in both model and non-model organisms (Chopra et al., 2014; Duan et al., 2012). When working with organisms that have known reference genomes, the mapping approach is used. This involves aligning reads to annotated references, followed by transcript assembly, SNP identification, and quantification of transcription expression levels based on this mapping information (Chopra et al., 2014). When there is no well-defined reference genome, various strategies can be employed, such as using
references from related species, EST assembly or de novo assembly of RNA-Seq data (Chopra et al., 2014). However, using related species as a reference may result in a loss of species-specific information, while EST assembly requires extensive information on ESTs or a genomic database (Chopra et al., 2014).

RNA-Seq proves to be a valuable option for de novo transcriptome assembly and it helps in understanding molecular and functional mechanisms (Hölzer & Marz, 2019; Madritsch et al., 2021). However, de novo transcriptome assembly has challenges due to the presence of various alleles, closely related paralogs, and homologs, leading to considerable diversity of isoforms (Chopra et al., 2014; Góngora-Castillo & Buell, 2013; Madritsch et al., 2021). This challenge is even more significant in complex polyploid plants due to gene duplications, dosage imbalance, and allele-specific gene expression and the presence of multiple homoeologs in allopolyploids, which adds a level of complexity (Madritsch et al., 2021; Voshall & Moriyama, 2020), as is the case with sugarcane and energy cane. These characteristics may lead to increased rates of fused or redundant transcripts (Madritsch et al., 2021).

To improve the accuracy of de novo assembly in transcriptome with polyploid characteristics, it is essential to utilize high-quality materials such as non-fragmented RNA and long sequencing reads (Gutierrez-Gonzalez & Garvin, 2017). It is also important to ensure satisfactory coverage and use an appropriate analysis pipeline to avoid confusion between homoeologous nucleotide differences and sequencing errors (Gutierrez-Gonzalez & Garvin, 2017). Polyploidy has significant impacts on transcript assembly and quantification, especially in separate expression levels between homoeologs or subgenomes (Voshall & Moriyama, 2020). Indeed, allopolyploids may exhibit unequal expression among duplicated genes, known as homoeolog expression bias, which can result in differences in proteome composition and gene expression, even in phenotypically similar allopolyploids (Voshall & Moriyama, 2020; Yoo et al., 2012). Although mapping to a reference genome can correct expression levels between genes in most cases sequence variations among homologs can still lead to inaccurate expression estimates (Voshall & Moriyama, 2020).

In response to these challenges, several strategies and tools have been developed. One alternative to traditional alignment for RNA-seq quantification analyses is the use of pseudo-alignment methods such as Salmon (Patro et al., 2017) and Kallisto (Bray et al., 2016). By employing an expectation-maximization approach (Li et al., 2010) to assign reads to transcripts in situations of ambiguity, these methods can resolve mappings between various homoeologs. This is possible when a sufficient number of reads are exclusively mapped to only one homoeolog, allowing for the estimation of overall expression levels (Voshall & Moriyama, 2020).
Furthermore, pseudo-alignment is more computationally efficient as it does not require exact alignment to genomic references, resulting in faster processing and reduced resource usage (Bray et al., 2016; Patro et al., 2017). By overlapping information between reads and transcripts, pseudo-alignment enables more precise assignment of reads, contributing to more reliable estimates of gene expression levels in polyploid organisms.

Most software developed for assembly based only on RNA-seq data, such as Trinity (Grabherr et al., 2011), employs de Bruijn graphs (Compeau et al., 2011) for assembly. This method involves breaking reads into shorter overlapping sequences of a given length k, called kmers. This approach helps in efficient traversal of overlaps to reconstruct the original sequence (Voshall & Moriyama, 2020). It is also essential to balance the length of kmers with their limitations. Shorter kmers are more likely to cover the transcription sequences that need to be assembled completely. Still, they also increase the risk of incorrect assemblies when the same k-mer is present in reads from multiple transcripts (Voshall & Moriyama, 2020). These limitations must be balanced with the length of kmers. Shorter kmers are more likely to completely cover the transcription sequences that need to be assembled, but they also have a higher likelihood of resulting in incorrect assemblies when the same k-mer is present in reads from multiple transcripts (Voshall & Moriyama, 2020). These issues become more complicated for polyploid transcriptome assembly, as different yet highly similar sequences between alleles, as well as homoeologs, lead to more ambiguity in the graph. This greatly enlarges the graph size and increases the probability of traversal errors (Voshall & Moriyama, 2020).

There are software tools available freely or commercially that have been successful in complex organisms (Chopra et al., 2014). These include SOAPdenovo-Trans (Xie et al., 2014), rnaSPAdes (Bushmanova et al., 2019) and Trinity (Grabherr et al., 2011). Trinity stands out as one of the most popular software tools for generating high-quality \textit{de novo} transcriptomes with low base error rates and the ability to capture multiple isoforms, which are crucial for maintaining acceptable levels of accuracy when characterizing genes (Chopra et al., 2014). Despite these strategies, assembling allopolyploid transcriptomes \textit{de novo} remains a challenging task. However, the \textit{de novo} assembly of the energy cane transcriptome still offers significant advantages for scientific studies. This is primarily attributed to the lack of a genomic sequence that fully represents a complete hybrid genome of these plants. Energy cane is a crop that faces these challenges as they are allopolyploid hybrids, and up to this moment, genomes or transcriptomes have not been reported, especially regarding interaction with the fungus \textit{S. scitamineum} or other pathogens. In this scenario, we aim to perform \textit{de novo} assembly of two transcriptomes, using genotypes contrasting in resistance to the pathogen. What made it possible to explore the
differences in response between the transcriptomes of both genotypes and, through comparative analyses with modern sugarcane cultivars and species related to energy cane, identify whether there is specific information regarding energy cane.

1.5. Objective

1.5.1. General objective

This work aims to construct a reference transcriptome of contrasting energy cane varieties for infection with \textit{S. scitamineum}. We aim to compare the molecular events involved in the response of resistant and susceptible plants and to compare them to those described for sugarcane. We anticipate that these signatures will differ from those previously described for sugarcane, particularly considering the discrepancy in initial growth rates between sugarcane and energy cane. By comparing these transcriptomic profiles, we aim to illuminate the path to understanding fundamental molecular mechanisms underlying resistance and susceptibility to \textit{S. scitamineum} in energy cane varieties.

1.5.2. Specific objectives

1) Process the sequencing data obtained from experiments involving plants infected and non-infected with \textit{S. scitamineum} regarding parameters such as quality, sequencing depth, completeness, and others for subsequent analyses.

2) Perform assemblies using custom programs for transcriptomic data analysis and select those that best fit the analysis of the obtained sequences.

3) Compare contrasting varieties regarding their response to infection among themselves and other selected genotypes of closely related species with available sequences.
REFERENCES


2. De novo assembly and analysis of energy cane transcripts for sugarcane smut disease studies

2.1. Abstract

The crescent demand for renewable energy has led to the development of sustainable strategies, such as energy cane (Saccharum spp. hybrids), which emerges as a promising source of biomass for second-generation ethanol (E2G) production. These cultivars result from the crossing between accessions of the Saccharum complex and modern sugarcane varieties, aiming to select materials with high biomass production. In addition to E2G production, energy cane has potential for various industrial uses, including sugar production, biodiesel, biopolymers, fibers for thermoplastics, and bioelectricity. However, the interaction between energy cane genotypes and pathogens, such as the fungus S. scitamineum, has not been widely explored. In this study, we aimed to define a transcriptome for two energy cane genotypes (Vertix 1 and Vertix 2) contrasting in response to sugarcane smut and created a reference aiming to maintain variability while reducing redundancy through orthology. We performed functional analysis of the transcriptome, identifying differentially expressed and functionally enriched genes. We identified that despite energy cane presenting similarities with modern sugarcane cultivars and the species S. officinarum, it still has its own characteristics. Additionally, we observed that the infection and colonization of the fungus caused modifications, and we identified differentially expressed transcripts related to auxin response, reactive oxygen species, lignification, among other defense mechanisms. Our results mainly highlight the complexity of the energy cane transcriptome and provide an initial basis for future investigations into the interactions between this crop and its pathogens, as well as its distinct characteristics compared to other sugarcane varieties.

Keywords: Saccharum spp., Sporisorium scitamineum, RNAseq, Ortology
REFERENCES


