University of São Paulo
"Luiz de Queiroz" College of Agriculture

Dissecting expression patterns in the transcriptome of immature sugarcane culms: from methodology to biology

Victor Hugo de Mello Pessoa

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

Piracicaba
2020

Victor Hugo de Mello Pessoa
Bachelor of Physical and Biomolecular Sciences

Dissecting expression patterns in the transcriptome of immature sugarcane culms: from methodology to biology

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. PhD. **GABRIEL RODRIGUES ALVES MARGARIDO**

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

Piracicaba
2020

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Pessoa, Victor Hugo de Mello

Dissecting expression patterns in the transcriptome of immature sugarcane culms: from methodology to biology / Victor Hugo de Mello Pessoa. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2020 .

86 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Saccharum 2. Partição de carbono 3. Expressão diferencial 4. Réplicas biológicas 5. RNA-Seq . I. Título.

DEDICATÓRIA

Qualquer ser humano que passou pela pós-graduação sabe que, a despeito do que reforçam os setores interessados em desinformação e ausência de pensamento crítico, a atividade científica é séria e essencial para manutenção da sociedade. Nosso trabalho nunca para, porque perpassa nossas vidas em todas as esferas: dentro e fora do laboratório, no bandejão, em casa, em família, com amigos, nas horas de lazer, no banho e similares, muitas vezes nos impedindo de viver uma vida "normal". Não é raro que a pressão torne este período ainda mais difícil do que é. Esta dedicatória se destina a todos que o reconhecem e fornecem os meios para que essa importante profissão seja defendida e difundida, acessível a todes. Dentre estes, destina-se em especial para os que contribuíram para o meu processo.

À Michelle, que lutou muito para que eu estivesse aqui desde que nasci. Você sempre me inspirou com seu exemplo e com seus sonhos, e felizmente incentivou muitas pessoas ao seu redor. É um privilégio imenso ter sido educado com tanto afeto e compreensão. Ao mesmo tempo é um desafio, porque ser crítico passa por discordar, debater, ficar de mal e fazer as pazes. Nisso, eu aprendi com uma especialista. Ah, e ainda te lembro: este documento é um passo mais próximo para o nosso acordo.

À minha família que me defende a unhas e dentes, que me surpreende e que me apoia em minhas (in)decisões. Sempre esteve ao meu lado, presencialmente sempre que possível, mas também compreensíveis quanto a minha ausência. Me encorajou e confiou em mim ainda que eu não confiasse. Logicamente, se estende a minha família consanguínea ou não.

Aos meus amigos de laboratório (inclusive os longínquos agregados que nos visitaram), dentre os quais estendo aos laboratórios adotivos que tentaram me roubar mais de uma vez. Também se estende aos caros roomies, que nutriram minha saúde mental. Me mostraram que às vezes é normal confundir caquis e carambolas, mesmo que você estude muito botânica. Vocês me nortearam, ensinaram, quase sempre sem perceber. Me proporcionaram aprendizados culturais, artísticos, técnicos e humorísticos. Neste último ponto, me ajudaram a regredir também, o que também é um progresso.

Aos meus companheiros de viagem das cidades por onde passei (ou que passaram por mim), que me deixam com saudades constantemente e me fazem prometer a mim mesmo que irei voltar para visitar. Vocês nunca me deixaram ficar só de verdade, mesmo eu sendo relapso. Sei que tenho muitas casas para onde posso correr em caso de um vírus apocalíptico.

Aos que construíram gerações de conhecimento antes de mim e aos que virão. A ciência se faz com trabalho coletivo.

E à Pandora e à Sassá.

ACKNOWLEDGEMENTS

I thank all the people that made this work viable, mainly my supervisor Gabriel Margarido, because of his efforts to clarify my doubts and to create a friendly environment for developing scientific criticism as well as my own research pathways. I also have a special acknowledgment to my laboratory friends and researchers from UFSCar Araras, which have performed an excellent experimental job with the sugarcane plants, providing a flawless dataset for the analysis I performed. I express my gratitude to the Graduate Program of Genetics and Plant Breeding and the USP teachers who have contributed to my scientific formation, besides everyone who encouraged me and allowed the development of university extension activities. Finally, I recognize the funding agencies CAPES and FAPESP, which promoted the scholarship maintenance, and to CNPq, which, together with FAPESP, have funded the projects wherein I participated.

CONTENTS

RESUMO . . . 7
ABSTRACT . . . 8
GENERAL INTRODUCTION . . . 9
REFERENCES . . . 11
OBJECTIVES . . . 13
CHAPTER ONE: An RNA-Seq-based comparison of biological replication strategies for differential gene expression in *Saccharum* . . . 15
1 Introduction . . . 17
2 Materials and Methods . . . 19
2.1 Biological material and RNA-Seq . . . 19
2.2 Downsampling and quality control . . . 19
2.3 *De novo* transcriptome assembly and functional annotation . . . 20
2.4 Comparison of differential expression results with the full dataset . . . 20
2.5 Impact of missing samples on differential expression results . . . 21
3 Results . . . 23
3.1 Gene identification in the sugarcane transcriptome . . . 23
3.2 Comparison of differential expression results between strategies . . . 23
3.3 Assessment of strategies using subsets of samples . . . 25
3.4 Contribution of SBDG exclusive genotypes for differential expression . . . 27
4 Discussion . . . 29
5 Conclusions . . . 33
6 Abbreviations . . . 35
REFERENCES . . . 36
CHAPTER TWO: Identifying expression profiles associated with diverse carbon partitioning phenotypes in young sugarcane stems . . . 39
1 Introduction . . . 41
2 Materials and Methods . . . 43
2.1 Biological samples and phenotypic characterization . . . 43
2.2 RNA-Seq and data preparation . . . 43
2.3 Transcriptome assembly and functional annotation . . . 44
2.4 Assessment of individual conserved orthologues in the transcriptome . . . 44
2.5 Differential expression and functional enrichment analyses . . . 44
3 Results . . . 47
3.1 Assembly statistics . . . 47
3.2 Assembly evaluation for conserved orthologues in Poales . . . 47
3.3 Differential expression and functional enrichment results . . . 49
3.3.1 Sugar conversion enzymes . . . 52
3.3.2 Cellulose and other fiber components . . . 54
3.3.3 Sugar transmembrane transporters . . . 54
3.3.4 Hormone-related genes . . . 54
3.3.5 Pollen recognition . . . 55
3.3.6 Transposable elements . . . 55
4 Discussion . . . 57
4.1 *De novo* assembly challenges . . . 57

4.2 Sucrose synthesis and breakdown in young internodes 57
4.3 Sugar transporters role on sucrose storage 60
4.4 Transposable elements and genotype relatedness 62
5 Conclusions 63
6 Abbreviations 65
REFERENCES 66
SUPPLEMENTARY 71

RESUMO

Dissecando padrões de expressão no transcriptoma de colmos imaturos de cana-de-açúcar: da metodologia à biologia

O genoma da cana-de-açúcar é, de várias formas, o mais complexo dentre as plantas cultivadas, devido à sua alta ploidia, heterozigosidade e histórico de eventos de hibridização. Apesar de esforços nos últimos anos para se construírem três referências genômicas por grupos distintos, estas sequências ainda representem uma informação incompleta sobre os genomas de cana-de-açúcar. A presente dissertação contempla duas análises centrais para explorar o transcriptoma de cana-de-açúcar, visando trazer informações biológicas sobre a expressão gênica em colmos imaturos bem como reflexões metodológicas para o planejamento de experimentos de expressão diferencial. O primeiro capítulo apresenta uma comparação metodológica de duas estratégias visando ilustrar a influência de réplicas biológicas para plantas propagadas vegetativamente, como é o caso da cana-de-açúcar. Estas análises compararam o uso de clones ao uso de um conjunto diverso de genótipos como componentes dos grupos contrastantes de amostras. Os resultados indicam que o uso de clones permitiu a detecção de um maior número de genes diferencialmente expressos, provavelmente incluindo genes de efetivo interesse entre genes induzidos em genótipos específicos. Por outro lado, o uso de genótipos diversos proporcionou menos genes diferencialmente expressos, mas com aparentemente maior proporção de genes biologicamente relevantes. Esta proposição foi corroborada tanto pelos resultados do enriquecimento funcional quanto pelo conjunto de genes detectados em comum pelas estratégias. O segundo capítulo apresenta uma investigação biológica sobre os mecanismos genéticos pelos quais ocorre a partição de carbono em colmos apicais, onde o processo de acúmulo de sacarose não está desenvolvido. Genes diferencialmente expressos foram identificados para metabolismo e transporte de sacarose, tais como os genes de sacarose sintase, invertases e transportador de sacarose. Entretanto, o fenômeno mais notável relativo à partição de carbono foi a biossíntese de componentes da parede celular. Estes estudos podem trazer novas perspectivas para pesquisas sobre genética de cana-de-açúcar, por apresentarem um conjunto de genes de interesse para o metabolismo de açúcares e fibra, bem como conduzindo a uma escolha consciente do delineamento experimental para análises de RNA-Seq.

Palavras-chave: Saccharum, Partição de carbono, Expressão diferencial, Réplicas biológicas, RNA-Seq

ABSTRACT

Dissecting expression patterns in the transcriptome of immature sugarcane culms: from methodology to biology

The sugarcane genome is, by all accounts, the most complex among the cultivated crops due to its high ploidy, heterozygosity and history of hybridization events. Despite substantial efforts in the past years to obtain three genomic references by different groups, these sequences still represent incomplete information about sugarcane genomes. The current master thesis presents two core analyses to explore the sugarcane transcriptome, with the goal of bringing both biological insights about gene expression in immature culms as well as methodological considerations for the planning of differential expression experiments. The first chapter presents a methodological comparison of two strategies aiming to illustrate the influence of biological replication for vegetatively propagated plants, such as sugarcane. These analyses compared the use of clones and a diverse set of genotypes as components of contrasting groups of samples. The results indicate that the use of clones yielded an increased number of differentially expressed genes, which likely include genes of actual biological interest amidst genotype-specific significant tests. On the other hand, the use of diverse genotypes provided fewer differentially expressed genes, but the proportion of biologically relevant genes was seemingly higher. This statement was supported by evidence from both functional enrichment tests as well as the set of shared genes detected between the strategies. The second chapter presents a biological inquiry about the genetic mechanics regarding carbon partitioning in apical culms, where the sucrose accumulation process has not yet unfolded. Differentially expressed genes were identified for sucrose metabolism and transport, such as sucrose synthase, invertases, and sucrose transporter. However, the most apparent phenomenon with regard to carbon partitioning was the biosynthesis of cell wall components. These studies could drive new insights into sugarcane genetic investigations, by providing a set of important genes for early fiber and sugar metabolism in sugarcane, as well as aid researchers in making a more careful choice of experimental design for RNA-Seq essays.

Keywords: Saccharum, Carbon partitioning, Differential expression, Biological replicates, RNA-Seq

GENERAL INTRODUCTION

Sugarcane is one of the most valuable crops worldwide due to its importance for sugar, ethanol, and, more recently, biomass production. However, sugarcane has a highly complex genome. Modern cultivars result from hybridization between *Saccharum officinarum*, a high-sugar domesticated grass, and *S. spontaneum*, a fibrous plant tolerant to a broad range of biotic and abiotic stresses. Both parental species are autopolyploids (Bremer, 1925; Panje & Babu, 1960) — *S. officinarum* likely has eight sets of ten chromosomes (D'Hont *et al.*, 1998), while the number of chromosomes in *S. spontaneum* ranges from 40 to 128 —, such that sugarcane hybrids are auto-allopolyploids with frequent aneuploidy. Moreover, the number of chromosomes varies among genotypes (Piperidis *et al.*, 2010), as well as within the same genotype (D'Hont *et al.*, 1996).

Among the species of the genus *Saccharum*, *S. spontaneum* includes accessions with the most variable morphological features and chromosome number, as well as with a broader geographic span, from Northeast Africa to the Pacific Islands. Ecological and morphological adaptations allowed the species to thrive in widely diverse habitats. For instance, plants exhibit a great variation in size, ratooning capacity, amount of juice, stalk color, as well as tolerance to growing in dry soil or submerged into river waters (Mary *et al.*, 2006). When compared to modern cultivars, it has lower sucrose content, higher fiber yield, increased ratooning performance, and thinner stalks. Its high tolerance to biotic and abiotic stresses led to efforts of hybridization at the end of the 19th century, to introgress these traits to the high-sugar *S. officinarum*.

The narrow genetic base of the hybrids from the first decades of the 20th century has driven to issues such as the need to introgress new traits and the decreased rate of genetic gain, which led to new attempts of crossing with the parental species. However, a study using molecular markers showed that the diversity currently captured by breeding programs is still low, when considering the contribution of *S. spontaneum* (Aitken *et al.*, 2018). The use of new accessions as sources of alleles can be particularly relevant for introducing desirable traits for improving the so-called energy cane, given the growing allocation of sugarcane resources to the production of ethanol.

Besides classical plant breeding programs, both public and private initiatives have been developing genetically modified sugarcane aiming to tackle several agronomical issues. Transgenic sugarcane harboring genes for resistance to insects (Gao *et al.*, 2016; Cristofoletti *et al.*, 2018) and viruses (Yao *et al.*, 2017), for conferring drought tolerance (Zhao *et al.*, 2020), and for increasing sugar yield (Anur *et al.*, 2020) are examples of these. Yet, the lack of detailed information about the sugarcane genome hinders the understanding of how molecular mechanisms happen and can be leveraged for sugarcane breeding.

Genomic or transcriptomic assays are two approaches that can be used to acquire data for this purpose. Genomic studies are particularly complex to be performed for sugarcane, due to the high ploidy numbers and heterozygosity. There are three major scientific studies for presenting a comprehensive view of *Saccharum* genomes. The hybrids R570 and SP80-3280, as well as the haploid *S. spontaneum* accession AP85-441, had their genome sequenced using different methodologies, providing our best knowledge of the sugarcane genome to date (Garsmeur *et al.*, 2018; Souza *et al.*, 2019; Zhang *et al.*, 2018). Despite the advance brought by these studies, they still represent incomplete assemblies and pose obstacles when used as references. On the other hand, transcriptomic studies can be performed with fewer complications when compared to genomic ones. In particular for RNA-Seq analyses in non-model species, the currently used de novo assemblers do not necessarily require genomic references and are able to deal with polymorphisms present in different alleles (at least for highly expressed genes).

One suitable approach for detecting genes involved in biological processes is using differential expression analyses to identify up and downregulated genes in comparisons of interest. For instance, this strategy has been used to identify genes related to sugar yield (Papini-Terzi *et al.*, 2009; Thirugnanasam-

bandam *et al.*, 2017), fiber content (Vicentini *et al.*, 2015), drought stress (Li *et al.*, 2016), and resistance to smut (Rody *et al.*, 2019). Because gene expression is cell, tissue, and organ-dependent, the part of the plant chosen to be sampled provides data to answer different biological questions. Historically, sucrose yield is at the spotlight of studies regarding not only gene expression, but also enzymatic activity, plant physiology, and cellular biochemistry. However, these studies frequently focus on mature internodes or immature internodes at a late developmental stage. The current dissertation comprises two projects using sampled immature internodes in the earliest stage of development, which is right below the apical meristem.

The RNA-Seq data used as input to identify differentially expressed genes often fits into three sources of comparison: different tissues (organs) of the same plant; plants under different treatment levels; or plants with different genotypes, often selected based on their phenotypes. The first two examples do not necessarily depend on the choice of genotype, because the genomic composition of the contrasting groups is often identical. On the other hand, the latter case has a marginal effect of the combination of genotypes, regardless of the experimental design, such as in comparisons of high versus low sugar genotypes or susceptible versus tolerant plants to a pathogen. Also, especially for sugarcane and other vegetatively propagated crops, the use of clones or elite lines as biological replicates is a frequently adopted sampling strategy for the comparison of expression patterns. Another equally valid strategy and also used in gene expression assays is selecting different genotypes, grouped by a shared phenotypic trait. In chapter one, we analyze these different sampling strategies for biological replication, in which the group of interest is formed by clones or by a diverse set of genotypes (Figure 1). This study aims to compare the outcomes of differential expression analyses corresponding to these strategies and to provide a reference for the experimental design of future researches using RNA-Seq data under these conditions.

Little is known about the immature internode +1, which is the youngest part of sugarcane stalk, and remains largely unexplored with regard to its transcriptome. Chapter two presents an investigation of expression patterns in this organ for plants contrasting in sucrose levels, presenting a set of differentially expressed genes and enriched biological functions. Here, we identified putative markers for carbon partitioning before the start of sucrose accumulation.

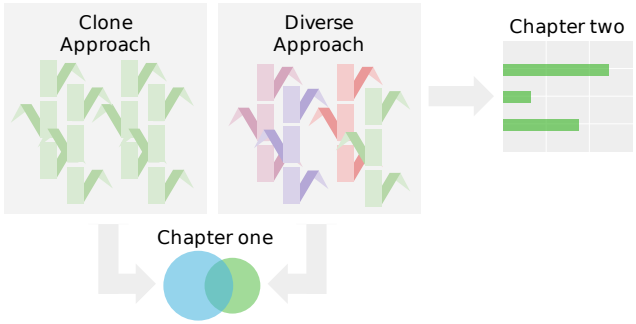


Figure 1: Scheme representing the use of biological data in the analyses performed in chapters one and two.

REFERENCES

Aitken, K., Li, J., Piperidis, G., Qing, C., Yuanhong, F., & Jackson, P. (2018). Worldwide genetic diversity of the wild species Saccharum spontaneum and level of diversity captured within sugarcane breeding programs. *Crop Science, 58*(1), 218-229.

Anur, R. M., Mufithah, N., Sawitri, W. D., Sakakibara, H., & Sugiharto, B. (2020). Overexpression of Sucrose Phosphate Synthase Enhanced Sucrose Content and Biomass Production in Transgenic Sugarcane. *Plants, 9*(2), 200.

Bremer, G. (1925). The cytology of the sugarcane. *Genetica, 7*(3), 293-322.

Cristofoletti, P. T., Kemper, E. L., Capella, A. N., Carmago, S. R., Cazoto, J. L., Ferrari, F., ... & Santos, N. Z. (2018). Development of transgenic sugarcane resistant to sugarcane borer. *Tropical Plant Biology, 11*(1-2), 17-30.

D'Hont, A., Grivet, L., Feldmann, P., Glaszmann, J. C., Rao, S., & Berding, N. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Molecular and General Genetics MGG, 250*(4), 405-413.

D'Hont, A., Ison, D., Alix, K., Roux, C., & Glaszmann, J. C. (1998). Determination of basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. *Genome, 41*(2), 221-225.

Gao, S., Yang, Y., Wang, C., Guo, J., Zhou, D., Wu, Q., ... & Que, Y. (2016). Transgenic sugarcane with a cry1Ac gene exhibited better phenotypic traits and enhanced resistance against sugarcane borer. *PLoS One, 11*(4), e0153929.

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., ... & Costet, L. (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature communications, 9*(1), 2638.

Li, C., Nong, Q., Solanki, M. K., Liang, Q., Xie, J., Liu, X., ... & Li, Y. (2016). Differential expression profiles and pathways of genes in sugarcane leaf at elongation stage in response to drought stress. *Scientific reports, 6*, 25698.

Mary, S., Nair, N. V., Chaturvedi, P. K., & Selvi, A. (2006). Analysis of genetic diversity among Saccharum spontaneum L. from four geographical regions of India, using molecular markers. *Genetic Resources and Crop Evolution, 53*(6), 1221-1231.

Panje, R. R., & Babu, C. N. (1960). Studies in Saccharum spontaneum distribution and geographical association of chromosome numbers. *Cytologia, 25*(2), 152-172.

Papini-Terzi, F. S., Rocha, F. R., Vêncio, R. Z., Felix, J. M., Branco, D. S., Waclawovsky, A. J., ... & Vicentini, R. (2009). Sugarcane genes associated with sucrose content. *BMC genomics, 10*(1), 120.

Piperidis, G., Piperidis, N., & D'Hont, A. (2010). Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. *Molecular Genetics and Genomics, 284*(1), 65-73.

Rody, H. V., Bombardelli, R. G., Creste, S., Camargo, L. E., Van Sluys, M. A., & Monteiro-Vitorello, C. B. (2019). Genome survey of resistance gene analogs in sugarcane: genomic features and differential expression of the innate immune system from a smut-resistant genotype. *BMC genomics, 20*(1), 809.

Souza, G. M., Van Sluys, M. A., Lembke, C. G., Lee, H., Margarido, G. R. A., Hotta, C. T., ... & Nishiyama Jr, M. Y. (2019). Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience, 8*(12), giz129.

Thirugnanasambandam, P. P., Hoang, N. V., Furtado, A., Botha, F. C., & Henry, R. J. (2017). Association of variation in the sugarcane transcriptome with sugar content. *BMC genomics, 18*(1), 909.

Vicentini, R., Bottcher, A., dos Santos Brito, M., dos Santos, A. B., Creste, S., de Andrade Landell, M. G., ... & Mazzafera, P. (2015). Large-scale transcriptome analysis of two sugarcane genotypes contrasting for lignin content. *PloS one, 10*(8).

Yao, W., Ruan, M., Qin, L., Yang, C., Chen, R., Chen, B., & Zhang, M. (2017). Field performance of transgenic sugarcane lines resistant to sugarcane mosaic virus. *Frontiers in Plant Science, 8*, 104.

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., ... & Wai, C. M. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nature genetics, 50*(11), 1565-1573.

Zhao, X., Jiang, Y., Liu, Q., Yang, H., Wang, Z., & Zhang, M. (2020). Effects of drought-tolerant Ea-DREB2B transgenic sugarcane on bacterial communities in soil. *Frontiers in Microbiology, 11*, 704.

OBJECTIVES

In chapter one, we aim to perform a comparison of sampling strategies for biological replication, namely the use of clonal plants or a diverse set of genotypes, grouped by common phenotypic traits. The analyses address biological insights in favor or against the use of each strategy, concerning both differential expression and functional enrichment tests. Our main objective is to evaluate these methodologies for rational planning of future experimental designs in sugarcane and other vegetatively propagated species.

In chapter two, we aim to investigate expression patterns of genes associated with carbon partitioning, using RNA-Seq data obtained from internode +1 of sugarcane stalks, which have not been thoroughly studied for this purpose. This work provides an annotated transcriptomic reference and identifies putative regulators of sucrose accumulation, while also describing the associated biological processes. Another objective is to examine the relationship between sugarcane phenotypic traits and expression levels of genes from classes such as sucrose-related enzymes, cell wall genes, and sugar transporters.

CHAPTER ONE: AN RNA-SEQ-BASED COMPARISON OF BIOLOGICAL REPLICATION STRATEGIES FOR DIFFERENTIAL GENE EXPRESSION IN *SACCHARUM*

ABSTRACT

One of the key procedures for ensuring statistical confidence in the results of differential gene expression studies is the use of biological replicates for comparisons among groups. Biological replicates allow the estimation of residual variation in the expression level among samples of a given experimental condition, for each expressed gene. For vegetatively propagated plants it is often possible to obtain an estimate of residual variability at two levels: among samples of distinct genotypes of the same experimental treatment, or among clonal replicates of the same genotype. However, the costs of sequencing are often a limitation to leveraging both these levels in the same study, stressing the relevance of efforts to determine an appropriate experimental design. Here we aim to investigate this question by comparing the transcriptional profiles of young sugarcane culms using strategies based on clones and on a diverse set of genotypes chosen to represent a common phenotypic group. The analyzed samples come from sugarcane genotypes differing in sugar accumulation. Our results show that the use of clonal replicates provided enough statistical power to identify nearly three times more differentially expressed genes than the more diverse strategy. However, the use of clones provided potentially less meaningful biological conclusions, because many of the significant genes were likely related to the particular genotype of choice, rather than representing a common expression profile for the groups we compared. We believe this study provides support for the development of sound experimental designs in new studies regarding differential expression for sugarcane.

Keywords: Transcriptome assembly, Gene ontology, Clonal replicates, Sugarcane

1 INTRODUCTION

The genus *Saccharum* comprises six species, of which *S. spontaneum* and *S. robustum* are the only wild representatives, spread over a large area in Asia and Indonesia, and the others are domesticated species — *S. officinarum*, *S. barberi*, *S. sinense*, and *S. edule*. Modern sugarcane cultivars are mainly descendant from the crossing of *S. officinarum* and *S. spontaneum*, followed by backcrossing to *S. officinarum*, such that they inherit the high sugar yield from the former species and the pathogen resistance, adaptability, and increased vigor of the latter (Irvine, 1999; Piperidis *et al.*, 2010). Sugarcane cultivation accounts for 86% of the worldwide production of sugar, despite the increasing allocation of its juice for ethanol production. Moreover, the sugarcane residue after juice extraction, called bagasse, is a byproduct that can be used for energy generation and production of bioplastics (OECD/FAO, 2019; Aguilar *et al.*, 2019). The crop is a renewable source of fuel and presents a significant advantage over fossil fuels due to the reduced emission of greenhouse gases (Goldemberg, 2008).

Sugarcane breeding programs usually rely on a few recurrent crosses between elite parents or wild germplasm to produce genotypes with desired traits, such as sugar or fiber yield and resistance to abiotic and biotic stresses (Heinz & Tew, 1987; de Souza Barbosa *et al.*, 2002; Jackson, 2005). As a consequence, scientific investigations in sugarcane are often based on elite lines as a source of plant material, such as two genome assemblies for the hybrids R570 — a major model in sugarcane genomic studies — and SP80-3280 (Garsmeur *et al.*, 2018; Souza *et al.*, 2019).

Also, these hybrids show a large variation in chromosome number and genome constitution. *S. officinarum* (2n = 8x = 80) and *S. spontaneum* (2n = 40-128), the parental species, have high levels of ploidy and complex genomes *per se* (Bremer, 1925; Panje & Babu, 1960). Chromosome number multiplicity and molecular evidence have led to the acceptance of the basic number of x = 8 for *S. spontaneum* (Liu *et al.*, 2016); however, the description of a wild accession with x = 10 brought a new panorama to the evolutionary history of the genus (Meng *et al.*, 2020). These facts reveal an intricate set of hurdles concerning the understanding of sugarcane genomics, which must be considered for data-driven experiments.

More specifically, the use of phenotypic trait variation between genotypes is a common approach found in differential expression studies. In the literature of sugarcane gene expression research, there are analyses conducted with a single genotype representing each group of interest (Casu *et al.*, 2007; Papini-Terzi *et al.*, 2009; Casu *et al.*, 2015; Vicentini *et al.*, 2015; Dharshini *et al.*, 2016), as well as with multiple genotypes per group (Papini-Terzi *et al.*, 2009; Ferreira *et al.*, 2016; Thirugnanasambandam *et al.*, 2017; Hoang *et al.*, 2017). Biological replicates provide more accurate estimates of transcript abundances when comparing samples from two treatment levels. Clones from the same genotype are subject to variability in their expression levels due to factors such as interactions with the environment and other organisms. Still, the transcriptional variation within clones is expected to be smaller when compared to plants from different genotypes, which increases the dispersion of gene quantification estimates. Statistical parameters such as means of expression levels and their residual variances are the main variables considered in modern differential expression tests, which highlights the relevance of the choice of approach for performing these studies. While the use of clones renders a more homogeneous set of samples, and consequently more statistical power to detect differences of expression between groups, it also restrains the set of samples to a limited number of genotypes.

Here, we evaluate the influence of using clonal replicates or multiple genotypes in differential gene expression analysis between contrasting groups. The comparison of approaches we propose relies both on quantitative estimates of differentially expressed genes and qualitative functional enrichment tests. We aim to present an information-based criterion for selecting biological replicates for further experiments using RNA-Seq, particularly for sugarcane, whose genomic properties can deviate dramatically

among genotypes.

2 MATERIALS AND METHODS

2.1 Biological material and RNA-Seq

The genotypes chosen for this study are part of the Brazilian Panel of Sugarcane Genotypes, located in Araras - Brazil (22.31602 S, 47.38929 W). They were selected from 266 genotypes to represent elite lines and commercial hybrids used in Brazil, as well as the parental species *S. officinarum* and *S. spontaneum*. First, for the strategy based on diverse genotypes (SBDG), we selected twelve distinct genotypes and separated them into four groups with three members each. This categorization divided genotypes based on their content of soluble solids, measured in o Brix: VLB (Very Low ºBrix), LB (Low ºBrix), HB (High ºBrix), VHB (Very High ºBrix). Next, for the strategy based on clones (SBC), we chose one representative of each of the four groups and used three clonal replicates of these genotypes to represent the corresponding phenotypic groups (Table 1).

Immature culms (internode +1) from all twenty-four plants were collected in June 2016, in Araras, under uniform experimental conditions, followed by extraction of total RNA with the RNeasy Plant Mini Kit (Qiagen) according to manufacturer's recommendations. We prepared the RNA-Seq libraries of polyadenylated transcripts using the TruSeq Stranded mRNA LT (Illumina) protocol. These libraries were sequenced in a HiSeq 2500 equipment (Illumina), resulting in paired-end reads 100 bp long. The twelve libraries of the SBC were sequenced in three lanes, in combination with other samples not used in this study, with final sequencing depth corresponding to eight samples per lane. For the SBDG we used a single lane exclusively for the twelve samples.

2.2 Downsampling and quality control

Because of the difference in sequencing depth between both datasets, we carried out a downsampling step for the SBC data, which showed higher average read counts per sample. This procedure aimed to ensure a balance of the amount of information in both strategies. For that, we applied the sample function of the Seqtk suite (`https://github.com/lh3/seqtk`), using as parameters a fixed random seed -s100 and the probability of removing a read proportional to the ratio of the average read counts of SBDG and SBC samples. After that, we used the programs Cutadapt v1.18 (Martin, 2011) and Trimmomatic v0.38 (Bolger *et al.*, 2014) to: *i)* trim residual sequences of Illumina adapters from raw reads; *ii)* remove base pairs with Phred score less than 20 in a window of 5bp; *iii)* trim the first 13bp of each read; and *iv)* remove paired reads shorter than 50 bp.

	VLB	**LB**	**HB**	**VHB**
SBC	IN84-58 R1	F36-819 R1	R570 R1	SP80-3280 R1
	IN84-58 R2	F36-819 R2	R570 R2	SP80-3280 R2
	IN84-58 R3	F36-819 R3	R570 R3	SP80-3280 R3
SBDG	IN84-58	F36-819	R570	SP80-3280
	SES205A	Criolla Rayada	White Transparent	White Mauritius
	Krakatau	IJ76-317	RB92579	RB835486

Table 1: Genotypes selected to compose each ºBrix group for the strategy based on clones (SBC) and based on diverse genotypes (SBDG). In the former strategy, we sampled the immature internode +1 of three clonal replicates (R1, R2, and R3) for each genotype per group, and samples from three different genotypes per group for the later. The genotypes IN84-58, F36-819, R570, and SP80-3280 were represented in both strategies, using samples from different plants.

2.3 *De novo* transcriptome assembly and functional annotation

We chose to perform a *de novo* transcriptome assembly based on all samples to be used as the reference for expression quantification, to minimize the potential effect of biases on genes and alleles represented. For that, we used the libraries after downsampling and quality control as input to Trinity v2.8.0 (Grabherr *et al.*, 2011), using the default parameters except by the normalization by readset. Functional annotation was carried out with blastx and blastp (Altschul *et al.*, 2010) significant hits (*e*-value < 1e-5) against the Swiss-Prot database, using ORFs identified in the transcriptome with Transdecoder (`https://github.com/TransDecoder/TransDecoder`). We also annotated protein domains using hmmscan v3.2.1 (Eddy, 2009) with the Pfam database. All these sources of information were compiled with the software Trinotate v3.1.1 (`https://github.com/Trinotate/Trinotate`) to produce the final annotation. This reference was further assessed by the identification of conserved orthologs among green plants and monocotyledons, using the software BUSCO v3 (Simão *et al.*, 2015) with databases in OrthoDB10.

Next, we used the *quasi-mapping* strategy of salmon v0.12.0 (Patro *et al.*, 2017) to quantify the expression of the assembled transcripts, separately for each sample. Quantification at the gene level used the sum of their corresponding transcripts counts weighted by each transcript length, because longer mRNA molecules tend to be sampled more often. The transcriptome file was used to build an index with a k-mer size of 31 bp, with the additional parameters of GC bias correction and validate mappings to achieve higher mapping rates and confidence levels.

2.4 Comparison of differential expression results with the full dataset

For differential expression analyses we initially excluded lowly expressed genes, by filtering out genes that did not show a count per million (CPM) greater than one for at least three samples. We did this filtering individually for each strategy, resulting in different sets of filtered genes for SBC and SBDG. Next, the following steps were repeated with the same criteria for both strategies, using the edgeR package (Robinson *et al.*, 2010). We normalized the gene counts with the trimmed mean of M-values method and built MDS plots using the top 2,000 genes with the greatest pairwise variation between samples.

For statistical tests of differential expression, we considered a model for gene counts parametrized as follows,

$$Y_{g,i} \sim NB(\mu_{g,i}, \Phi_g)$$

for sample i in a experimental group, gene g, $\pi_{g,i}$ the fraction of gene counts per gene and sample, dispersion ϕ_g, libraries size N_i, average counts $\mu_{g,i} = N_i\pi_{g,i}$, and variance $\Phi_g = \pi_{g,i}(1 + \pi_{g,i}\phi_g)$. The common dispersion is the squared Biological Coefficient of Variation (BCV), which takes into account the common dispersion from all genes. The use of a local regression of genewise dispersion provides an additional level of information for dispersion estimates for each gene. As a result, ϕ_g represents a compromise between the dispersion of counts for gene g and the borrowed genewise dispersion from genes with close average CPM.

We designed three orthogonal contrasts to test for differential expression for each gene, namely VLB x VHB.HB.LB, corresponding to the null hypothesis $H_0 : \pi_{g,VLB} = \frac{\pi_{g,VHB}+\pi_{g,HB}+\pi_{g,LB}}{3}$, VHB x HB.LB to $H_0 : \pi_{g,VHB} = \frac{\pi_{g,HB}+\pi_{g,LB}}{2}$, and HB x LB to $H_0 : \pi_{g,HB} = \pi_{g,LB}$. A likelihood ratio test was performed for each combination of gene, contrast, and strategy to identify the differentially expressed genes (DEGs), with *p*-values adjusted by the false discovery rate (FDR, Benjamini & Hochberg, 1995) at a 0.05 significance threshold.

Using the sets of DEGs and the annotated transcriptome, we performed functional enrichment analyses considering the frequency of Gene Ontology (GO) terms in the background reference and each set. Because the average gene length may vary among GO categories, care was taken to calculate effective gene lengths, based on the average length of genes in each sample weighted by their expression levels. We used the goseq package (Young *et al.*, 2010) to perform the functional enrichment test for each represented GO term ($p < 0.01$, after adjusting for multiple tests with the FDR approach).

2.5 Impact of missing samples on differential expression results

In addition to using all samples of each strategy, we also analyzed the effect of systematically removing samples on the differential expression results. This procedure can provide a better understanding of the effect of individual samples on the downstream analysis, as well as establishing a comparison between this approach and the use of full data. We have developed a methodology to compare different combinations of subsets of samples, under the condition that valid combinations must have at least two samples per group. This restriction is necessary because minimal replication per group is required to properly calculate gene dispersions, even if the estimates are less accurate. Because there are four groups with three samples each, 255 possible combinations exist, all of which were individually tested for differential expression with the same contrasts previously designed. The number of combinations of different numbers of removed samples is given by the binomial factor:

$$n_i = \binom{k}{i} g^i$$

in which k represents the number of samples per group ($k = 3$), g represents the number of groups ($g = 4$), and i represents the number of removed samples, ranging from one to four. For each combination, we removed genes with low expression levels (CPM > 1 in less than two samples) and recorded the differential expression result as one of the following categories: *(a)* upregulated, *(b)* downregulated, *(c)* not significant, or *(d)* filtered. One result was obtained for each gene, combination of samples, contrast and sampling strategy. We applied the same workflow for performing differential expression and functional enrichment tests as in the full data analyses.

Among all tested combinations of samples in our subsampling evaluation, one of special interest is that composed of genotypes present exclusively in SBDG. The strategy based on clones comprised a single genotype per group of soluble solids content, namely, IN84-58, F36-819, R570, and SP80-3280 (Table 1). For SBDG, we chose another eight genotypes in addition to these, which we call exclusive genotypes of SBDG, specifically SES205A, Krakatau, Criolla Rayada, IJ76-317, White Transparent, RB92579, White Mauritius, and RB835486. We also performed analyses of differential expression with this subset of samples.

3 RESULTS

3.1 Gene identification in the sugarcane transcriptome

Because our main objective was to compare the strategy based on clones (SBC) and the strategy based on diverse genotypes (SBDG), it was necessary to first establish a common reference for gene quantification. To that end, we performed a *de novo* transcriptome assembly using all 24 samples from both sampling strategies. The resulting transcriptome included 598,874 transcripts for a total of 262,281 assembled genes. Genes had an average size of 932.63 bp and the transcript N50 was 1,687 bp. The majority of genes had a single corresponding transcript isoform (64.3%). Our transcriptome assessment approach considered the representation of conserved single-copy orthologs from Viridiplantae and Liliopsida clades — green plants and monocotyledons, respectively. We identified 95.1% of the 430 orthologs conserved in green plants without sequence fragmentation. For the set of orthologs in monocots, 93.1% of 3,278 orthologs were fully represented. We performed a single transcript and gene quantification for all downstream analyses, using Salmon (Supplementary Table 1).

3.2 Comparison of differential expression results between strategies

Our goal in analyzing these datasets was to follow a standard pipeline of differential expression analysis with the R package edgeR, followed by functional enrichment tests with goseq. The quantification outputs, despite being ready for analysis after normalization by library size, still contain a large amount of lowly expressed genes. We selected genes with CPM > 1 in at least three samples, for each strategy separately, resulting in different sets of kept genes for SBC and SBDG. The former presented 42,566 genes after filtering, and 41,934 remained in the latter.

An initial exploratory investigation allowed assessing the main characteristics of expression profiles with a multidimensional scaling plot (Figure 1). For the SBC, we observed a clear clustering of replicates from each genotype, indicating high similarity in the expression profiles of clonal replicates. As expected, the first dimension of the plot separated replicates of genotype IN84-58 from the remaining three groups, reflecting their contrasting genetic backgrounds. On the other hand, we noted that the biological variance of gene expression was much higher in the diverse approach than in the clone approach. In the SBDG we found little overlap of samples from the same phenotypic group, with the exception of the VLB genotypes, which again were isolated from the others by differences in the first component. In this case, the distances between samples from the same category made it hard to find clear clustering patterns for the other groups. We highlight the separation of two *S. officinarum* accessions, Criolla Rayada and IJ76-317, while the other two (White Transparent and White Mauritius) were located among the hybrids.

The MDS analysis provided a broad view of the overall patterns of transcription abundances for the set of samples, but does not allow a closer assessment of individual genes. We then used the differential expression testing approach for a detailed investigation of the transcriptome expression profiles. We arranged the four groups of samples into three orthogonal contrasts. Hence, we conducted three tests of differential expression for each gene. Mean-difference plots display the relationship between log fold change for each contrast and the average log CPM, as well as the differential expression test results per gene (Figure 2). The quantity of DEGs identified via the SBC largely surpassed that of SBDG for all contrasts, especially in VHB x HB.LB and HB x LB. In these two contrasts, we can observe a mass of significant DEGs beginning in relatively low absolute logFC values for SBC. Conversely, only a few DEGs were significant for SBDG, even for genes showing fold changes of large magnitude. We identified non-significant genes even at $|\text{logFC}| > 10$, standing for more than a thousand-fold variation of read counts. This is possible because of the characteristics of the adopted likelihood model for gene abundance, which considers gene counts and variance within groups for the likelihood ratio test. We can

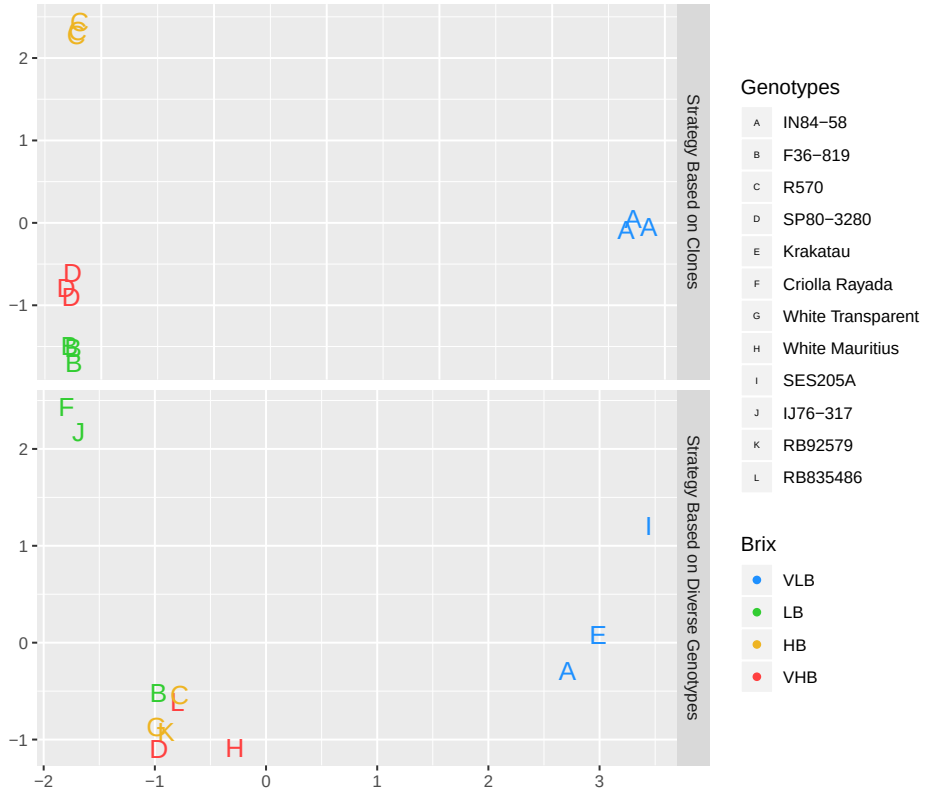


Figure 1: Multidimensional scaling (MDS) plot showing pairwise distances between samples based on the most divergent genes for each pair. The panels represent the MDS plot for the strategy based on clones (SBC) and the strategy based on diverse genotypes (SBDG), respectively.

thus (at least partly) attribute the lower number of DEGs for the SBDG to the higher residual variance in gene counts observed with this strategy. An indicator of dispersion with a meaningful interpretation is the Biological Coefficient of Variation (BCV), calculated as the square root of the negative binomial dispersion of counts. The average BCV for all filtered genes of the SBC was 0.087, and 0.440 for the SBDG, representing a five-fold variation between strategies. In addition, for the set of genes in common retained after filtering for both strategies (37,535 genes), 98% of them showed higher BCV in the SBDG. These numbers reinforce the role of dispersion as a key parameter that distinguishes the approaches regarding differential expression.

The intersection of sets of DEGs between strategies revealed that the majority of genes identified as significant in the SBDG was also significant in the SBC, but the inverse was not true (Figure 3). About 71% of DEGs detected with the SBDG were shared with the other strategy, for each of the three contrasts. This fact suggests that the use of more diverse genotypes favored the identification of genes with similar expression patterns among the group members. The observation regarding the high residual variance for VHB x HB.LB and HB.LB also strengthens this hypothesis, because only the more homogeneously expressed genes achieved significance. On the other hand, in addition to most of the DEGs in common with SBDG, the use of clones was also able to identify many other genes as differentially expressed, which are possibly genotype-specific and may not be directly associated with the phenotype of interest.

The current work presents a systematic analysis of the effects of competing strategies of biological replication over gene expression studies. Our goal is not to provide a biological interpretation of expression patterns, but to justify with biological reasoning the use of each methodology. Therefore, we chose the functional enrichment analysis as a meaningful approach for understanding the consequences of data-mining over the sets of filtered and differentially expressed genes. Within each set of genes that

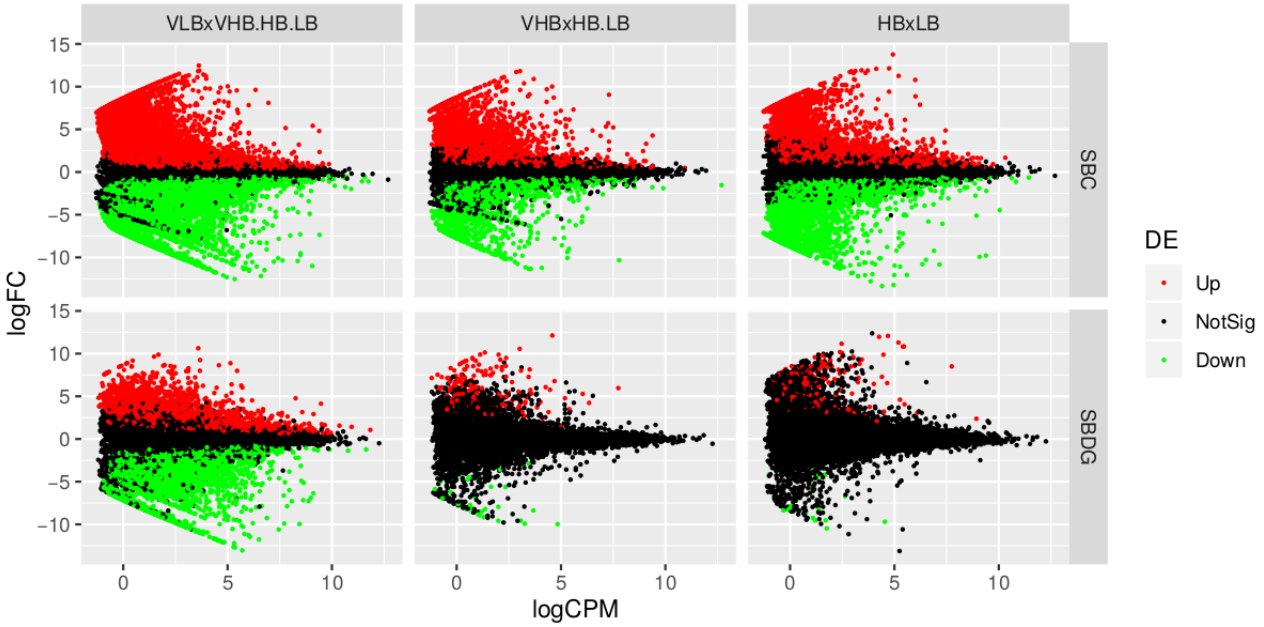
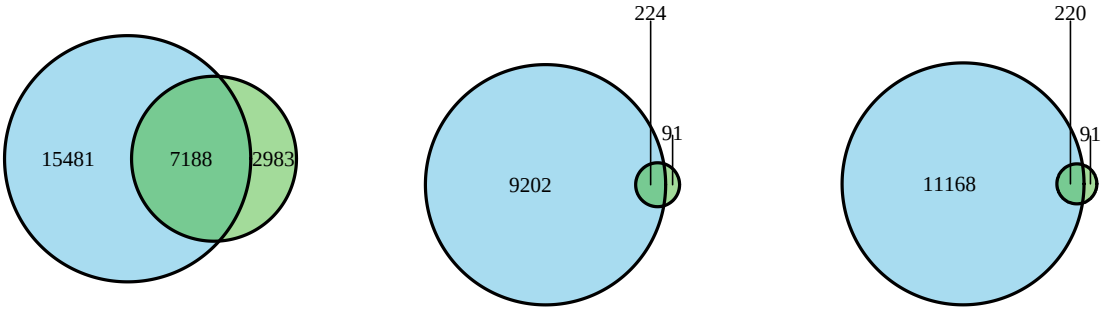


Figure 2: Mean-difference plot grid showing differentially expressed genes for all contrasts and strategies. Fold changes (logFC) and average expression levels in counts per million (logCPM) are shown in base 2 log scales. The columns indicate the three orthogonal contrasts, while the lines correspond to the strategies of sampling biological replicates. Colors represent the result of differential expression tests ($p < 0.05$, after FDR correction for multiple tests).

passed the expression filter and were used for further investigation, we found 12,364 and 11,979 genes containing at least one attributed GO term for SBC and SBDG, respectively. Using these genes as a background reference, we performed a functional enrichment analysis to identify GO terms more frequent among DEGs than expected by chance alone (Figure 4). For the SBDG, the contrast VHB x HB.LB resulted in only one enriched term (adenosine diphosphate binding), and HB x LB had no enriched GO.

3.3 Assessment of strategies using subsets of samples

When removing a fraction of samples from the experimental design, the average values of gene counts and variance are modified and less precise, such that the resulting set of DEGs may be different. For this reason, we adopted the strategy of systematically removing samples as a validating procedure of



(a) VLB x VHB.HB.LB (b) VHB x HB.LB (c) HB x LB

Figure 3: Differentially expressed genes shared by the strategies based on clones and on diverse genotypes. The strategy based on clones is represented in blue, while the strategy based on diverse genotypes is indicated in green. The diagrams represent the number of genes detected as significantly differentially expressed in the contrasts.

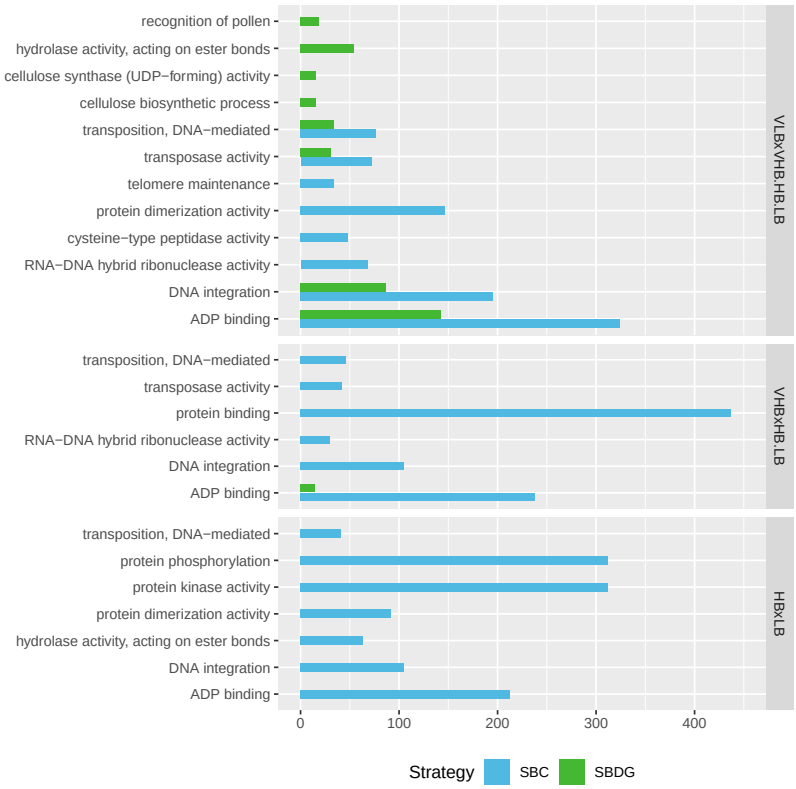


Figure 4: Enriched gene ontology terms by strategy and contrast. There was no significant test result for HB x LB in the strategy based on diverse genotypes ($p < 0.01$, after FDR adjustment). The numeric axis represents the number of differentially expressed genes for each particular gene ontology term.

the expression results. Also, this approach evaluates the effect of variation on the number of samples per group, such as in unbalanced experiments, and missing samples. Mistakes while handling the samples, low volumes of biologic material, difficulties in preparing the sequencing libraries, and other unexpected events often cause (random) loss of samples. Because each strategy includes twelve samples divided into four groups, and there must be at least two samples per group for estimating the dispersion parameter, we could jointly remove a maximum of four samples. These restrictions produced 255 combinations of samples, which were individually tested for differential expression.

We observed that as the number of removed samples grew from one to four, the more the results of differential expression disagreed with the results obtained with the full set of samples. Albeit at low rates, we could identify genes with an inverted result of differential expression, *i.e.*, miscalls of up or downregulation, which occurred from 10^{-6} to 10^{-5}% of genes for the SBC, and from 10^{-5} to 10^{-4}% for the SBDG. Using the original data results as a gold standard (full set of samples), the strategy based on clones showed a relatively lower percentage of false negatives and a higher percentage of false positives — green and purple curves in (Figure 5), respectively.

Because our systematic removal of samples provided a large number of differential expression tests for each gene, we could establish a high confidence set of DEGs — those with at least 95% of tests with the same results (Table 2). We then used this high confidence set for performing a functional enrichment analysis (Figure 6). The enriched GO terms for the full set of DEGs and the high confidence set were essentially different, once there were only three enriched terms for the SBDG, of which two had also been detected with the full dataset, and the other was only significant for the SBC. Given the low number of annotated and differentially expressed genes for the contrasts VHB x HB.LB and HB x LB, it was not possible to detect any enriched term for the SBDG. Analyzing exclusively the SBC, nearly

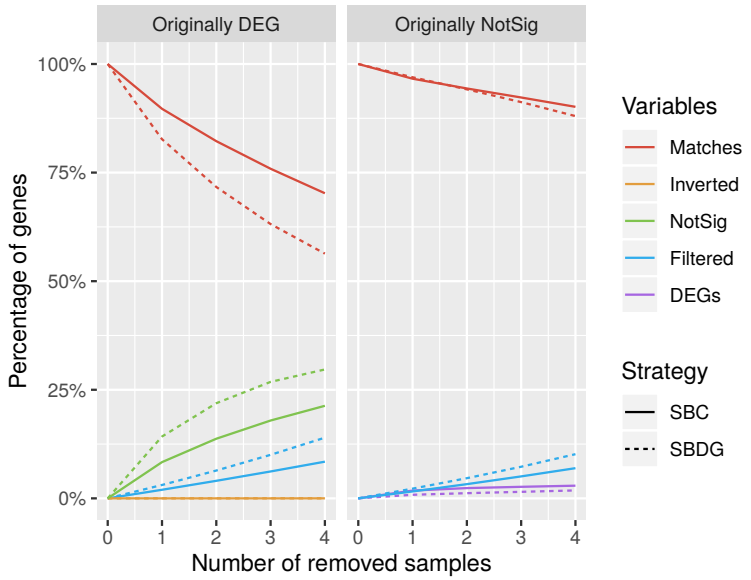


Figure 5: Effect of sample removal on the results of differential expression. The values represented by the continuous (SBC) and dashed (SBDG) curves are the averages of differential expression tests for all possible combinations and contrasts, as a function of the number of removed samples. The red curve indicates the concordant genes in the original and subsampled datasets; in yellow, the differentially expressed genes with inverted results, whether up or downregulated; in green, genes that were not significant due to subsampling; in blue, filtered genes after subsampling; and in purple, genes that appeared as spuriously differentially expressed with subsampling.

Contrast	**SBC**		**SBDG**	
	Total DEGs	**Annotated DEGs**	**Total DEGs**	**Annotated DEGs**
VLB x VHB.HB.LB	14240	2688	2960	458
VHB x HB.LB	5774	829	44	7
HB x LB	5371	994	34	7

Table 2: The differentially expressed genes (DEGs) in the high confidence set. We identified these genes as differentially expressed in at least 95% of the subsampling combinations when removing from one to four samples in each strategy. We considered only the combinations which presented a minimum of two samples per experimental group. For each strategy, we show the total number of DEGs and those annotated with gene ontology terms.

73% of the terms were also enriched in the full dataset for VLB x VHB.HB.LB, 50% for VHB x HB.LB and 75% for HB x LB. Also, the number of enriched terms was high, even with fewer DEGs for the test. Some of the terms were exclusive for the high confidence set, such as zinc ion binding, proteolysis, and negative regulation of translation. The opposite also occurred, such as for kinase activity.

3.4 Contribution of SBDG exclusive genotypes for differential expression

We compared the DEGs identified in the subgroup of genotypes absent in SBC with the data from SBC and SBDG, using the same parameters for the analysis (Figure 7). Here, it was possible to observe distinct patterns between the contrast VLB x VHB.HB.LB and the others, regarding the number of DEGs called by each approach. In the first contrast, the total number for SBDG was greater than for the exclusive set, in opposition to the results for the last two contrasts. We also highlight in these comparisons that the larger fraction of DEGs detected in SBDG was concentrated in the intersection with the other approaches.

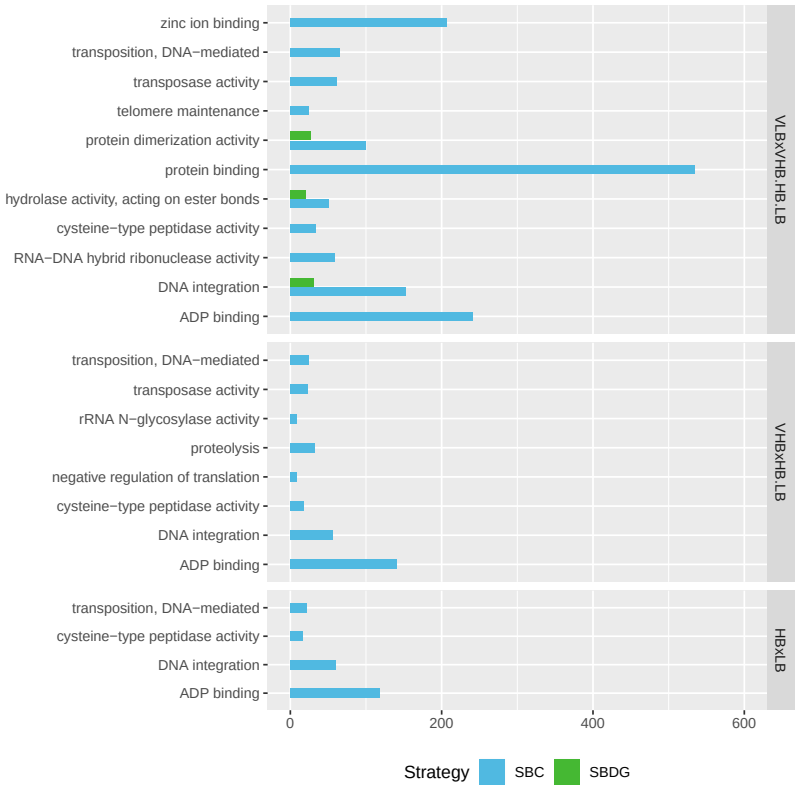
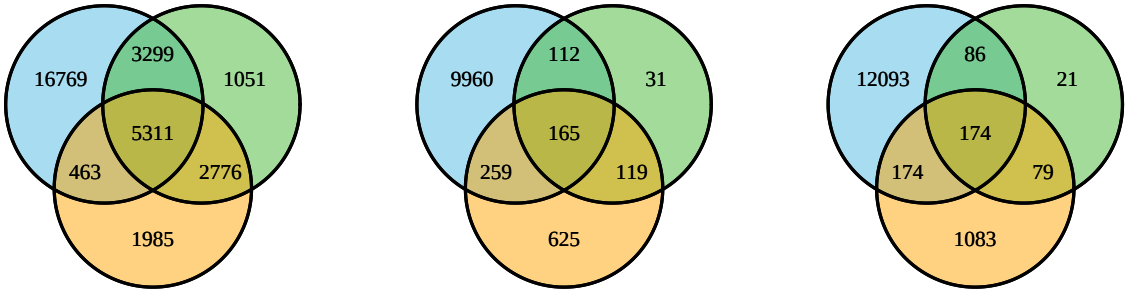


Figure 6: Enriched GO terms by strategy and contrast for the high confidence set of DEGs. This set contains genes with a significant test for differential expression in more than 95% of combinations of samples ($p < 0.01$, after FDR adjustment).



(a) VLB x VHB.HB.LB (b) VHB x HB.LB (c) HB x LB

Figure 7: Differentially expressed genes shared by three sets of samples: SBC, SBDG, and SBDG-exclusive genotypes. The method for filtering genes with low expression was the same for the three sets, adopting a more permissive criterion due to the lower number of samples in the exclusive set (CPM > 1 for at least two samples).

4 DISCUSSION

Analyzing the patterns found in the MDS plots, which were based on the most extreme genes in terms of differential expression, we can infer that transcriptional profiles agreed only partially with the phenotypic assortment of genotypes into four categories of soluble solids content (Figure 1). This plot also shows a recurrent observation in the other analyses regarding the sharp disparities found between VLB and the other groups. This fact is evident in the separation of samples in (Figure 1), the increased number of DEGs from the VLB x VHB.HB.LB tests, when compared to the other contrasts (Figure 3), and the functional enrichment from SBDG (Figure 4). A straightforward likely explanation is the genetic background of the genotypes, because VLB comprises *S. spontaneum* accessions, while VHB, HB, and LB comprise *S. officinarum* and commercial hybrids. Despite having a genomic contribution from both parental species, commercial hybrids underwent backcrossing to *S. officinarum* in order to enhance sugar yield, which makes them more alike to this species in terms of expression. This conclusion also agrees with cytogenetic information from R570, because about 80% of its chromosomes presented similarity to *S. officinarum* and 10% to *S. spontaneum*, using probes with fluorescent in situ hybridization (D'Hont *et al.*, 1996).

SBDG yielded a result which agrees with this assumption by providing a low number of DEGs for contrasts VHB x HB.LB and HB x LB, when compared to SBC. When considering the fraction of significant DEGs in common between strategies using the full dataset, the amount of shared significant tests was nearly constant over the three contrasts in SBDG (Figure 3). The same was not true for SBC, which had a rate of shared DEGs ranging from 2 to 32%. A feasible explanation is that sugarcane genotypes have high variability of expression among each other, and the use of clones provides enough statistical power to detect it. However, a substantial proportion of these genes might not be actually related to the biological phenomenon of interest, because the lower variability in the SBC led to the identification of DEGs with lower fold-change magnitudes. Extrapolating these results, we can suggest that the SBC was not fully representative of the groups of interest, because of the low agreement of DEGs identified in common with the SBDG.

Several of the identified enriched GO terms fit in molecular mechanisms with no explicit relationship to the accumulation of sugars or carbon partition. For instance, the contrast HB x LB in SBC, which represents a direct comparison of the genotypes R570 and F36-819, showed a significant enrichment of kinase activity, which may indeed represent an important mechanism that distinguishes the phenotypes of these plants. However, phosphorylation is a broad molecular mechanism of signal transduction, and it could be related to other processes other than sugar accumulation. Besides, this term was not significant for the same contrast in SBDG, showing that the expression patterns of genes associated with protein phosphorylation were not consistent among the other genotypes. On the other hand, some of the terms found in the SBDG functional enrichment are coherent with observable phenotypic traits, *e.g.*, recognition of pollen and cellulose biosynthetic process. The recognition of pollen is a potentially vital activity for genotypes in the VLB group, because it is composed uniquely of wild plants, which probably depend on sexual reproduction. Also, the discrepant levels of fiber in VLB x VHB.HB.LB groups corroborate the enrichment of cellulose biosynthetic activity. With the outcomes of functional enrichment for the high confidence set, we could recognize several GO terms in disagreement with the DEGs based on the full set. The terms discussed above such as pollen recognition, cellulose biosynthetic process, and kinase activity were not significant for these high confidence genes. These examples highlight the lack of similar expression patterns among all samples. Besides the biological and residual sources of variation in gene expression quantitation, stochastic processes also have a contribution to the variance of RNA-Seq data, such as the random sampling of transcripts in library preparation. For SBDG, we could also consider that the genotypes in each group have different contributions to the differential expression result. More

precisely, combining a diverse set of genotypes into an experimental group increases the overall variability of expression levels for most of the genes, and modifies the average counts per group.

We presented a selection of four genotypes for the SBC, which is one particular choice among 81 (3^4) possible combinations if maintaining the same categories from the SBDG. Examining the wide distribution of genotypes in the MDS plot for SBDG (Figure 1), we can presume that the choice of genotypes can lead to sharply discordant sets of DEGs. This is a result of the faulty coherence of genotypes inside the groups VHB, HB, and LB. Furthermore, the combination-sensitive set of identified DEGs could drive mistaken conclusions regarding the biological issue of interest. For example, a specific gene might be called as differentially expressed due exclusively to the choice of sampled genotypes, instead of representing a general phenomenon for other genotypes with similar phenotypic characteristics. The outcomes of the analyses using subsets of samples reinforce this hypothesis (Figure 5). There we can observe an increasing number of genes with contradictory results of differential expression tests when compared to the full-data tests. This fact implies that simply including or not some genotypes may lead to changes in the list of DEGs. Another result that supports the reasoning about the caveats on genotype choice is the number of DEGs for VHB x HB.LB and HB x LB contrasts (Figure 7). In the former contrast, the exclusive genotypes of the SBDG showed 625 DEGs that could only be found with these samples, versus 31 in SBDG. The difference was even more prominent for HB x LB. Notably, Criolla Rayada and IJ76-317 are *S. officinarum* accessions that integrate the LB group, both with a discrepant expression profile according to the MDS analysis. The simple inclusion of F36-819 in this group might have been enough to disrupt the homogeneity detected between the other two genotypes. These observations show how the lack of uniformity in the SBDG genotypes leads to a low number of significant DEGs. Moreover, they indicate that this uniformity may be sensible to the choice of genotypes to form the experimental groups.

The behavior of (mis)matches in the detection of DEGs (Supplementary Figures 1 and 2) can be helpful to illustrate some properties of each strategy. First, they illustrate the more robust response of the SBC regarding the removal of samples by the lower rate of mismatches. This fact reinforces that SBC showed increased statistical power to detect DEGs. Second, these results also suggest that individual samples can have a determinant role on the identification of differential expression for a considerable number of genes, mainly for the SBDG. As shown in (Supplementary Figure 2), the $n_i \times n_i$ grids did not reveal a uniform or linear distribution pattern of the power to detect differential expression. There were both rows and columns densely occupied by DEGs, in patterns contingent on the number of removed samples. They occurred in multiples of 25% for three samples and 33% for four, which correspond to the fractions of combinations without a specific sample. (Supplementary Figure 1) revealed a similar pattern, noticeable by the steep inclines of cumulative distributions for particular mismatch rates.

We argued that the SBDG yielded fewer DEGs as a consequence of combining genotypes with more variable expression patterns than the SBC. Also, our interpretation of results presented evidence towards the prevalence of more biologically meaningful DEGs for SBDG, instead of simply revealing genotype-specific profiles. However, a feasible criticism over these hypotheses is that the use of a collection of genotypes per phenotypic group could still lead to genotype-specific DEGs, but for more than one genotype at once. A necessary step to avoid this issue is to choose a diverse set of genotypes for the experimental groups, which should be unrelated and representative of the population of interest. For tackling this question, we performed the complete analysis pipeline using the set of genotypes exclusive to the SBDG, because we can then assess the direct contribution of the genotypes shared with the SBC, the genotypes absent in SBC, and the intersection between them. Interestingly, this analysis showed that the intersection between SBC and the exclusive set concentrated most of the SBDG genes in all contrasts (from 37 to 48% of DEGs), which agrees with the expectation of a shared set of genes among all twelve genotypes. Moreover, the correlations of logFC in VLB x VHB.HB.LB among the approaches revealed

that SBDG had an intermediate pattern for differential expression between SBC and the exclusive set of samples (Supplementary Figure 3). Another important observation is that the increased number of samples for SBDG compared to the exclusive set of samples led to a larger number of DEGs in the first contrast and a smaller number in the other two. Thus, we can hypothesize that as the number of genotypes per group increased, the issue of detecting genotype-specific DEGs and genes with reduced biological meaning decreased.

5 CONCLUSIONS

With the increasing application of next generation sequencing to investigate complex transcriptomes, such as that of sugarcane, recent studies aim to apply these techniques to unravel the molecular mechanisms controlling several phenotypic traits. However, a single biological replicate in each contrasting group is not enough for performing this sort of analysis, leaving for the researcher the choice of a suitable experimental design. Our present study intended to illustrate the strengths and caveats inherent to two sampling strategies for biological replication, namely by using a diverse group of genotypes with common phenotypic characteristics or clones from the same genotype, chosen to be representative of this group. The results have provided evidence of discrepancies in *(i)* quantitative terms, regarding the number of genes detected as differentially expressed, *(ii)* consistency, when subjected to self-validation using subsampling, and *(iii)* inferred biological conclusions from the functional annotation of differentially expressed genes. These analyses suggest that the use of clones as biological replicates may yield somewhat restricted results, biased by the particular choice of genotypes. Regardless of these concerns, the direct comparison of two genotypes can still be useful in particular situations. On the other hand, the presence of a representative set of genotypes within the same group can lead to more reasonable biologic outcomes. In any case, it is possible to combine these strategies to refine the level of details, if economically viable. This research offers support to the experimental design of new studies using differential expression as a method of investigation in sugarcane and other plants with high genomic complexity.

6 ABBREVIATIONS

BCV - Biological coefficient of variation, CPM - Counts per million; DEG - Differentially expressed gene; GO - Gene ontology; HB - High ºBrix; LB - Low ºBrix; MDS - Multidimensional scaling; SBC - Strategy based on clones; SBDG - Strategy based on diverse genotypes; VHB - Very High ºBrix; VLB - Very Low ºBrix.

REFERENCES

Aguilar, N. M., Arteaga-Cardona, F., de Anda Reyes, M. E., Gervacio-Arciniega, J. J., & Salazar-Kuri, U. (2019). Magnetic bioplastics based on isolated cellulose from cotton and sugarcane bagasse. *Materials Chemistry and Physics, 238*, 121921.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology, 215*(3), 403-410.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Bremer, G. (1925). The cytology of the sugarcane. *Genetica, 7*(3), 293-322.

Casu, R. E., Jarmey, J. M., Bonnett, G. D., & Manners, J. M. (2007). Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling. *Functional & integrative genomics, 7*(2), 153-167.

Casu, R. E., Rae, A. L., Nielsen, J. M., Perroux, J. M., Bonnett, G. D., & Manners, J. M. (2015). Tissue-specific transcriptome analysis within the maturing sugarcane stalk reveals spatial regulation in the expression of cellulose synthase and sucrose transporter gene families. *Plant molecular biology, 89*(6), 607-628.

de Souza Barbosa, G. V., de Menezes Cruz, M., Soares, L., Rocha, A. M. C., Ribeiro, C. A. G., Sousa, A. J. R., ... & dos Santos, A. V. P. (2002). A brief report on sugarcane breeding program in Alagoas, Brazil. *Crop Breeding and Applied Biotechnology, 2*(4).

Dharshini, S., Chakravarthi, M., Manoj, V. M., Naveenarani, M., Kumar, R., Meena, M., ... & Appunu, C. (2016). De novo sequencing and transcriptome analysis of a low temperature tolerant Saccharum spontaneum clone IND 00-1037. *Journal of biotechnology, 231*, 280-294.

D'Hont, A., Grivet, L., Feldmann, P., Glaszmann, J. C., Rao, S., & Berding, N. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Molecular and General Genetics MGG, 250*(4), 405-413.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Informatics 2009: Genome Informatics Series Vol. 23* (pp. 205-211).

Ferreira, S. S., Hotta, C. T., de Carli Poelking, V. G., Leite, D. C. C., Buckeridge, M. S., Loureiro, M. E., ... & Souza, G. M. (2016). Co-expression network analysis reveals transcription factors associated to cell wall biosynthesis in sugarcane. *Plant molecular biology, 91*(1-2), 15-35.

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., ... & Costet, L. (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature communications, 9*(1), 1-10.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology, 29*(7), 644.

Goldemberg, J. (2008). The Brazilian biofuels industry. *Biotechnology for biofuels, 1*(1), 6.

Heinz, D. J., & Tew, T. L. (1987). Hybridization procedures. In *Developments in crop science* (Vol. 11, pp. 313-342). Elsevier.

Hoang, N. V., Furtado, A., O'Keeffe, A. J., Botha, F. C., & Henry, R. J. (2017). Association of gene expression with biomass content and composition in sugarcane. *PLoS One, 12*(8).

Irvine, J. E. (1999). Saccharum species as horticultural classes. *Theoretical and Applied Genetics, 98*(2), 186-194.

Jackson, P. A. (2005). Breeding for improved sugar content in sugarcane. *Field Crops Research, 92*(2-3), 277-290.

Liu, X., Li, X., Liu, H., Xu, C., Lin, X., Li, C., & Deng, Z. (2016). Phylogenetic analysis of different ploidy Saccharum spontaneum based on rDNA-ITS sequences. *PloS one, 11*(3).

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal, 17*(1), 10-12.

Meng, Z., Han, J., Lin, Y., Zhao, Y., Lin, Q., Ma, X., ... & Wang, K. (2020). Characterization of a Saccharum spontaneum with a basic chromosome number of x= 10 provides new insights on genome evolution in genus Saccharum. *Theoretical and Applied Genetics, 133*(1), 187-199.

Panje, R. R., & Babu, C. N. (1960). Studies in Saccharum spontaneum distribution and geographical association of chromosome numbers. *Cytologia, 25*(2), 152-172.

Papini-Terzi, F. S., Rocha, F. R., Vêncio, R. Z., Felix, J. M., Branco, D. S., Waclawovsky, A. J., ... & Vicentini, R. (2009). Sugarcane genes associated with sucrose content. *BMC genomics, 10*(1), 120.

Piperidis, G., Piperidis, N., & D'Hont, A. (2010). Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. *Molecular Genetics and Genomics, 284*(1), 65-73.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212.

Souza, G. M., Van Sluys, M. A., Lembke, C. G., Lee, H., Margarido, G. R. A., Hotta, C. T., ... & Nishiyama Jr, M. Y. (2019). Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience, 8*(12), giz129.

Thirugnanasambandam, P. P., Hoang, N. V., Furtado, A., Botha, F. C., & Henry, R. J. (2017). Association of variation in the sugarcane transcriptome with sugar content. *BMC genomics, 18*(1), 909.

Vicentini, R., Bottcher, A., dos Santos Brito, M., dos Santos, A. B., Creste, S., de Andrade Landell, M. G., ... & Mazzafera, P. (2015). Large-scale transcriptome analysis of two sugarcane genotypes contrasting for lignin content. *PloS one, 10*(8).

CHAPTER TWO: IDENTIFYING EXPRESSION PROFILES ASSOCIATED WITH DIVERSE CARBON PARTITIONING PHENOTYPES IN YOUNG SUGARCANE STEMS

ABSTRACT

Sugarcane is a crop with well-known genomic complexity among plant breeders and geneticists, due to its auto and allopolyploidy, and hybridization events among *Saccharum* species. The intricate genome poses problems going from difficulties in improving yield via classic plant breeding programs to a lack of molecular information regarding its sugar accumulation mechanisms. Despite advances such as the first genome assemblies of a *S. spontaneum* accession and two hybrids, carbon partitioning in sugarcane is still poorly understood. More specifically, there are no genetic studies describing the regulation of carbon partitioning in an early stage of development, in the first culm below the apical meristem. Here we present a *de novo* transcriptome assembly comprised of expression profiles of apical culms from a broad range of genotypes contrasting in sugar content. We were able to quantify the expression of currently putative sucrose content regulators such as sucrose synthases, invertases, and sucrose phosphate synthases. Our data point to cellulose synthesis as a major process, which segregates low sugar and high sugar individuals, and how it communicates with sugar processing and transport in the storage parenchyma tissue. We also found *in silico* evidence for a novel splicing pattern for the sugar transporters SWEET2b and SWEET16 being expressed in all investigated genotypes, resulting in a protein with an unreported number of transmembrane helices and a new signal peptide. This work contributes to understanding the underlying mechanisms responsible for phenotypical variation among *Saccharum spontaneum*, *S. officinarum*, and hybrids in traits such as sugar and fiber yield.

Keywords: Immature culm, SWEET transporter, Sucrose synthase, Cellulose synthase, RNA-Seq

1 INTRODUCTION

Sugarcane cultivation has a large contribution to Brazilian agribusiness in terms of production, economic importance, and cultivated area. Its relevance is mainly due to the use of culms to produce sugar and ethanol. Brazil has over 8 million ha cultivated with sugarcane, or roughly 1% of the total land area, with an estimated production of 642.7 million tons in 2019/2020 (`https://www.conab.gov.br/`). It is thus the major worldwide producer of this crop. Using vegetable biomass from sugarcane represents an alternative source of biofuel generation, which has a lower environmental impact than fossil fuels, due to carbon sequestration. Therefore, concerns involving environmental pollution and long-term fuel reserves are responsible for the interest in using fuels from renewable energy sources. This fact drives genetic breeding of this crop aiming to increase production with phenotypes of agronomic and economic interest, such as drought resistance, sugar yield, biomass content, high tillering, and ratooning (Rooney *et al.*, 2007; Morris *et al.*, 2013).

The complexity of the sugarcane genome is one of the main challenges for performing computational analyses based on molecular information. Panje & Babu (1960) performed an extensive manual characterization of the number of chromosomes in more than three hundred *S. spontaneum* accessions, concluding that the diploid number varies from 40 to 128 for this species. For *S. officinarum*, the karyotype 2n = 80 is found among most of the accessions since the first evaluation by Bremer (1925), with the exception of aneuploid variation from atypical clones (Heinz, 1991). The previous interspecific hybridization events between *S. spontaneum* and *S. officinarum* aimed to create crops with high sugar yield in the culms and resistance to pathogens. From a genomic standpoint, these hybrids resulted in polyploid and aneuploid crops with 2n = 100-130, with chromosomes from both parental species as well as recombinants (D'Hont *et al.*, 1996). High ploidy levels increase the number of gene copies, which can lead to the emergence of new alleles by sequence divergence, pseudogenes by loss of function, and neofunctionalization in exceptional cases. Aneuploidy is also a determinant factor on phenotypic variation, because genic dosage ratios are affected for genes in different chromosomes (Makarevitch & Harris, 2010).

These obstacles have hindered advances into understanding the roles of key genes related to physiological characteristics of sugarcane, such as sucrose accumulation, one of its most important agronomic traits. The molecular mechanisms of sucrose concentration in mature sugarcane culms depend on sucrose balance and availability, passing through several biological processes, such as photosynthesis, cellular growth, respiration, and sugar transport. Sugar accumulation in sugarcane works differently from most of the plants, because its storage carbohydrate is sucrose, instead of complex and insoluble polysaccharides, such as starch, and it occurs in stem parenchymal tissue (McCormick *et al.*, 2008 b). Sacher *et al.* (1963) described that sucrose undergoes conversion to hexoses by acid invertases acting at the apoplast of parenchyma cells in young culms. By doing so, these enzymes can control the uptake of sugars for posterior degradation or resynthesis for accumulation. Moore (1995) evaluated the metabolite dynamics through each compartment — apoplast, cytosol, and vacuole of storage parenchyma cells in culms. In his work, the main hypothesis is that the maintenance of low turgor gradient of solutes could regulate the transport of metabolites, mediated by a turgor-sensing system, which has been endorsed by Wang *et al.* (2013). However, Wu *et al.* (2007) were able to develop transgenic sugarcane carrying a gene for sucrose isomerase (SI) that can double its sugar content by converting sucrose to isomaltulose, with no decrease in sucrose concentration and increased photosynthetic rates. In this experiment, the authors showed that signaling pathways in the source-sink system can regulate photosynthetic activity on leaves and that osmotic restraints are not the main limiters to the sucrose accumulation process.

Studies indicate invertases as key regulators of sugar levels in sugarcane culms (Wang *et al.*, 2013). Three invertases synthesized in storage parenchyma — cell wall bound invertase (CWI), neutral invertase (NI), and soluble acid invertase (SAI) — are redirected to different cell compartments

and prevail in different stages of culm maturation (Wang *et al.*, 2013). Alongside with the invertases, sucrose synthase (SuSy) and sucrose phosphate synthase (SPS) have been identified as putative markers of sucrose accumulation. SuSy catalyzes a reversible reaction for cleaving sucrose into fructose and UDP-glucose. The breakdown/synthesis ratio of SuSy increases with internode maturity, with activity in young internodes mainly in the synthetic direction (Schäfer *et al.*, 2004). Lower sucrose concentration in immature culms also corroborates these results, supporting sucrose synthesis instead of breakdown by SuSy. Overall sucrose synthesis occurs by SuSy and SPS catalysis (Botha & Black, 2000). However, other factors might be related to sucrose levels in sugarcane, such as redirecting of UDP-glucose to cell wall synthesis by cellulose synthase A (CesA) complexes and intricate regulation of sugar trafficking through the phloem, apoplast, storage parenchyma cytoplasm, and vacuole, via apoplasmic or symplasmic paths (Casu *et al.*, 2015).

RNA-Seq and EST sequencing projects have been performed in sugarcane to understand genes in terms of their tissue-specificity, to determine differences between mature and immature culms, and to detect genes affected by abiotic stresses. Several of these studies also combined sequencing and hybridization approaches to identify genes of interest and estimate the abundance of their corresponding transcripts (de Araujo *et al.*, 2005; Rae *et al.*, 2005; Papini-Terzi *et al.*, 2009; Casu *et al.*, 2015; Thirugnanasambandam *et al.*, 2019). Previous studies, using sugar-contrasting genotypes at distinct levels of stem maturity, reported several classes of genes as differentially expressed. Genes related to signaling such as kinases, phosphatases, and transcription factors, and to cell wall biosynthesis, as well as SuSy, SPS, and bidirectional sugar transporters (SWEETs) were either up or downregulated in high sugar genotypes (Papini-Terzi *et al.*, 2009; Thirugnanasambandam *et al.*, 2017). Moreover, genes whose expression was affected by abiotic stresses, such as drought, were also related to sugar accumulation processes in the case of abscisic acid (ABA) signaling and biosynthesis, for example (Papini-Terzi *et al.*, 2009). This observation may be due to the fact that sugarcane shows increased sugar levels in response to abiotic stresses. Genotypes contrasting in lignin content also exhibited a consistent differential expression pattern of genes in the phenylpropanoid biosynthetic pathway and cell wall proteins (Vicentini *et al.*, 2015). Regardless of efforts to identify genes associated with sucrose accumulation in sugarcane culms, little is known about how carbon partitioning takes place in the apical section of its stalks.

In this work, we investigate gene expression in internode +1 of different sugarcane genotypes contrasting in sugar levels, representing a wide variety of phenotypes and origins of *Sachharum* accessions. Because this section of the culm is directly below the apical meristem, the sugar accumulation process has not yet taken place. The main biological activity in this section of the stem is the expanding of cell wall surface for cellular growth (Rose & Botha, 2000). We used RNA-seq data to analyze expressed transcripts in sugarcane to understand the mechanisms involved in carbon partitioning at an early stage of development. Gene expression studies using immature internodes often involve comparisons to other organs (Casu *et al.*, 2003, Casu *et al.*, 2004; Papini-Terzi *et al.*, 2013; de Barros Dantas *et al.*, 2020) or between culms at a later developmental stage and low levels of sucrose accumulation (Thirugnanasambandam *et al.*, 2017). However, to the best of our knowledge there is no study using RNA-Seq data to assess the expression patterns in internodes at the most immature stage of development. Internode +1 may provide useful information regarding the process of carbon partitioning in young sucrose storage cells. Using a diverse set of genotypes with a broad range of sugar yield, we could assess whether there was consistency in the expression patterns of sugar-related genes. Instead of using a single accession per group, this approach is suitable to reveal alternative routes to achieve similar phenotypes, because we can evaluate genotypic specificities regarding the expression levels of sugar regulators. Here, our main purpose was to characterize putative key regulators of carbon partitioning that could initiate the phenotypical differentiation between high and low sugar sugarcane genotypes.

2 MATERIALS AND METHODS

2.1 Biological samples and phenotypic characterization

From the 266 accessions available in the Brazilian Panel of Sugarcane Genotypes, we selected 12 genotypes (Table 1) to be sampled for the RNA-Seq experiment, with the goal of representing a large range of phenotypes and characteristics of interest. This panel consists of a diverse set of germplasm available for sugarcane plant breeding programs in Brazil and is grown in field conditions in the city of Araras (22.31602 S, 47.38929 W). The selected genotypes include plants from the parental species *S. spontaneum* and *S. officinarum*, which were intercrossed to produce most of the genetic basis of modern Brazilian cultivars. In June 2016, we sampled six-month-old plants from a field trial in a complete block design, collecting one sample per genotype of the internode +1. This internode is located just below the apical meristem, representing an early stage in the development of sugarcane culms.

Barreto *et al.* (2019) had previously evaluated these genotypes for their content of soluble solids (measured as ºBrix) and fiber content. We used this data with the goal of establishing possible associations between the expression levels of identified genes and the traits of interest (Supplementary Figure 4). Based on the content of soluble solids, we gathered the genotypes SES205A, Krakatau, and IN84-58 in the Very Low ºBrix group (VLB); Criolla Rayada, F36−819, and IJ76−317 represented the Low ºBrix group (LB); R570, White Transparent, and RB92579 with High ºBrix (HB); and RB835486, SP80-3280, and White Mauritius showed the highest levels of soluble solids, forming the Very High ºBrix set (VHB). We planned the experimental design so that plants of different genotypes represent biological replicates from the same phenotypic group. All genotypes in VLB are *S. spontaneum* accessions, showing low sugar and high fiber content in comparison with *S. officinarum* and commercial sugarcane hybrids. It is apparent from the distribution of phenotypic values that the VLB genotypes were markedly different from the other groups, particularly with regard to fiber content (Supplementary Figure 4 b). On the other hand, differences between the LB, HB and VHB groups were of smaller magnitude.

2.2 RNA-Seq and data preparation

We first extracted the total RNA from each sample with the RNeasy Plant Mini Kit (Qiagen) and prepared poly-A (+) sequencing libraries with the TruSeq Stranded mRNA kit (Illumina), following the manufacturer's recommendations. Sequencing was performed on the Illumina HiSeq 2500 platform,

Genotype	Species	Group	Observations
IN84-58	*S. spontaneum*	VLB	
Krakatau	*S. spontaneum*	VLB	Originally from Indonesia, n = 60 (Panje & Babu, 1960)
SES205A	*S. spontaneum*	VLB	Originally from India, West Bengal, n = 32 (Panje & Babu, 1960)
Criolla Rayada	*S. officinarum*	LB	
F36-819	*S. spp - hybrid*	LB	North american germplasm
IJ76-317	*S. officinarum*	LB	
R570	*S. spp - hybrid*	HB	Widely used in genomics studies and used to assemble a monoploid genome draft (Garsmeur *et al.*, 2018)
RB92579	*S. spp - hybrid*	HB	Important in Northeast Brazil, tolerates high temperatures and water stress (Gazaffi *et al.*, 2016)
White Transparent	*S. officinarum*	HB	
RB835486	*S. spp - hybrid*	VHB	Economically important to the Brazilian central region (EMBRAPA, 2008)
SP80-3280	*S. spp - hybrid*	VHB	Genotype sequenced to obtain a gene-space assembly (Souza *et al.*, 2019)
White Mauritius	*S. officinarum*	VHB	

Table 1: Sugarcane genotypes selected for the RNA-Seq experiment. The table shows whether each genotype is a wild accession or a hybrid and their origin (when known). Separation into the four groups was based on the average content of soluble solids. VLB - Very Low ºBrix, LB - Low ºBrix, HB - High ºBrix, VHB - Very High ºBrix

using a single lane for a pool with the 12 libraries. The 2 x 100 bp paired-end fragments underwent a quality control procedure for removing adapters and cleaning low-quality bases. We cleaned the raw cDNA sequencing data using Cutadapt v1.18 (Martin, 2011) and Trimmomatic v0.38 (Bolger *et al.*, 2014). Cutadapt was used to remove residual sequences of adapters from the sequencing protocol. With Trimmomatic we cropped the first 13 bases of each read and cleaned bases with Phred quality score less than 20 in a sliding window of 5 bp. Only reads longer than 50 bp after preprocessing were kept.

2.3 Transcriptome assembly and functional annotation

The twelve high quality read sets were used to perform a *de novo* transcriptome assembly with the software Trinity v2.8.0 (Grabherr *et al.*, 2011), after normalization by read set to more uniformly represent each biological sample, using the parameter for strand specific assembly. We used TransDecoder (`https://github.com/TransDecoder/TransDecoder`) to obtain predicted long ORFs from the assembly. These data were used to get significant (E-value < 1e-5) blastx and blastp hits (Altschul *et al.*, 2010) against the Swiss-Prot database. Protein homologues were also identified with HMMER v3.2.1 (Eddy, 2009) against the Pfam database. We used Trinotate v3.1.1 (`https://github.com/Trinotate/Trinotate`) to combine these sources of information and compile the final annotation of Gene Ontology (GO) terms. We also assigned KEGG Orthologues (KO) via pathway mapping with KAAS (Moriya *et al.*, 2007), using the predicted proteins as queries and the available databases from the species *Oryza sativa*, *Zea mays*, and *Aegilops tauschii*. We removed the generic pathways K01100 (metabolic pathway) and K01110 (biosynthesis of secondary metabolites) due to lack of specificity. For identifying the name of putative homologues to the differentially expressed genes (DEGs), in addition to the automated annotation, we also manually included selected proteins from GenBank and RefSeq (Supplementary Table 2) in our local Swiss-Prot-based reference and assigned names from the best blastx hit to each gene.

2.4 Assessment of individual conserved orthologues in the transcriptome

In order to evaluate assembly properties of individual genes in such a complex transcriptome we performed multiple alignments of four conserved orthologues in grasses. To select candidate genes for this investigation we first selected conserved orthologues from BUSCO v4.0.6 (Simão *et al.*, 2015) that had multiple hits in the assembled transcriptome, namely Cyclin P2-1, SWEET16, SWEET2b, and a mitochondrial Arginase 1. We considered two different scenarios regarding the status of the assembled genes: *i)* single-copy orthologues mapped to a single Trinity gene containing multiple transcripts and *ii)* single-copy orthologues mapped to multiple Trinity genes. The software msa (Bodenhofer *et al.*, 2015) was used to align cDNA sequences of each transcript belonging to the target gene, as well as the amino acid sequences corresponding to their ORFs with significant homology to the target protein. From the several genes found in the listed cases, we selected Arginase 1 because of its ubiquity in every grass dataset available in the BUSCO database; Cyclin P2-1, due to the duplicity of blast hits in the assembled transcriptome; and SWEETs because of their role in sugar transport. We also studied the presence of conserved domains among proteins using the webserver version of HMMER (Eddy, 2009), aiming to evaluate the structure and function of each transcript. Besides, we used the results from BUSCO to analyze the whole transcriptome and the subset of longest transcripts for each gene, aiming to evaluate the completeness of the conserved orthologues among the clades of green plants, monocots, and grasses.

2.5 Differential expression and functional enrichment analyses

For each read set, we quantified the expression both at the transcript and gene levels using a quasi-mapping strategy, as implemented in Salmon v0.12.0 (Patro *et al.*, 2017) (Supplementary Table

3). The quasi-mapping strategy of Salmon is optimized for estimating transcript abundance with high accuracy while taking into account several sources of bias, such as using the length of each transcript as weights on gene abundance calculation. We first created an index for the transcriptome with a k-mer size of 31 and then used the default EM algorithm with GC bias correction to estimate transcript abundances. For downstream steps we filtered for genes with count per million (CPM) greater than one in at least three samples. Three orthogonal contrasts were designed, representing meaningful biological comparisons regarding the four groups of genotypes. The first contrast tested for differences in expression between VLB and the average of the other three groups, namely VLB x VHB.HB.LB for the null hypothesis $H_0 : \pi_{g,VLB} = \frac{\pi_{g,VHB}+\pi_{g,HB}+\pi_{g,LB}}{3}$; the second compared LB with the average of VHB and HB, denoted by LB x VHB.HB, $H_0 : \pi_{g,LB} = \frac{\pi_{g,VHB}+\pi_{g,HB}}{2}$; and the last one compared HB and VHB, HB x VHB, $H_0 : \pi_{g,HB} = \pi_{g,VHB}$, for each gene g. The group listed after the versus signal for each contrast was considered the reference group. Filtered genes were subjected to the likelihood ratio test using the edgeR library (Robinson *et al.*, 2010), and genes were called as differentially expressed if the corresponding p-value was less than 0.05, after correcting for multiple tests with the false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995). Enriched Gene Ontology and KEGG Orthology terms were found using the library goseq (Young *et al.*, 2010) after correcting by mean gene weighted lengths.

Additionally, we designed three contrasts to better understand the sets of differentially expressed genes identified in the contrast VLB x VHB.HB.LB. Following the notation mentioned above, we represented them as VLB x VHB ($H_0 : \pi_{g,VLB} = \pi_{g,VHB}$), VLB x HB ($H_0 : \pi_{g,VLB} = \pi_{g,HB}$), and VLB x LB ($H_0 : \pi_{g,VLB} = \pi_{g,LB}$).

Heatmaps were constructed for sets of genes of particular interest. We selected genes from the custom databased formed by combining Swiss-Prot and additional genes from NCBI, excluding chloroplastic and mitochondrial genes. The heatmaps were displayed in a $log_2(CPM + 1)$ scale to avoid issues with zero counts. The genotypes were clustered based on the correlations between the heatmap columns.

3 RESULTS

3.1 Assembly statistics

The *de novo* transcriptome assembly resulted in a list of 476,721 transcripts with an average length of 940.79 bp, N50 of 1,692 bp, and distributed in 219,725 putative genes. Most of the genes contained a small number of transcripts (66.8% had a single transcript, 13.0% had two, and 5.4% had three transcripts).

For assessing its completeness of conserved protein orthologues, we used the software BUSCO v4 as a diagnostic tool (Table 2). Considering the database containing single-copy orthologues among Viridiplantae (OrthoDB10), which includes 425 single-copy genes from 57 species of plants, we found that 96% of them were included in the assembled transcriptome. For the databases corresponding to Liliopsida and Poales (OrthoDB10), nearly 95% of the orthologues were represented in both of them. The low rates of fragmented orthologues (3.3%, 6.9%, and 3.0%, respectively) indicate high confidence in the assembly quality. We note that a single-copy orthologue from the reference database may result in a duplicated hit even if the matched transcripts belong to the same assembled gene. As a consequence, multiple hits using all transcripts are not informative for transcriptome gene assembly quality. In this sense, we also performed BUSCO analysis for the subset of the longest transcripts in the transcriptome, such that each assembled gene be represented only once. However, this approach is not necessarily accurate in every case, because Trinity transcripts may have different annotation as well as low-scoring alignments for the longest transcripts. The results show high levels of single hits against the longest transcripts, with 77.6% for green plants, 60.5% for monocots, and 69.4% for grasses. Increased rates of missing values in comparison with the analysis for all transcripts might also indicate that a few orthologues were misassembled and collapsed as the same gene.

3.2 Assembly evaluation for conserved orthologues in Poales

We identified one gene assembled by Trinity with homology to Uniprot protein Q7X7N2, identified as *arginase 1, mitochondrial*. With the cDNA and predicted amino acid sequence of the translated protein, we performed a multiple alignment for its four transcripts to explore their homology and investigate the gene structure in sugarcane (Supplementary Figure 5 a). The cDNA sequence presented blocks of conserved base pair residues flanked by exclusive regions for each transcript. Sequences i6, i4, i14, and i8 have blocks of divergent nucleotides in their chains spread by five positions over the multiple alignment. Only the former has a variant sequence in one of the edges. Besides, all pairwise combinations showed at least one single nucleotide polymorphism (SNP), except for the pair i8 and i4, which disagreed

Category	Viridiplantae		Liliopsida		Poales	
	All transcripts	**Longest transcript**	**All transcripts**	**Longest transcript**	**All transcripts**	**Longest transcript**
Single hit	20.0%	77.6%	20.9%	60.5%	18.7%	69.4%
Multiple hit	76.0%	0.5%	68.8%	0.9%	70.8%	1.6%
Fragmented	3.3%	16.7%	6.9%	22.3%	3.0%	10.4%
Missing	0.7%	5.2%	3.4%	16.3%	7.5%	18.6%

Table 2: Main statistics of BUSCO v4 analysis. The *de novo* assembled transcriptome was compared to datasets of 425 single-copy orthologues of green plants (Viridiplantae), as well as 3,236 genes found in monocotyledons (Liliopsida), and 4,896 single-copy genes common to grasses (Poales). We carried out analyses both with every transcript in the assembled reference and with the subset of the longest transcript of each gene. Single and multiple hits were identified as orthologous blast hits against one or more transcripts in our reference; fragmented hits represent orthologues whose tblastn alignment had low coverage of the target gene; missing orthologues were not represented in the reference.

by a large exclusive region in i4 and 3' extremities. Those sequence particularities resulted in predicted polypeptides different lengths (Table 3), but with a common N-terminus. According to the results of protein domain prediction, only i14 had an arginase family motif (PF00491.21). The protein sizes and identified protein domains suggest that i6 and i8 cDNAs do not include 3' edges for the functional protein C-terminus, and the long i4-exclusive region introduced a premature stop codon.

Cyclin P2-1, another single-copy orthologue, had two assembled genes identified with BUSCO. Evidence from both alignment and ORF identification implies that their cDNAs come from different strands of the double helix. For the assembled gene TRINITY_DN34761_c0_g1, the ORF coding Cyclin P2-1 was in the direct cDNA orientation, while for TRINITY_DN67488_c1_g1 it was in the reverse complement orientation. Multiple alignment also implied sequence homology when using the complementary sequence of the latter (Supplementary Figure 5 b). The three transcripts from TRINITY_DN34761_c0_g1 varied only by an indel (insertion or deletion of a base pair) in the residue 1342 of the sequence i1, which possibly represents a sequencing error, and a 15 bp gap in i3. Thus, it is reasonable to assume that i1 and i2 have the same splicing pattern. Moreover, i2 and i3 may belong to the same allele with alternative splicing patterns. Several polymorphisms in the TRINITY_DN67488_c1_g1 transcripts suggest they do not belong to the same allele nor match any sequence of the other gene. A duplication of 48bp, found exclusively in TRINITY_DN67488_c1_g1_i2, spans the reverse sequence from base pairs 179 to 226 and 382 to 429. The resulting proteins had the same amino acid composition, despite all variations in cDNAs. Regarding their expression levels, the protein coding gene had an average CPM of 5.0, while the other had 0.3 CPM, and therefore it was removed from differential expression analysis due to low expression.

We also examined the assembly results of two putative sugar transporters, namely SWEET2b and SWEET16. A single gene aligned against the former transporter, with eight transcripts that we assorted into three groups according to their 5' edges. Transcripts i4, i5, and i6 formed group A; in group B were i1, i2, i7, and i8; finally, i13 was the only transcript in group C (Supplementary Figure 5 c). Because portions of the 5' sequences were highly variable, the alignment quality was poor with moderate evidence of homology in these regions. After the base pair residue number 250 for group A and 109/111 for B/C, the cDNA sequences were nearly identical for all the transcripts. In this region, three SNPs supported the grouping of transcripts i1 and i5, a 3 bp indel grouped i1 and i8, and an indel of 12 bp grouped i2 and i5. The homologous ORFs yielded a 231 aa protein for group A and a 90 aa protein for B/C. One of the three SNPs in the conserved residues produced an amino acid variation of leucine for i1 and i5 and phenylalanine for the others, both with non-polar side chains. Remarkably, HMMER predicted two repeated sugar efflux (PF03083.16) domains, with seven transmembrane helices for group A proteins, while proteins in groups B and C had a single PF03083.16 domain, two transmembrane helices, and a signal peptide. Average expression levels from group A transcripts were roughly 2.8 TPM, 0.6 TPM for group B, and 1.4 TPM for group C.

Transcript	**Transcript size**	**ORF size**	**Transcripts per million**
TRINITY_DN8277_c0_g1_i4	2957 bp	77 aa	1.5 ± 0.7
TRINITY_DN8277_c0_g1_i6	504 bp	102 aa	1 ± 2
TRINITY_DN8277_c0_g1_i8	660 bp	131 aa	0.9 ± 0.6
TRINITY_DN8277_c0_g1_i14	1623 bp	340 aa	40 ± 10

Table 3: General characteristics of TRINITY_DN8277_c0_g1, the *Arginase 1, mitochondrial* orthologue in sugarcane. This gene represents the only sequence in the transcriptome whose best blastx hit was against the *Oryza sativa* arginase, and it is formed by a group of four transcripts. The ORF of each transcript was selected by homology to the arginase amino acid sequence. Transcript abundance is represented by the average and standard deviation over all genotypes.

SWEET16 followed a pattern similar to SWEET2b. Trinity assembled the SWEET16 gene with three transcripts, namely i2, i3, and i4 (Supplementary Figure 5 d). Sequences i2 and i4 differed only by a large indel of 125 bp. The i3 cDNA had a 112 bp long low-identity alignment to the other sequences in its 3' end, which is possibly due to the absence of homology rather than sequence polymorphism. Thus, there is no substantial evidence supporting the occurrence of multiple alleles for this gene, because it can be argued that all observed variation was due to alternative splicing. Similarly to group A protein of SWEET2b, the longest ORF of SWEET16 i4 also codes a protein with two PF03083.16 domains and seven transmembrane helices, despite having a predicted long disordered C-terminus. The insertion in i2 cDNA resulted in an early stop codon, reducing the ORF length to only 24 coding amino acids. However, a second ORF results in a 226 aa long protein, which included a single sugar efflux domain, three transmembrane helices, and a signal peptide, resembling groups B and C of SWEET2b proteins. At last, the i3 isoform yielded a protein with one PF03083.16 domain and five predicted transmembrane helices. Transcript quantification resulted in 1.2 TPM for i4, 0.1 TPM for i2, and 0.3 for i3.

3.3 Differential expression and functional enrichment results

We filtered the assembled genes by expression level, resulting in 41,650 genes available for differential expression analysis, with 11,910 of them annotated with GO terms, and 4,017 annotated with KO terms. A multidimensional scaling plot (MDS) was performed to provide an overview of the expression profiles, considering the most divergent genes between pairwise combinations of samples (Figure 1). The first component of the MDS analysis corroborated the separation of *S. spontaneum* genotypes from the others, as seen in (Supplementary Figure 4 b), whereas the distribution pattern of the *S. officinarum* and hybrids was not consistent. Also, the hybrid genotypes did not present an intermediary position between the parental species, grouping with two accessions from *S. officinarum* (White Mauritius and White Transparent), while Criolla Rayada and IJ76−317 remained isolated from the other plants. The results show no evidence of clustering by sugar yield in the maturity stage.

To perform differential gene expression analyses we first designed three orthogonal contrasts, namely VLB x VHB.HB.LB, LB x VHB.HB, and HB x VHB. Using the methodology proposed by McCarthy *et al* (2012), we performed a likelihood ratio test for each contrast, resulting in the identification of 10,111 DEGs after false discovery rate correction of p-values (Benjamini & Hochberg, 1995). For the first contrast, we found 4,256 genes upregulated and 5,662 downregulated in VLB; the second showed 82 up- and 325 downregulated genes in LB; and for the latter, we found genes 23 up- and 41 downregulated in HB (Supplementary Figure 6). The divergence between VLB and the other groups was also explored using individual contrasts. The comparison VLB x LB yielded 8,624 DEGs, followed by 5,828 in VLB x HB, and 5,520 in VLB x VHB, with a larger proportion of genes repressed in VLB for all contrasts (Supplementary Figure 7). From the intersection of 3,911 DEGs in common among these contrasts, only one was not detected as differentially expressed in the original VLB x VHB.HB.LB test.

The functional enrichment test is a strategy for summarizing information from a large number of DEGs based on their biological roles. Our analyses revealed enriched terms only for the VLB x VHB.HB.LB and LB x VHB.HB contrasts. In the first case, the terms were related to diverse groups of molecular functions and biological processes. In the second, the only enriched term was adenosine diphosphate (ADP) binding.

Regarding the enriched terms of the first contrast, genes annotated with terms DNA integration, transposase activity, and hydrolase activity on ester bonds were more highly expressed in the groups of higher sugar content (Figure 2). Of these, only the last term is not related to transposable element activity, because it represents a diverse group of proteins with varying biological roles. On the other hand, there was a clear overrepresentation of DEGs with cellulose synthase activity with higher expression in VLB

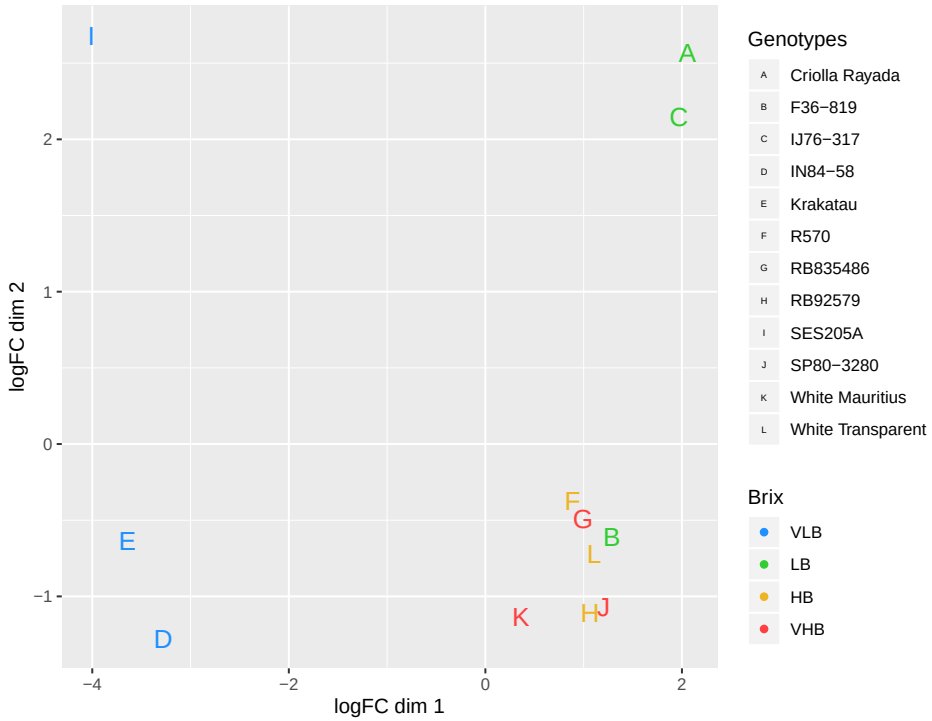


Figure 1: Multidimensional scaling plot based on gene expression profiles of sugarcane genotypes. The clustering of biological replicates partially matches that based on phenotypic data, with separation of the Very Low ºBrix group from the others in the first component. However, separation of the genotypes with higher sugar levels is not clear, suggesting that the most marked differences in expression among these samples were not related to their sugar content.

— which is redundant with the also enriched biological process of cellulose biosynthesis. Also, we found the same set of enriched GO terms for the DEGs found in common in the VLB x VHB, VLB x HB, and VLB x LB contrasts.

In addition, we performed separate functional enrichment tests with the subset of up and down-regulated genes. For genes overexpressed in VLB, terms related to oxidoreductase activity, membrane components, protein phosphorylation, and binding to iron ion and chitin were only significant using this approach. Protein dimerization, transposition mediated by DNA, telomere maintenance, RNA-DNA hybrid ribonuclease activity, and binding to zinc ion and DNA were significantly enriched only in high-sugar groups (VHB, HB, and LB).

The enrichment of genes within specific pathways showed significant KO terms only in the VLBxVHB.HB.LB contrast, in agreement with gene ontology analysis. The DEGs which have a more extensive contribution to KEGG pathway enrichment were the ones upregulated in VLB. Six KO terms were significant among these genes (Figure 3), while there was no significant enrichment for the downregulated ones. The outcome for the complete set of DEGs for this contrast resulted in four detected terms, namely alpha-linolenic acid metabolism, MAPK signaling pathway, photosynthesis (antenna proteins), and plant hormone signal transduction. Among the DEGs in the upregulated set for sucrose and starch metabolism, we highlight genes that will be discussed with further details such as sucrose synthase and cell-wall invertase (beta-fructofuranosidase). Carbon fixation process includes genes from Calvin-Benson and C4 cycles. The antenna proteins orthology term was composed of only 11 genes associated with light-harvesting complexes I and II, nine of which were upregulated in VLB. The KOs for hormone signaling represented genes related to several hormone classes, *e.g.*, abscisic acid, auxin, gibberellin, ethylene, brassinosteroids, jasmonate, as well as protein kinases and transcription factors.

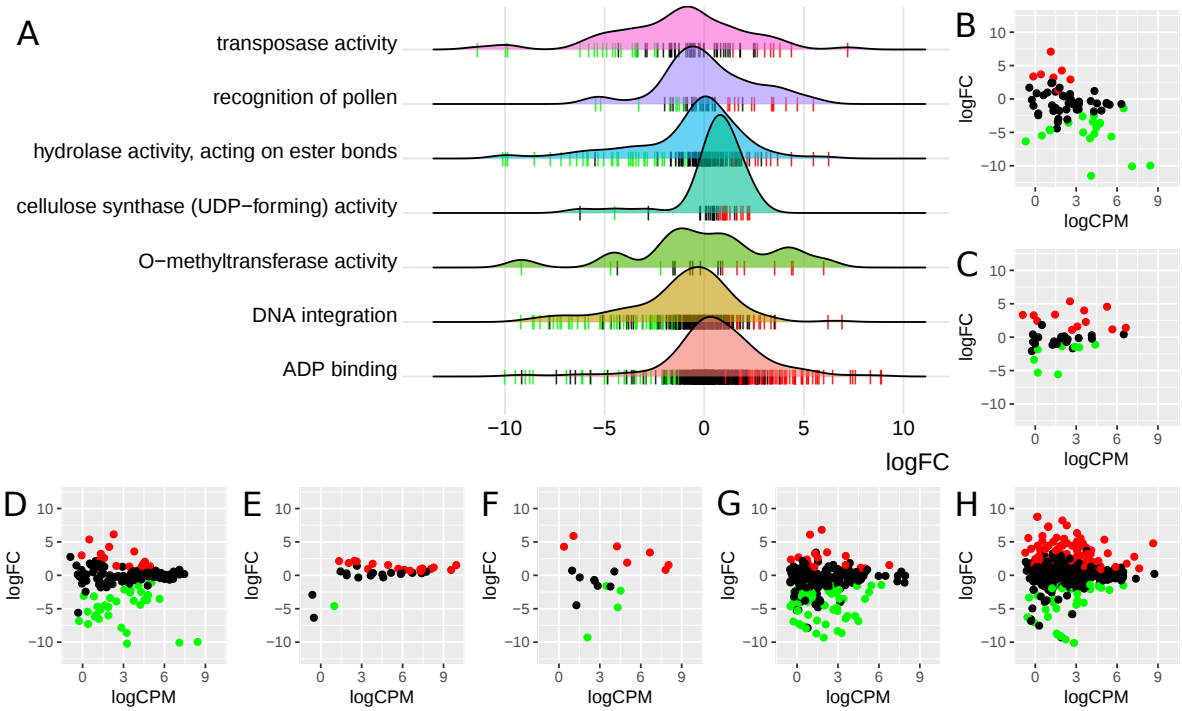


Figure 2: Gene ontology terms enriched among differentially expressed genes for the VLB x VHB.HB.LB contrast. Upregulated (downregulated) genes are marked in red (green), and not significative in differential expression as black. (A) LogFC density plots show relative expression rates in favor of VLB (positive logFCs, in red) or the average of VHB, HB, and LB (negative logFC, in green). From B to H, individual mean-difference plots for the enriched terms transposase activity (B); recognition of pollen (C); hydrolase activity, acting on ester bonds (D); cellulose synthase (UDP-forming) activity (E); O-methyltransferase activity (F); DNA integration (G); and ADP binding (H).

To better understand the expression patterns of specific genes of interest, based on previous studies and for well-established molecular processes, we constructed heatmaps showing gene abundances per genotype, including DEGs and non-DEGs related to carbon partitioning in sugarcane (Figure 4). We investigated three particular groups of genes of interest, namely cellulose synthase A subunits, genes involved in sucrose synthesis and/or breakdown, and sugar transport genes. The heatmaps show the expression levels for all genotypes in terms of z-scores of gene counts, in order to visualize the biological variation of each gene both within and among groups.

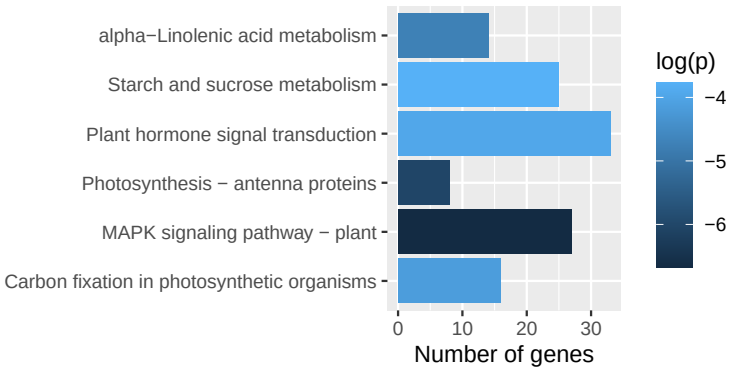
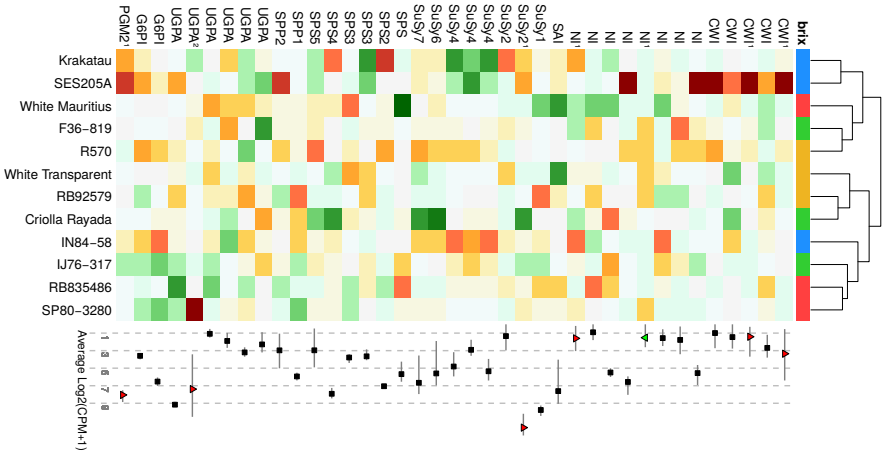


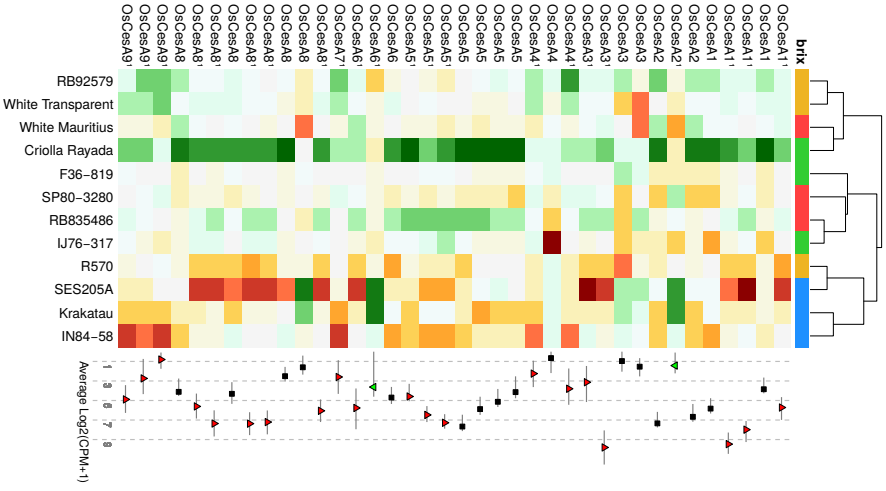
Figure 3: KEGG pathway enrichment for genes upregulated in VLB. The genes used to identify these KEGG Orthology terms were significantly differentially expressed in the VLB x VHB.HB.LB contrast, with a positive log fold-change.

3.3.1 Sugar conversion enzymes

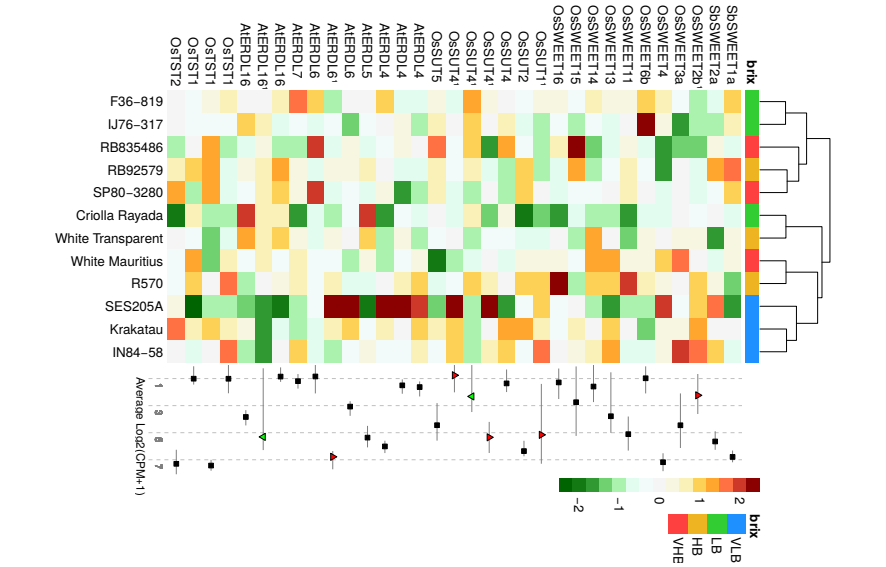
The sucrose-related enzymes belong to four main categories depending on their activity. Our results show that sugarcane harbors several variant forms of expressed invertases (Figure 4 a). There were two out of eight differentially expressed NI and two out of five CWI in the comparison between VLB and the other groups, with higher expression in VLB for three of these four genes. Interestingly, one of the NI was not expressed by any of the *S. spontaneum* genotypes or White Mauritius. Despite there being more expressed NI genes than acid invertases, the overall CPM of NI genes was only higher than SAI for eight samples. While invertases act by cleaving sucrose into simple hexoses, SuSy enzymes are responsible for both degradation and synthesis of sucrose, depending on their reaction substrates. One SuSy2 was differentially expressed and showed a high average expression level — $log_2(CPM)$ ranging from 10.22 to 12.96 —, with a two-fold higher variation in VLB genotypes. Three representatives of SuSy4 showed a consistent pattern among the *S. spontaneum* accessions, with lower expression levels in Krakatau and SES205A, but higher expression in IN84-58. SPS and SPP form a third group of genes acting over sucrose synthesis in sugarcane stalks, but showed no significant genes in the differential expression tests. The fourth set of enzymes is responsible for the conversion of hexoses with covalent modifications, connecting SuSy and SPS reactions in a cycle. UTP–glucose-1-phosphate uridylyltransferase, also known as UGPA, had six assembled genes in the transcriptome, including two highly expressed ones. From the other genes in this set, namely glucose-6-phosphate isomerase and phosphoglucomutase, only the latter was differentially expressed, being upregulated in all genotypes of VLB.



(a) enzymes involved in the synthesis, breakdown, and conversion of sugars



(b) cellulose synthase A subunits



(c) sugar transporters

Figure 4: Heatmaps of expression levels for genes related to carbon partitioning in sugarcane. On the left, plots for z-scores based on $log_2(CPM + 1)$ values of selected genes, separately for each genotype. Columns were clustered based on the Euclidean distance to highlight groups with similar transcriptional profiles. The sets contain both DEGs and non-DEGs for the designed contrasts, with DEGs indicated by a superscript next to the corresponding gene name (index one indicates DEGs from the VLB x VHB.HB.LB contrast). On the right, the minimum, maximum, and average levels of $log_2(CPM + 1)$ are displayed, with upregulated (downregulated) genes shown by red (green) triangles, and non-DEGs in black dots.

3.3.2 Cellulose and other fiber components

Genes involved in cellulose synthesis were not only enriched as shown by gene ontology functional activity, but also showed increased expression in the VLB group in most cases (Figure 2 e). Of the 39 CesA genes identified in the transcriptome, 38 had the best blast hit against *Oryza sativa* genes and 22 of them were significantly differentially expressed in at least one contrast (Figure 4 b). The contrast VLB x VHB.HB.LB also revealed several families of genes associated with cell wall components upregulated in the VLB genotypes. Among them, we identified two putative cellulose synthase-like proteins; ten xyloglucan associated genes such as endotransglucosylases, hydrolases, glycosyltransferases, and a fucosyltransferase; nine subtilisin-like homologues from eight genes, which are related to cell wall proliferation; cell wall expansion related proteins, e.g., nine expansins, three expansin-like proteins, and three COBRA-like proteins; as well as six endoglucanases. From the mentioned families, only one expansin-like A1 was downregulated in VLB. Some classes of cell wall genes showed genes with differential expression in both directions, for example, pectinesterases/pectinesterase inhibitors, with three upregulated and three downregulated genes, and two upregulated and one downregulated callose synthases.

Phenylpropanoids are also a relevant class of compounds for cell wall biosynthesis, because they are precursors of lignin, one of the main fiber components (Vicentini *et al.*, 2015). We observed several genes of the phenylpropanoid biosynthetic pathway upregulated in the high-fiber, low-sugar group VLB, starting with the amino acids phenylalanine and tyrosine as substrates. For instance, we detected four phenylalanine ammonia-lyases, two phenylalanine/tyrosine ammonia-lyases, one trans-cinnamate 4-monooxygenase, one 4-coumarate–CoA ligase 3 and two 4-coumarate–CoA ligase like, one caffeoyl-CoA O-methyltransferase, one cinnamoyl-CoA reductase 1 and cinnamoyl-CoA reductase-like SNL6, one caffeic acid 3-O-methyltransferase, and two cinnamyl alcohol dehydrogenase. On the other hand, we only found one caffeic acid 3-O-methyltransferase downregulated in VLB.

3.3.3 Sugar transmembrane transporters

The classes of sugar transporters selected in this study were sucrose transport protein (SUT), bidirectional sugar transporter (SWEET), sugar transporter early response to dehydration 6-like (ERDL), tonoplast sugar transporter (TST), and vacuolar glucose transporter (VGT). We identified in our transcriptome genes of sugar transporters reported to be relevant to phloem unloading and uptake in other grasses, except for VGT. Interestingly, sugar transporters were the only category of genes among those investigated for which the VLB genotypes clustered together as a group (Figure 4 c), consistently with their low content of soluble solids. We also note that some of these genes showed extreme expression patterns for SES205A, which resulted in increased fold changes of the differentially expressed genes in the VLB x VHB.HB.LB contrast.

3.3.4 Hormone-related genes

Several genes reported as related to hormone biosynthesis, degradation, response, transport, and signal transduction showed evidence of significant differential expression for the VLB x VHB.HB.LB contrast, leading to an enrichment of the hormone signal transduction pathway. We found 36 DEGs related to auxins, 39 to ethylene, 12 to gibberellin, and 8 to ABA. Among auxin-related genes, transcription factors included response factors, responsive protein IAA, responsive protein SAUR, auxin-induced protein 5NG4, and auxin-induced protein X10A, of which 21 were upregulated and eight were downregulated in VLB. We also identified two classes of auxin transporters with a total of seven genes, all of them upregulated in VLB. For ethylene, we found three families of binding receptors with four upregulated genes and six families of transcription factors with 24 up and five downregulated genes. We also found six copies of 1-aminocyclopropane-1-carboxylate oxidase, four of them upregulated. The set of genes

responsive to gibberellins revealed six upregulated enzymes for their synthesis/breakdown, three GID1 intracellular receptors, two extracellular receptors, and a chitin-inducible transcription factor. Finally, the ABA-related DEGs comprise three transcription factors, three receptors, and two ABA degrading hydroxylases.

3.3.5 Pollen recognition

It is interesting to note that, unlike the other groups, VLB consists exclusively of wild *S. spontaneum* accessions, which have not undergone selection against flowering. On the other hand, the sugarcane hybrids present in the other groups are less prone to produce flowers and to undergo sexual reproduction. By analyzing genes annotated with the gene ontology term GO:0048544, named recognition of pollen, we found 21 DEGs in the VLB x VHB.HB.LB contrast. Among those genes, there were eight up- and four downregulated G-type lectin S-receptor-like serine/threonine-protein kinase; four receptor-like serine/threonine-protein kinase upregulated and one downregulated in VLB; two downregulated PAN domain-containing protein; as well as a downregulated S-locus-specific glycoprotein S13.

3.3.6 Transposable elements

We detected transposable elements as a significant term in gene ontology analysis, with a prevalence of genes upregulated in VHB, HB, and LB rather than in VLB, revealing that transposition may be an essential molecular phenomenon distinguishing these genotypes. We found 1,000 expressed genes related to transposable elements in the assembled transcriptome, assigned into 24 unique gene names from the Uniprot database. From these, 19 had at least one differentially expressed gene (Figure 5). A total of 9.3% of the identified transposable element genes were upregulated and 23.2% were downregulated in the VLBxVHB.HB.LB contrast.

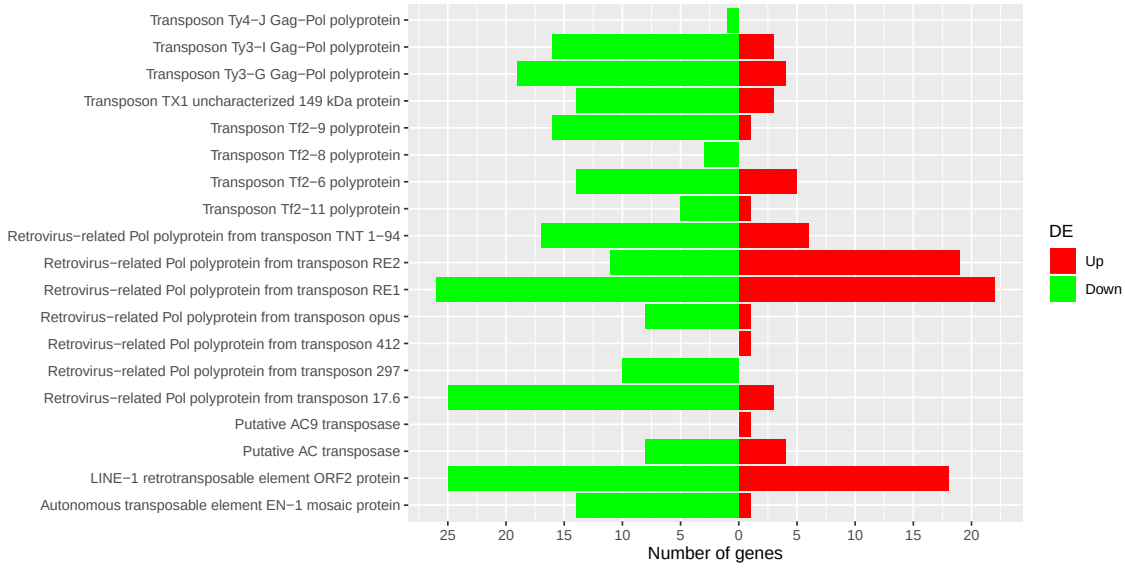


Figure 5: Differentially expressed genes associated with transposable elements in the VLB x VHB.HB.LB contrast. The DEGs were identified by homology to known genes of transposons and retrotransposons in Swiss-Prot. Colored bars represent up and downregulated genes in VLB.

4 DISCUSSION

4.1 *De novo* assembly challenges

The different hierarchical levels of organization of transcripts — such as gene, locus of origin, allele, and isoform — can be confounded due to sequence variation, especially in sugarcane, because of its complex genome. The following manually analyzed genes illustrate some of these effects over gene assembly. We used conserved single-copy genes with no closely related paralogues or pseudogenes in several grasses, such as Arginase 1, Cyclin P2-1, and the SWEET transporters 2b and 16, to assess their assembly in the sugarcane transcriptome. Even with the presence of genotypes from different species and genetic backgrounds in the dataset, and the complexity of the sugarcane transcriptome, we consider that this *in silico* evaluation can be useful to raise hypotheses regarding the assembly quality and overall expression patterns. Arginase 1 (Uniprot ID Q7X7N2) is a rice enzyme with 340 aa, of which 24 correspond to a transit peptide for mitochondrial destination. The other grasses arginase orthologues identified in OrthoDB v10.1 range from 337 to 342 aa, except for the *Triticum aestivum* arginase, which has 193 aa. In the current study, we obtained a single assembled gene with homology to rice arginase, supporting the hypothesis that this gene is also single-copy in sugarcane. Polymorphisms in cDNA sequences suggest that the gene has three to four alleles among the genotype samples we sequenced, being possible that one allele admits two splicing isoforms. The transcript i14 produces the only predicted protein with a compatible size with the other grasses (Table 3). Moreover, it is 35-fold more abundant than the others. This fact indicates that the variant alleles are not frequent in all sampled genotypes, or that there is a regulatory mechanism responsible for the increased abundance of this transcript.

Cyclin P2-1 matched two genes in the *de novo* assembled transcriptome. With evidence from both ORF identification and multiple sequence alignment, we can argue that one assembled gene produces to a protein-coding cDNA and an antisense cDNA of Cyclin P2-1. We found no predicted known protein for the putative antisense sequence, which represents another evidence for the antisense expression hypothesis. Moreover, the library preparation protocol uses a strategy for stranded sequencing of the original mRNA. Even considering possible failures in the protocol, this would not explain the lack of assembled antisense genes for more abundant transcripts, given the low expression of both Cyclin P2-1 genes. Although we have not found molecular studies investigating antisense expression of cyclins in plants, the phenomenon has already been identified for cell cycle genes in sugarcane. Sense-antisense transcript pairs for the cell cycle process were enriched among differentially expressed genes in water deprivation experiments for aerial tissues (Lembke *et al.*, 2012). Thus, the hypothetical existence of antisense regulation for Cyclin P2-1 should be evaluated in further molecular experiments. Nevertheless, it is feasible to postulate that different layers of regulation might control cyclin levels in plant cells, because the abundance of these proteins varies along the cell cycle. With the number of polymorphisms alone, it is not possible to check whether they belong to different alleles or loci. By analyzing SWEETs found in the BUSCO scan, we could also identify putative markers of allelic variation as well as alternative splicing isoforms for each single gene.

4.2 Sucrose synthesis and breakdown in young internodes

(Figure 6) shows a putative simplified scheme of the main reactions that collaborate directly to carbon partitioning in the sugarcane sink tissue. Although several authors have proposed a variety of proteins as central regulators of sucrose content, the mechanics are not fully clarified yet. Invertases have been pointed as main regulators because they degrade sucrose, leading to lower concentrations in the sink tissue and thus increasing its osmotic potential. Our results indicate that sugarcane expresses at least 14 variant forms of genes with homology to known invertases, which can account as actual different

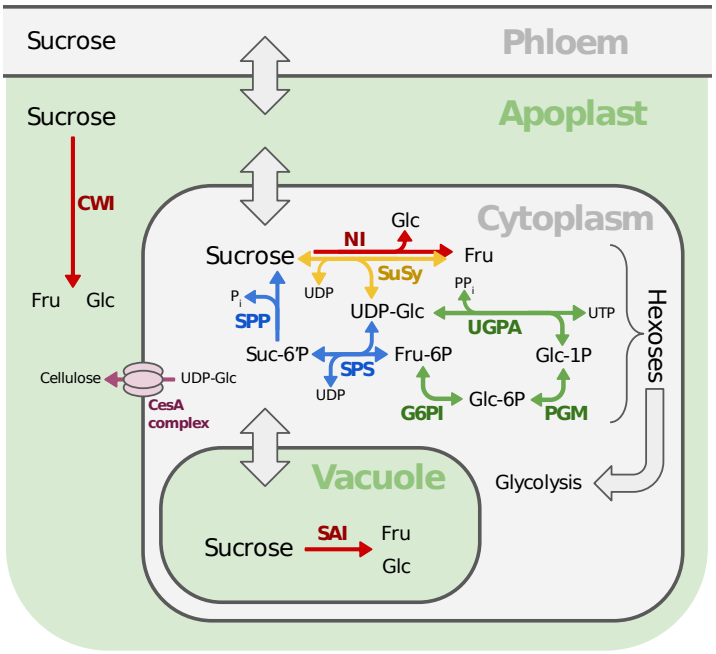


Figure 6: Schematic representation of biochemical reactions related to sucrose synthesis and breakdown at the parenchyma cells of sugarcane culms. Acronyms displayed in black are metabolites found in the cell, while those shown in bold and different colors represent the enzymes that catalyze each correspondent reaction. Arrows in red represent the invertase catalyzed reaction, *i.e.*, sucrose breakdown in different cell compartments. Arrows in yellow represent a pathway for reversible breakdown/synthesis of sucrose, providing UDP-glucose as a final product, which can be directly used for cellulose assembly or reused in the cycle. Although both SuSy and SPS are reversible catalyzers, SPP unidirectionally cleaves sucrose-phosphate into sucrose and pyrophosphate, orienting the blue reactions toward sucrose synthesis. Green arrows represent a pathway for conversion of fructose-6P to UDP-glucose, which can contribute to communicating metabolic substrates for the reactions of SuSy and SPS. Metabolites: Suc - sucrose, Glc - glucose, Fru - fructose, UTP - uridine triphosphate, UDP - uridine diphosphate, P_i - phosphate, PP_i – diphosphate. Enzymes: SAI - soluble acid invertase, NI - neutral invertase, CWI - cell wall invertase, SuSy - sucrose synthase, SPS - sucrose phosphate synthase, SPP - sucrose phosphatase, UGPA - UTP–glucose-1-phosphate uridylyltransferase, PGM - phosphoglucomutase, G6PI - glucose-6-phosphate isomerase.

genes from the same family or alleles separated in the assembly process. SAI expression levels were not consistent with those obtained by Jain *et al* (2017), because we could not establish any relationship between SAI expression and sugar content, which might be due to the early stage of development of our samples. Wang *et al* (2017) have identified and characterized several of the invertase genes in sugarcane, suggesting that NI had suffered a stronger conservation constraint during evolution, and it has a major enzymatic activity compared to the other invertases. Despite being at the core reactions of sugar conversion in sugarcane culms, the exact role of invertases in sucrose accumulation and sink strength is still unknown. Based on the expression patterns that we found, apical culms of distinct genotypes have particular profiles independent of their sugar content, at this early stage of development. Transgenic sugarcane was developed with a lower expression and activity of SAI, producing an elite cultivar with no significant difference in sugar yield than the non-modified plants (Botha *et al.*, 2001). On the other hand, a recent study showed that SAI silencing through the siRNA technique significantly increased the soluble solids content of transformed sugarcane plants (Khaled *et al.*, 2018). Both strategies should not be directly compared because they used distinct methodologies for gene suppression and for assessing the

results (enzymatic activity and expression fold change measured by qRT-PCR, respectively). As shown in (Figure 6), the final substrates of invertases are fructose and glucose, which cannot be reinserted in the sucrose breakdown/synthesis cycle directly. These monosaccharides must previously undergo conversion to any of the phosphorylated forms, which can be done via catalysis by kinases such as hexokinases.

Expression analysis shows SuSy2 as one of the SuSy genes related to top internodes and roots, suggesting that it has a role in sink strength in these tissues rather than a housekeeping function (Thirugnanasambandam *et al.*, 2019). The main activity of SuSy in young internodes happens toward sucrose synthesis instead of breakdown, because sucrose levels are lower in top culms (Schäfer *et al.*, 2004). However, breakdown catalysis is still present and is also more pronounced at the bottom of internodes, wherein more intense metabolic activity is detected (Rose & Botha, 2000). Moreover, SuSy was identified as a component of the catalytic unity associated with the cellulose synthase rosettes, using an immunoblotting assay (Fujii *et al.*, 2010). Thus, even with an overall balance towards sucrose synthesis, the microenvironment close to CesA complexes may demand a constant supply of UDP-glucose from SuSy. The higher expression level of SuSy in VLB may indicate a source for the UDP-glucose demand, which is the substrate for cellulose synthesis. The abundance of SPS and SPP transcripts were not statistically different in any of the analyzed contrasts, and this may be due to the function overlap with the synthetic activity of SuSy in young internodes. As internodes approach maturity and reach higher sucrose concentration, SPS surpasses SuSy in expression levels and activity, possibly because sucrose content drives SuSy enzymatic activity to breakdown instead of synthesis (Botha & Black, 2000; Schäfer *et al.*, 2004).

Inconsistencies among many sugarcane studies might be related to particularities in the expression patterns of distinct genotypes. As an example, Botha & Black (2000) showed opposing results and conclusions compared to Lingle & Irvine (1994) with regard to the role of SuSy in sucrose sink strength. Authors of the first study found no significant correlation between sugar accumulation and the activity of SuSy as the maturity of culms increased. Meanwhile, the others showed that SuSy activity is a reliable marker for sink strength due to a positive correlation between the same two variables. Botha & Black (2000) suggest that a reasonable explanation is due to inherent genotypic differences between these two studied sugarcane plants, which means that the found results are a product of a genotype-specific feature, instead of representing a conserved feature in sugarcane. This could as well be the reason for the differential expression of SuSy2 we observed, in agreement with results found by Thirugnanasambandam *et al.* (2017), who showed that lower sugar genotypes had higher expression levels of this gene.

Regarding the expression patterns of sucrose/hexose cycling enzymes (Figure 4 a) and their association with soluble solids content (Supplementary Figure 4 a), we can argue that differences in sugar content were not fully explained by transcriptional regulation of these enzymes in internode +1. We can consider this phenomenon as a likely genotype-specific expression, in which the experimental groups do not share a similar expression pattern. For example, SES205A showed greater expression levels for some of the CWI, NI, SPP, and PGM genes, which was not the case for the other VLB genotypes, even though it is phenotypically similar to IN84-58 and Krakatau. Another remarkable feature from the enzymes related to sugar processing is that most of them possess a large number of copies, such as CWI, NI, SuSy, SPS, and UGPA. This fact may reveal an expansion of the family members by gene duplication events or separation of alleles when assembling the transcriptome. Regarding the high ploidy levels found in sugarcane, it is feasible that it could harbor multiple alleles for each gene, besides allowing relaxed selection constraints due to the multiplicity of copies. The copies of genes mapped to the same annotation present a broad range of logCPM as well as different expression patterns among the genotypes, which may also be a case of genotype-specific or allele-specific expression.

We identified three KO terms directly related to carbon metabolism in plants by exploring the upregulated genes in VLB, namely photosynthesis (antenna proteins), carbon fixation in photosynthetic

organisms, and starch/sucrose metabolism. For the first, transcripts coding for chlorophyll-binding proteins from light-harvesting complexes I and II, which enable absorption of light with more efficiency, were more abundant in VLB. Next, we identified genes for enzymes of the Calvin-Benson and C4 cycles, responsible for the assimilation of carbon from atmospheric CO_2. Finally, the last pathway involves enzymatic conversions of sugars, for instance, isomerization of monosaccharides and synthesis/breakdown of polysaccharides. Thus, starch and sucrose metabolism represents a more complete set of the enzymatic reactions depicted in Figure 6. This sequence of KOs (organized as an ordered biological process) reveals a higher transcriptional activity of *S. spontaneum* regarding the acquisition and metabolism of sugars in the youngest internode of the sugarcane stalk. Previous studies have shown evidence that sink demands on culms can regulate the photosynthetic rates in leaves, mediated by local hexose concentration (McCormick *et al.*, 2006; McCormick *et al.*, 2008 a). A possible inference based on our data suggests that VLB genotypes have a higher demand for sugars, despite the low levels of accumulated sucrose in later stages of development. Besides, this supply-demand relationship can be controlled at a transcriptional level.

We identified cellulose synthase activity as one enriched molecular function in the contrast VLB x VHB.HB.LB (Figure 2 e). This implies that genes involved in cellulose synthesis were statistically more frequent than expected by chance among DEGs separating VLB from the other groups. Ding & Himmel (2006) proposed a model of CesA proteins spontaneously arranged in a rosette conformation formed by three distinct subunits of cellulose synthase. Of all the OsCesA genes we found in the transcriptome, 53% were differentially expressed with higher levels in VLB. This observation is consistent with the higher fiber content in VLB genotypes (Supplementary Figure 4). The hybrid R570, however, showed high expression levels of CesA genes despite its lower fiber content when compared to the VLB genotypes, which caused it to cluster with this group (Figure 4 b).

Casu *et al.* (2015) identified sugarcane homologues to *O. sativa* genes CesA1, -3, and -8 related to the synthesis of primary cell wall, while CesA4, -7, and -9 were associated with the secondary cell wall (Tanaka *et al.*, 2003). By analyzing both the average expression per gene and global expression regarding each CesA subunit, we found that the ones related to the primary cell wall deposition were among the most expressed, with ten upregulated genes in this set. The downregulated cellulose synthases, namely OsCesA2 and OsCesA6, have not yet been functionally characterized. Because we jointly sampled all tissues in the internode +1, the expression patterns represent a mosaic of mRNA from parenchyma, epidermis, and vascular tissues. We also identified several genes related to cell wall organization and remodeling upregulated in VLB. This fact reinforces the allocation of cellular resources, such as cellulose and other fibrous molecules, aiming to coordinate the expansion of cell walls in fiber-rich genotypes. Still, our results agree with more marked thickening of primary cell walls, as is expected for a young internode.

4.3 Sugar transporters role on sucrose storage

Proposing a model for sugar transport in plants is not a simple task due to the diversity of transmembrane transporter families, the multiple routes for import and export, and the existence of both mono and disaccharide substrates. Also, there is no single strategy of phloem unloading among plants, because they may vary in source/sink ratios and the presence of apoplasmic barriers (Milne *et al.*, 2018). Studies with dye localization in *Sorghum bicolor* show that transport can occur by different strategies depending on the context. Bihmidine *et al.* (2015) state that storage parenchyma cells are symplasmically isolated from sieve elements and companion cells, hence an apoplasmic step is required for transport between phloem and storage cells. On the other hand, Milne *et al.* (2015) point that apoplasmic unloading happens exclusively at meristematic and elongating portions of internodes. Secondary wall thickening, lignification, and suberization of elongated cells cause suppression of the apoplasmic flow,

and symplasmic transport must occur. Physiological evidence supports this hypothesis on sugarcane stems (Jacobsen *et al.*, 1992). Given that we analyzed the internode +1 from sugarcane stems, it is reasonable to admit an apoplasmic path for sucrose uptake, although plasmodesmal interconnections may also perform this task. Under this assumption, identifying the genes responsible for transport of sugars across membranes is essential to understand sucrose availability for storage and supply.

Immunolabeling and gene expression experiments suggest that OsSUT members 1 and 2 are involved in sucrose transport at the phloem and storage parenchyma vacuoles, respectively (Rae *et al.*, 2005; Casu *et al.*, 2015; Eom *et al.*, 2011). Because SUT acts as a sucrose/H^+ symporter, the transport flow depends on a pH gradient, promoting sugar export from phloem to apoplasm and from vacuole to cytoplasm in storage parenchyma. Considering this model for sugar movement, the high expression of OsSUT2 inhibits sucrose accumulation in vacuoles, which would be appropriate for cellular metabolism and cell wall synthesis. Expression data analysis for OsSUT1 suggests a regulation role of cell wall thickening, due to it overexpression in the VLB group. This fact agrees with the expression patterns of both SuSy2 and CesA genes. We also identified four expressed OsSUT4 genes in sugarcane transcriptome, three of which were significantly differentially expressed in the VLB x VHB.HB.LB contrast. This result agrees with the findings of Zhang *et al.* (2016 b), that the set of SUT4 genes is more highly expressed in *S. spontaneum* than in *S. officinarum*. Proteomic and green fluorescent protein reporter approaches have characterized SUT4 as a vacuole transporter in Arabidopsis and barley (Endler *et al.*, 2006).

Studies suggest that SWEETs originated from gene duplication or fusion of SemiSWEET genes from prokaryotes, which has three transmembrane helices (Hu *et al.*, 2016). Those genes can also transport sugar when in a dimer, forming a pore wherein sugar can be transported (Chen *et al.*, 2010; Xu *et al.*, 2014; Tao *et al.* 2015). Despite the presence of transcripts coding for products with seven transmembrane helices in all SWEET genes, we found in the assembled transcriptome putative isoforms for both SWEET2b and SWEET16 with a lower number of helices and a signal peptide. In view of SemiSWEET structures, it is feasible to suggest that variant splicing isoforms could have a functional role in sugar transport in sugarcane. Moreover, the presence of a signal peptide in these variant isoforms indicates that the protein products could have a different cellular destination than their complete counterparts. The existence of novel splicing variants of both SWEET genes might be related to the sucrose accumulation processes in sugarcane. Also, it could offer a new perspective on the evolutionary dynamics of the SWEET family.

From the set of key candidate genes responsible for sugar accumulation, OsSWEET4 stands out due to the abundance of its transcript in all genotypes. The strong expression pattern of SWEET4 is also found in *S. bicolor* gene SbSWEET4-3, which is upregulated in stems (Mizuno *et al.*, 2016). Genomic information from maize, sorghum, and rice suggests that SWEET4 underwent two duplications after the divergence of rice from the two other grasses, indicating that this gene might also have multiple copies in sugarcane (Mizuno *et al.*, 2016). Therefore, SbSWEET4-3 homologue(s) in sugarcane probably play a central role in sugar transport in culms. In terms of differential expression, OsSWEET2b was the only SWEET upregulated in the VLB group. Also, OsSWEET2b promotes glucose uptake into the vacuole (Tao *et al.*, 2015). This fact suggests that *S. spontaneum* genotypes might store glucose in vacuoles at a higher concentration than high sugar *S. officinarum* and hybrids in young culms.

Besides SUT and SWEET, other genes have been reported as sugar transporters in vacuoles, such as early response to dehydratation 6 like family, tonoplast sugar transporters (previously called tonoplast monosaccharide transporters), and vacuolar glucose transporter (Hedrich *et al.*, 2015). Two members of the ERD6-like family, ESL1 and ERDL6, were first identified as hexose exporters from the vacuole in *A. thaliana* (Yamada *et al.*, 2010; Poschet *et al.*, 2011). Despite that, knowledge about this gene family has not increased substantially in the last decade. The assembled transcriptome has about 11 homologues of Arabidopsis ERD6-like annotated genes. Three of them belong to the ERDL6 family, known as vacuole glucose/H^+ symporter. We found one AtERDL6 gene with high transcripts, and it is

upregulated in VLB. AtERDL16 was found in *S. spontaneum* at very low expression levels, resulting in a 130-fold difference between groups, but it did not have its function explained until now.

4.4 Transposable elements and genotype relatedness

Genomic studies carried out with hybrids and *S. spontaneum* suggest that the wild accessions harbor more transposon genes than R570 and SP80-3280 (Souza *et al.*, 2019). However, a complete understanding of the distribution of transposable elements in the genomes of sugarcane hybrids depends on a full genome assembly of these genotypes. The currently available references rely on homology to the sorghum genome, as is the case of R570 (Garsmeur *et al.*, 2018), or on long reads, used in the SP80-3280 gene space assembly, which tend to underrepresent high copy regions, such as insertions of transposable elements (Souza *et al.*, 2019).

The expression patterns of transposase activity-annotated genes (Figures 2 b and 5) reveal that they might be more active in the groups of higher sugar content. Of the nine combined genotypes in VHB, HB, and LB, five are interspecific hybrids, and the other four represent *S. officinarum* genotypes. A comparative study using BACs from both species showed that *S. officinarum* has a higher number of base pairs masked as sequences from transposons and retrotransposons. On the other hand, insertions of LTR retrotransposons in *S. spontaneum* are more recent, in general (Zhang *et al.*, 2016 a). However, we cannot interpret this as evidence of more intense gene expression of transposable elements in *S. spontaneum*. In general, gene expression is required for transposition activity, but it does not guarantee its occurrence. It can be argued that recent hybridization, in terms of sexual reproduction generations, might have a contribution to the overexpression of transposase genes in the high sugar content groups (de Araujo *et al.*, 2005).

5 CONCLUSIONS

In this study, we investigated possible genetic reasons that explain sugarcane variation in terms of sucrose accumulation, which can reach the highest concentration known in a plant. More specifically, we focused on understanding how regulation of gene expression in apical culms could be responsible for differences in sugar storage in the target tissue. With the information acquired from the set of differentially expressed genes, we observed that several genes pointed as main regulators of sucrose yield were not significantly upregulated in the group of sugar-rich genotypes. Despite the complexity of factors that might contribute to this phenomenon, our analyses suggest that even at this early development stage, cellulose synthesis plays a vital role in the differentiation of sugarcane genotypes with lower and higher levels of sucrose and fiber. Also, SuSy2 was among the few differentially expressed genes responsible for sucrose synthesis/breakdown, and it might be useful to provide the required substrate for cellulose synthetic activity. The transport of sugars by transmembrane proteins was a noticeable process that distinguished samples with very low sugar concentration from the others. We also found evidence of a novel alternative splicing form in two SWEET genes expressed in the internode +1, which encodes a protein isoform with a different number of transmembrane helices. This discovery could bring new information to the sugar transport process in sugarcane. However, for a better glimpse of sugar transporters in general, it is necessary to acquire both tissue-specific expression patterns and cellular localization of these proteins to understand the exact trafficking of photoassimilates from source to sink. As a whole, gene expression patterns indicate a route for carbon partitioning in these young culms, provided by *i)* driving sucrose to the cytoplasm instead of apoplasm or vacuole; *ii)* breaking down of sucrose according to the metabolic demand; and *iii)* synthesizing cellulose for cell wall expansion. This molecular workflow may partially explain the relationship between fiber and sucrose yield that distinguishes *S. officinarum* and *S. spontaneum* phenotypes. Remarkably, our study pointed to several particularities possessed by each genotype in expression levels. This result simultaneously shows that sugar yield depends on multiple genes, and that genotypes with similar phenotypes might not have common grounds when it comes to expression profiles of well-established genes as the main regulators of the sugar accumulation process, at an early stage of development.

6 ABBREVIATIONS

ABA - Abscisic acid, CesA - Cellulose synthase A, CPM - Counts per million, CWI - Cell wall invertase, DEG - Differentially expressed gene, ERDL - Sugar transporter early response to dehydration 6-like, EST - Expressed sequence tag, FC - Fold-change, FDR - False discovery rate, G6PI - glucose-6-phosphate isomerase, GO - Gene ontology, HB - High ºbrix, KO - KEGG orthology, LB - Low ºbrix, NI - neutral invertase, ORF - Open reading frame, PCA - Principal component analysis, PGM - phosphoglucomutase, SAI - soluble acid invertase, SI - Sucrose isomerase, SNP - single nucleotide polymorphism, SPP - sucrose phosphatase, SPS - sucrose phosphate synthase, SuSy - sucrose synthase, SUT - Sucrose transport protein, SWEET - Sugars Will Eventually be Exported Transporter (Bidirectional sugar transporter), TPM - Transcript per million, TST - Tonoplast sugar transporter, UGPA - UTP–glucose-1-phosphate uridylyltransferase, UDP - uridine diphosphate, VGT - Vacuolar glucose transporter, VHB - Very high ºbrix, VLB - Very low ºbrix.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology, 215*(3), 403-410.

de Araujo, P. G., Rossi, M., de Jesus, E. M., Saccaro Jr, N. L., Kajihara, D., Massa, R., ... & Ulian, E. C. (2005). Transcriptionally active transposable elements in recent hybrid sugarcane. *The Plant Journal, 44*(5), 707-717.

Barreto, F. Z., Rosa, J. R. B. F., Balsalobre, T. W. A., Pastina, M. M., Silva, R. R., Hoffmann, H. P., ... & Carneiro, M. S. (2019). A genome-wide association study identified loci for yield component traits in sugarcane (Saccharum spp.). *PloS one, 14*(7), e0219843.

de Barros Dantas, L. L., Almeida-Jesus, F. M., de Lima, N. O., Alves-Lima, C., Nishiyama-Jr, M. Y., Carneiro, M. S., ... & Hotta, C. T. (2020). Rhythms of transcription in field-grown sugarcane are highly organ specific. *Scientific reports, 10*(1), 1-12.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Bihmidine, S., Baker, R. F., Hoffner, C., & Braun, D. M. (2015). Sucrose accumulation in sweet sorghum stems occurs by apoplasmic phloem unloading and does not involve differential Sucrose transporter expression. *BMC plant biology, 15*(1), 186.

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., & Hochreiter, S. (2015). msa: an R package for multiple sequence alignment. *Bioinformatics, 31*(24), 3997-3999.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Botha, F. C., & Black, K. G. (2000). Sucrose phosphate synthase and sucrose synthase activity during maturation of internodal tissue in sugarcane. *Functional Plant Biology, 27*(1), 81-85.

Botha, F. C., Sawyer, B. J. B., & Birch, R. G. (2001). Sucrose metabolism in the culm of transgenic sugarcane with reduced soluble acid invertase activity. In *Proceedings of the International Society of Sugar Cane Technologists* (Vol. 24, pp. 588-591).

Bremer, G. (1925). The cytology of the sugarcane. *Genetica, 7*(3), 293-322.

Casu, R. E., Dimmock, C. M., Chapman, S. C., Grof, C. P., McIntyre, C. L., Bonnett, G. D., & Manners, J. M. (2004). Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant molecular biology, 54*(4), 503-517.

Casu, R. E., Grof, C. P., Rae, A. L., McIntyre, C. L., Dimmock, C. M., & Manners, J. M. (2003). Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. *Plant molecular biology, 52*(2), 371-386.

Casu, R. E., Rae, A. L., Nielsen, J. M., Perroux, J. M., Bonnett, G. D., & Manners, J. M. (2015). Tissue-specific transcriptome analysis within the maturing sugarcane stalk reveals spatial regulation in the expression of cellulose synthase and sucrose transporter gene families. *Plant molecular biology, 89*(6), 607-628.

Chen, L. Q., Hou, B. H., Lalonde, S., Takanaga, H., Hartung, M. L., Qu, X. Q., ... & Chermak, D. (2010). Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature, 468*(7323), 527-532.

Ding, S. Y., & Himmel, M. E. (2006). The maize primary cell wall microfibril: a new model derived from direct visualization. *Journal of Agricultural and Food Chemistry, 54*(3), 597-606.

D'Hont, A., Grivet, L., Feldmann, P., Glaszmann, J. C., Rao, S., & Berding, N. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Molecular and General Genetics MGG, 250*(4), 405-413.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Informatics 2009: Genome Informatics Series Vol. 23* (pp. 205-211).

Endler, A., Meyer, S., Schelbert, S., Schneider, T., Weschke, W., Peters, S. W., ... & Schmidt, U. G. (2006). Identification of a vacuolar sucrose transporter in barley and Arabidopsis mesophyll cells by a tonoplast proteomic approach. *Plant Physiology, 141*(1), 196-207.

Eom, J. S., Chen, L. Q., Sosso, D., Julius, B. T., Lin, I. W., Qu, X. Q., ... & Frommer, W. B. (2015). SWEETs, transporters for intracellular and intercellular sugar translocation. *Current Opinion in Plant Biology, 25*, 53-62.

Eom, J. S., Cho, J. I., Reinders, A., Lee, S. W., Yoo, Y., Tuan, P. Q., ... & Bhoo, S. H. (2011). Impaired function of the tonoplast-localized sucrose transporter in rice, OsSUT2, limits the transport of vacuolar reserve sucrose and affects plant growth. *Plant Physiology, 157*(1), 109-119.

Fujii, S., Hayashi, T., & Mizuno, K. (2010). Sucrose synthase is an integral component of the cellulose synthesis machinery. *Plant and cell physiology, 51*(2), 294-301.

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., ... & Costet, L. (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature communications, 9*(1), 2638.

Gazaffi, R., Cursi, D. E., Chapola, R. G., Santos, J. D., Fernandes, A. R., Carneiro, M. S., & Charoenwong, S. (2016). RB varieties: A major contribution to the sugarcane industry in Brazil. *Proceedings of the 29th International Society of Sugar Cane Technologists, Chang Mai, Thailand. Dec. 2016*, 1677-1682.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology, 29*(7), 644.

Hedrich, R., Sauer, N., & Neuhaus, H. E. (2015). Sugar transport across the plant vacuolar membrane: nature and regulation of carrier proteins. *Current opinion in plant biology, 25*, 63-70.

Heinz, D. J. (1991). Sugarcane cytogenetics. *Chromosome engineering in plants: Genetics, Breeding, Evolution, Part B. Elsevier, Amsterdam*, 279-293.

Hu, Y. B., Sosso, D., Qu, X. Q., Chen, L. Q., Ma, L., Chermak, D., ... & Frommer, W. B. (2016). Phylogenetic evidence for a fusion of archaeal and bacterial SemiSWEETs to form eukaryotic SWEETs and identification of SWEET hexose transporters in the amphibian chytrid pathogen Batrachochytrium dendrobatidis. *The FASEB Journal, 30*(10), 3644-3654.

Jacobsen, K. R., Fisher, D. G., Maretzki, A., & Moore, P. H. (1992). Developmental changes in the anatomy of the sugarcane stem in relation to phloem unloading and sucrose storage. *Botanica Acta, 105*(1), 70-80.

Jain, R., Singh, S. P., Singh, A., Singh, S., Kishor, R., Singh, R. K., ... & Solomon, S. (2017). Soluble Acid Invertase (SAI) Activity and Gene Expression Controlling Sugar Composition in Sugarcane. *Sugar Tech, 19*(6), 669-674.

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., & Gao, G. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic acids research*, gkw982.

Khaled, S. K., Abdel-Tawab, F. M., Fahmy, E. M., Amer, E. A. M., & Khaled, K. A. (2018). The siRNA efficacy of soluble acid invertase down-regulation in Sugarcane (Saccharum spp.). *Arab Universities Journal of Agricultural Sciences, 26*(Special issue (2C)), 2011-2017.

Lembke, C. G., Nishiyama, M. Y., Sato, P. M., de Andrade, R. F., & Souza, G. M. (2012). Identification of sense and antisense transcripts regulated by drought in sugarcane. *Plant molecular biology, 79*(4-5), 461-477.

Lingle, S. E., & Irvine, J. E. (1994). Sucrose synthase and natural ripening in sugarcane. *Crop Science, 34*(5), 1279-1283.

Makarevitch, I., & Harris, C. (2010). Aneuploidy causes tissue-specific qualitative changes in global gene expression patterns in maize. *Plant physiology, 152*(2), 927-938.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal, 17*(1), 10-12.

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research, 40*(10), 4288-4297.

McCormick, A. J., Cramer, M. D., & Watt, D. A. (2006). Sink strength regulates photosynthesis in sugarcane. *New Phytologist, 171*(4), 759-770.

McCormick, A. J., Cramer, M. D., & Watt, D. A. (2008). Regulation of photosynthesis by sugars in sugarcane leaves. *Journal of plant physiology, 165*(17), 1817-1829.

McCormick, A. J., Watt, D. A., & Cramer, M. D. (2008). Supply and demand: sink regulation of sugar accumulation in sugarcane. *Journal of experimental botany, 60*(2), 357-364.

Milne, R. J., Grof, C. P., & Patrick, J. W. (2018). Mechanisms of phloem unloading: shaped by cellular pathways, their conductances and sink function. *Current opinion in plant biology, 43*, 8-15.

Milne, R. J., Offler, C. E., Patrick, J. W., & Grof, C. P. (2015). Cellular pathways of source leaf phloem loading and phloem unloading in developing stems of Sorghum bicolor in relation to stem sucrose storage. *Functional plant biology, 42*(10), 957-970.

Mizuno, H., Kasuga, S., & Kawahigashi, H. (2016). The sorghum SWEET gene family: stem sucrose accumulation as revealed through transcriptome profiling. *Biotechnology for biofuels, 9*(1), 127.

Moore, P. H. (1995). Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. *Functional Plant Biology, 22*(4), 661-679.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research, 35*(suppl 2), W182-W185.

Morris, D. R., Gilbert, R. A., Rainbolt, C. R., Perdomo, R. E., Powell, G., Eiland, B., & Montes, G. (2007). Sugarcane yields and soil chemical properties due to mill mud applications to a sandy soil. In *Proc. Int. Soc. Sugar Cane Technol* (Vol. 26, pp. 444-448).

Panje, R. R., & Babu, C. N. (1960). Studies in Saccharum spontaneum distribution and geographical association of chromosome numbers. *Cytologia, 25*(2), 152-172.

Papini-Terzi, F. S., Rocha, F. R., Vêncio, R. Z., Felix, J. M., Branco, D. S., Waclawovsky, A. J., ... & Vicentini, R. (2009). Sugarcane genes associated with sucrose content. *BMC genomics, 10*(1), 120.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods, 14*(4), 417.

Poschet, G., Hannich, B., Raab, S., Jungkunz, I., Klemens, P. A., Krueger, S., ... & Büttner, M. (2011). A novel Arabidopsis vacuolar glucose exporter is involved in cellular sugar homeostasis and affects the composition of seed storage compounds. *Plant physiology, 157*(4), 1664-1676.

Rae, A. L., Perroux, J. M., & Grof, C. P. (2005). Sucrose partitioning between vascular bundles and storage parenchyma in the sugarcane stem: a potential role for the ShSUT1 sucrose transporter. *Planta, 220*(6), 817-825.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140.

Rooney, W. L., Blumenthal, J., Bean, B., & Mullet, J. E. (2007). Designing sorghum as a dedicated bioenergy feedstock. *Biofuels, Bioproducts and Biorefining, 1*(2), 147-157.

Rose, S., & Botha, F. C. (2000). Distribution patterns of neutral invertase and sugar content in sugarcane internodal tissues. *Plant Physiology and Biochemistry, 38*(11), 819-824.

Sacher, J. A., Hatch, M. D., & Glasziou, K. T. (1963). Sugar accumulation cycle in sugar cane. III. Physical & metabolic aspects of cycle in immature storage tissues. *Plant Physiology, 38*(3), 348.

Schäfer, W. E., Rohwer, J. M., & Botha, F. C. (2004). Protein-level expression and localization of sucrose synthase in the sugarcane culm. *Physiologia Plantarum, 121*(2), 187-195.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212.

Souza, G. M., Van Sluys, M. A., Lembke, C. G., Lee, H., Margarido, G. R. A., Hotta, C. T., ... & Nishiyama Jr, M. Y. (2019). Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience, 8*(12), giz129.

Tanaka, K., Murata, K., Yamazaki, M., Onosato, K., Miyao, A., & Hirochika, H. (2003). Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. *Plant physiology, 133*(1), 73-83.

Tao, Y., Cheung, L. S., Li, S., Eom, J. S., Chen, L. Q., Xu, Y., ... & Feng, L. (2015). Structure of a eukaryotic SWEET transporter in a homotrimeric complex. *Nature, 527*(7577), 259-263.

Thirugnanasambandam, P. P., Hoang, N. V., Furtado, A., Botha, F. C., & Henry, R. J. (2017). Association of variation in the sugarcane transcriptome with sugar content. *BMC genomics, 18*(1), 909.

Thirugnanasambandam, P. P., Mason, P. J., Hoang, N. V., Furtado, A., Botha, F. C., & Henry, R. J. (2019). Analysis of the diversity and tissue specificity of sucrose synthase genes in the long read transcriptome of sugarcane. *BMC plant biology, 19*(1), 160.

Vicentini, R., Bottcher, A., dos Santos Brito, M., dos Santos, A. B., Creste, S., de Andrade Landell, M. G., ... & Mazzafera, P. (2015). Large-scale transcriptome analysis of two sugarcane genotypes contrasting for lignin content. *PloS one, 10*(8).

Wang, J., Nayak, S., Koch, K., & Ming, R. (2013). Carbon partitioning in sugarcane (Saccharum species). *Frontiers in plant science, 4*, 201.

Wang, L., Zheng, Y., Ding, S., Zhang, Q., Chen, Y., & Zhang, J. (2017). Molecular cloning, structure, phylogeny and expression analysis of the invertase gene family in sugarcane. *BMC plant biology, 17*(1), 109.

Wu, L., & Birch, R. G. (2007). Doubled sugar content in sugarcane plants modified to produce a sucrose isomer. *Plant biotechnology journal, 5*(1), 109-117.

Xu, Y., Tao, Y., Cheung, L. S., Fan, C., Chen, L. Q., Xu, S., ... & Feng, L. (2014). Structures of bacterial homologues of SWEET transporters in two distinct conformations. *Nature, 515*(7527), 448-452.

Yamada, K., Osakabe, Y., Mizoi, J., Nakashima, K., Fujita, Y., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2010). Functional analysis of an Arabidopsis thaliana abiotic stress-inducible facilitated diffusion transporter for monosaccharides. *Journal of Biological Chemistry, 285*(2), 1138-1146.

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology, 11*(2), R14.

Zhang, J., Sharma, A., Yu, Q., Wang, J., Li, L., Zhu, L., ... & Ming, R. (2016). Comparative structural analysis of Bru1 region homeologs in Saccharum spontaneum and S. officinarum. *BMC genomics, 17*(1), 446.

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., ... & Wai, C. M. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nature genetics, 50*(11), 1565.

Zhang, Q., Hu, W., Zhu, F., Wang, L., Yu, Q., Ming, R., & Zhang, J. (2016). Structure, phylogeny, allelic haplotypes and expression of sucrose transporter gene families in Saccharum. *BMC genomics, 17*(1), 88.

SUPPLEMENTARY

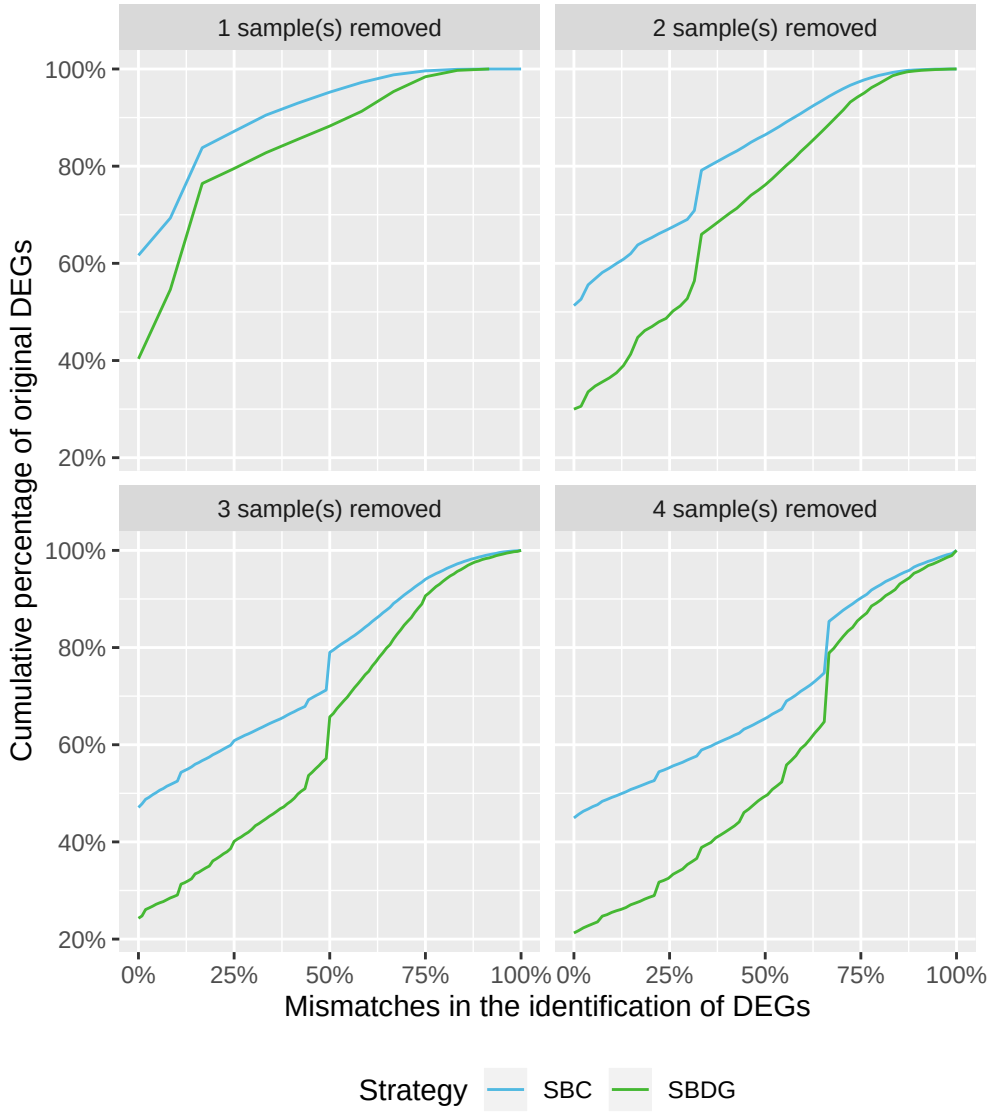


Figure 1: Cumulative distribution of identified mismatches. The total number of combinations of subsets of samples are: $n_1 = 12$, $n_2 = 54$, $n_3 = 108$, and $n_4 = 81$, in which the index i represents the number of removed samples. Thus, the original DEGs and the subsamples may match from 0 to n_i times. We consider a match when a given iteration has the same differential expression result for a gene. Hence, the curves indicate the percentage of DEGs which showed the minimum mismatch rate in the x-axis. For example, with a single sample removed for the SBDG (top left panel), 40% of genes did not have a single mismatch for all twelve combinations, and 80% of genes presented a maximum mismatch rate of 25%.

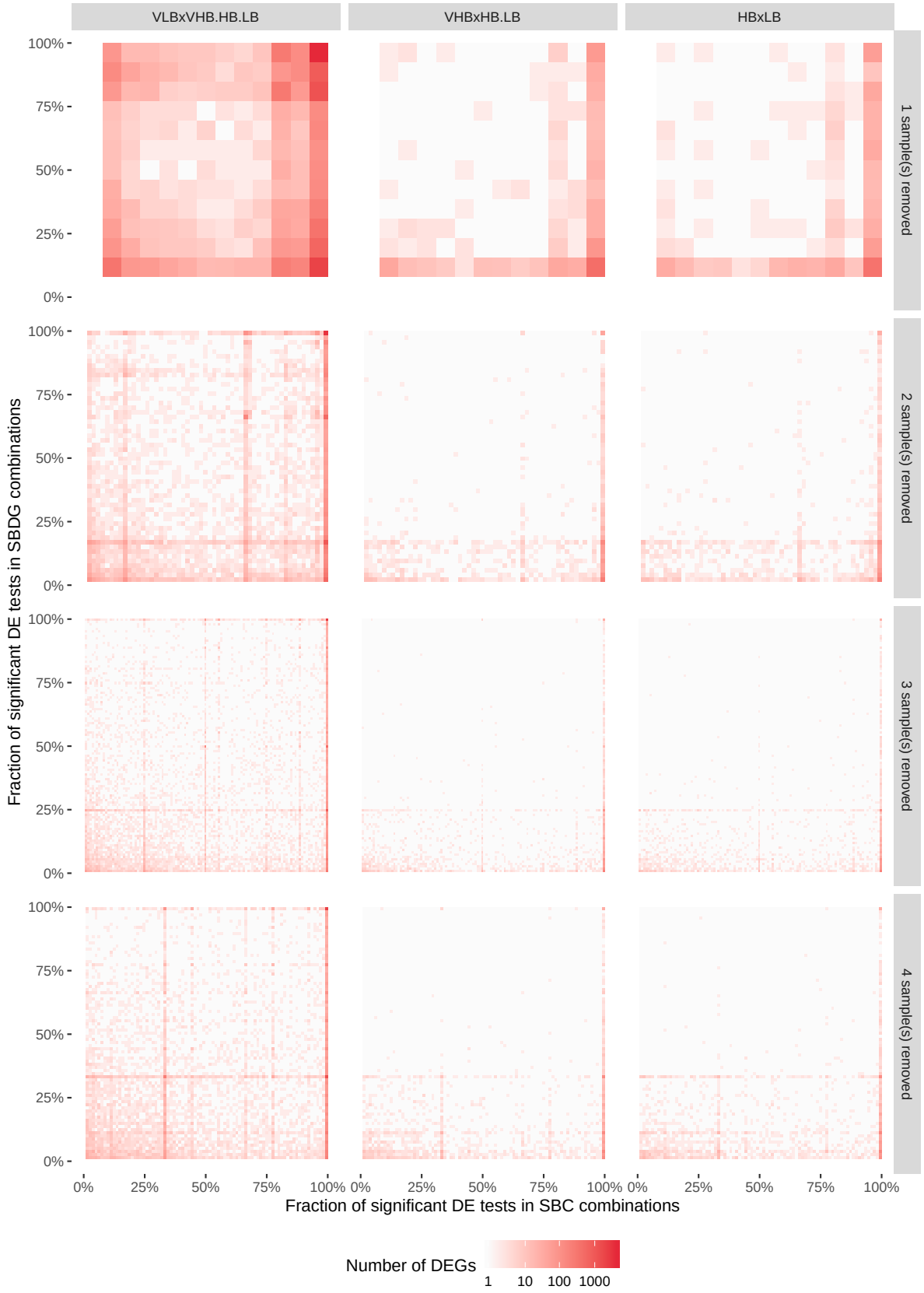


Figure 2: Correspondence between differentially expressed genes found by SBC and SBDG per contrast and number of removed samples. The number of combinations for i removed sample(s) is n_i, such that each gene ranges from 0 to n_i chances to be called as differentially expressed. The heatmap shows the number of DEGs identified by both strategies, for at least one combination of samples, normalized by n_i.

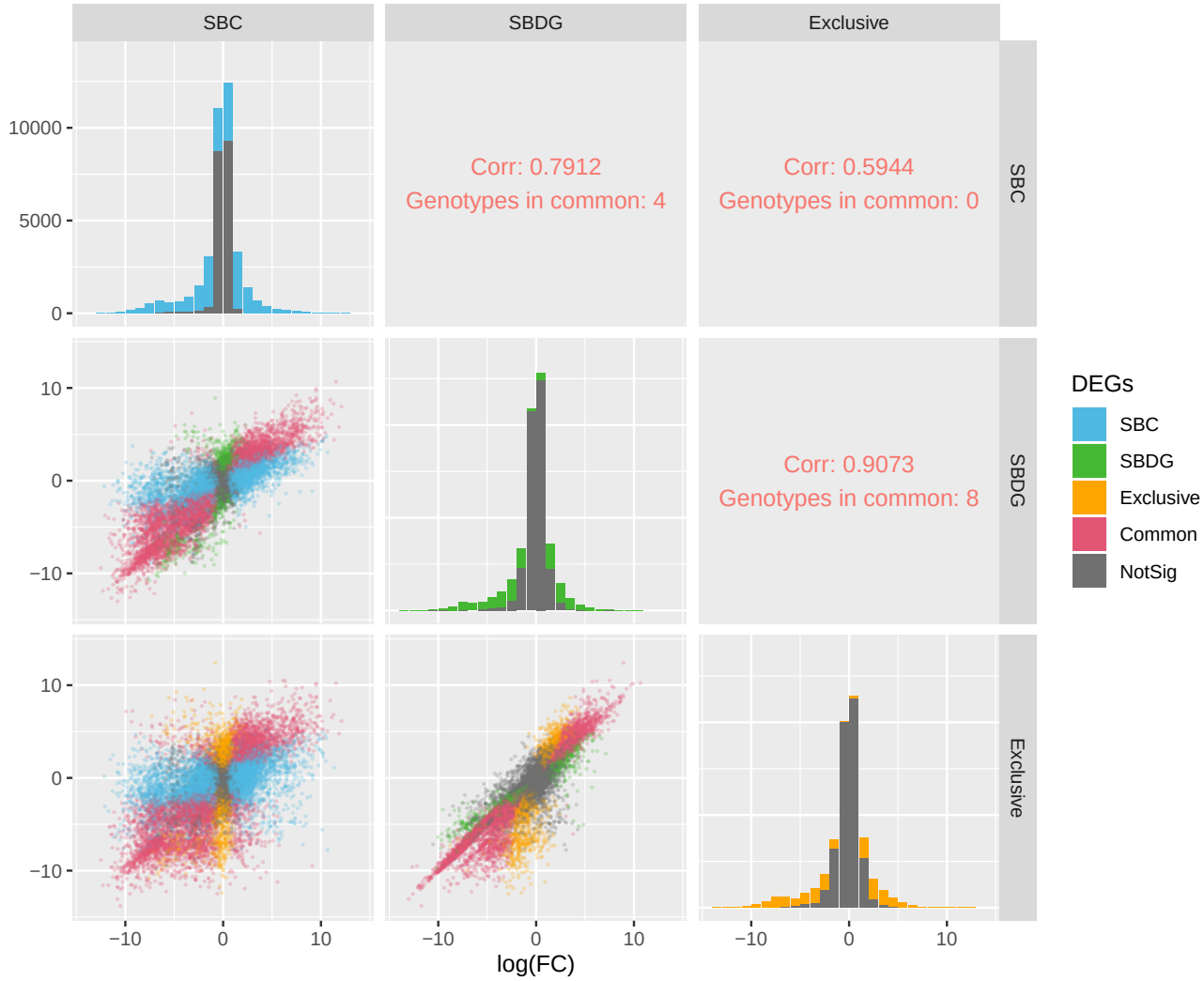
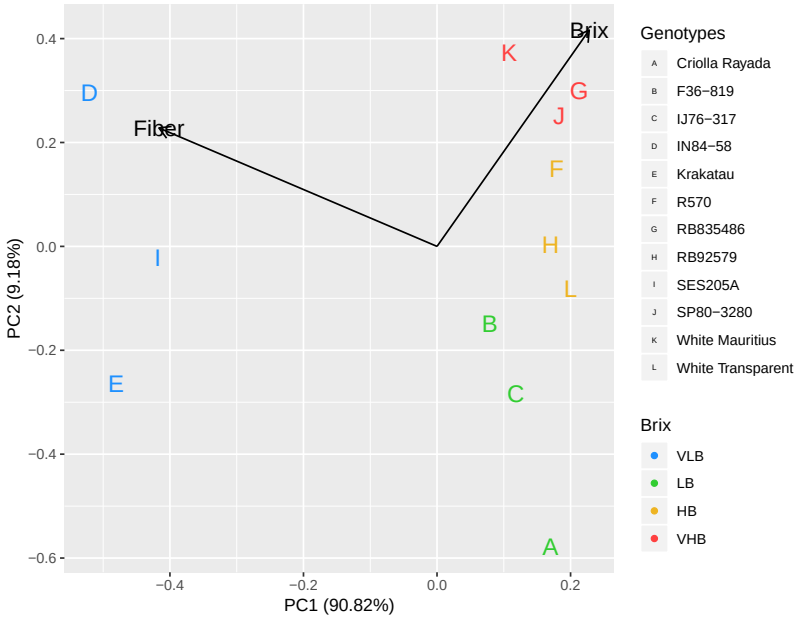
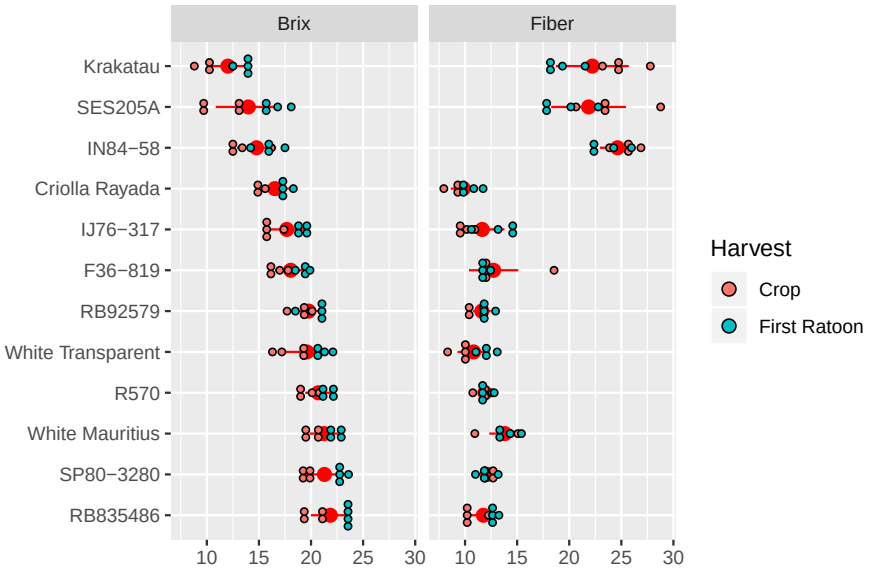


Figure 3: Correlation of log fold changes among the two strategies and the set of SBDG-exclusive samples, for contrast VLB x VHB.HB.LB. In the main diagonal, the histograms show how the distribution of differentially expressed genes (DEGs) and non-DEGs, as a function of the log fold change (logFC). The classification of genes considered the result of differential expression for all pairwise groups of strategies (SBC, SBDG, and SBDG-exclusive genotypes). Pink points represent genes with a significant test for a given pair of groups, and gray points represent non-significant genes for the same pair. We only considered the genes passing the low expression filtering criterion in all sets (38,420 genes).



(a)



(b)

Figure 4: Phenotypic characterization of contrasting genotypes used for transcriptional profiling. (A) Principal Component Analysis biplot for fiber percentage and content of soluble solids (measured in ºBrix) data separates the VLB group from the others in the first component, while the second component separates the remaining three groups. (B) The panels show values of the content of soluble solids and fiber percentage from individual plants, displayed by increasing average of soluble solids. VLB - Very Low ºBrix, LB - Low ºBrix, HB - High ºBrix, and VHB - Very High ºBrix.

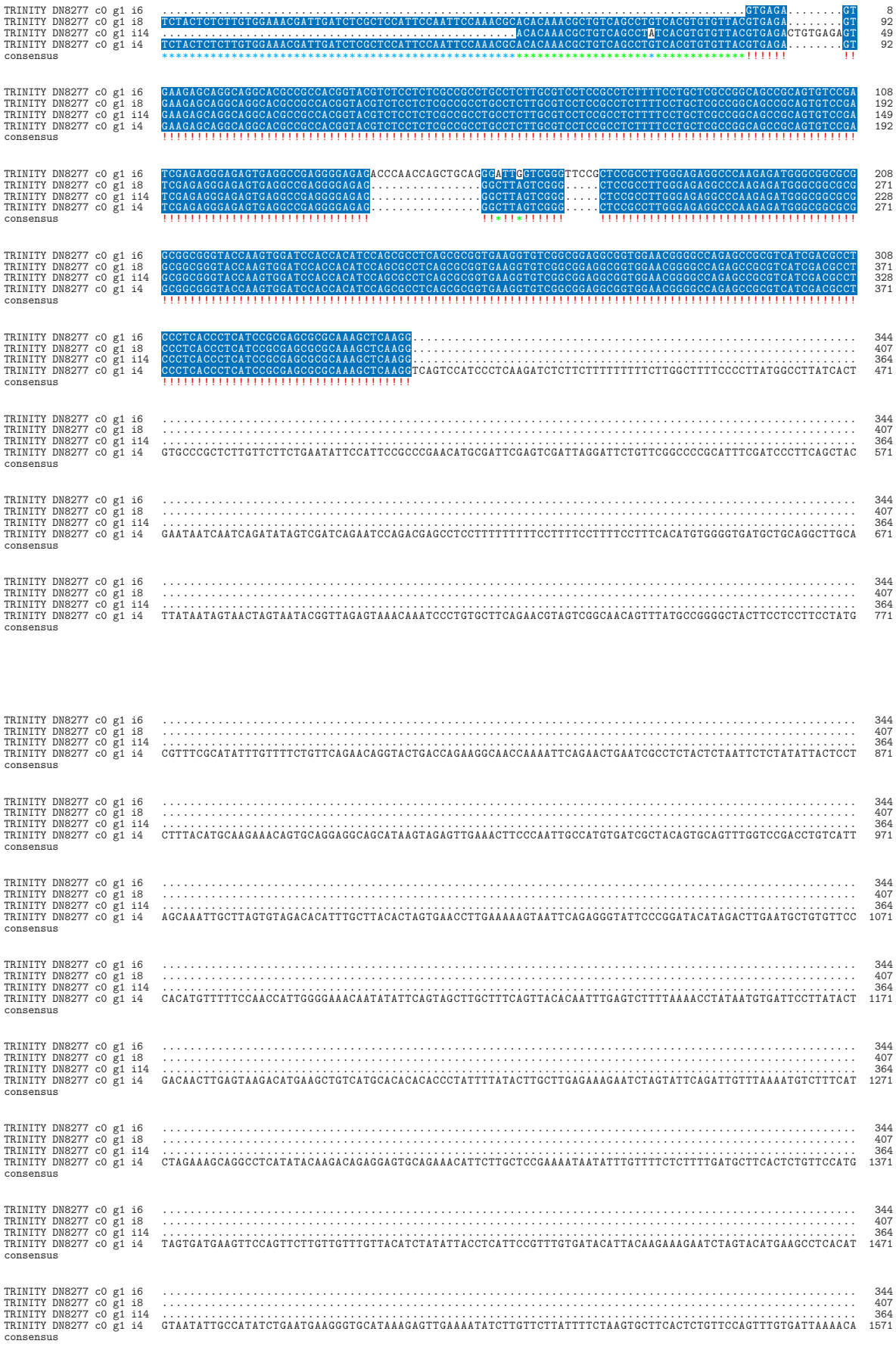

TRINITY DN8277 c0 g1 i6
TRINITY DN8277 c0 g1 i8
TRINITY DN8277 c0 g1 i14
TRINITY DN8277 c0 g1 i4
consensus

Figure 5: (a)

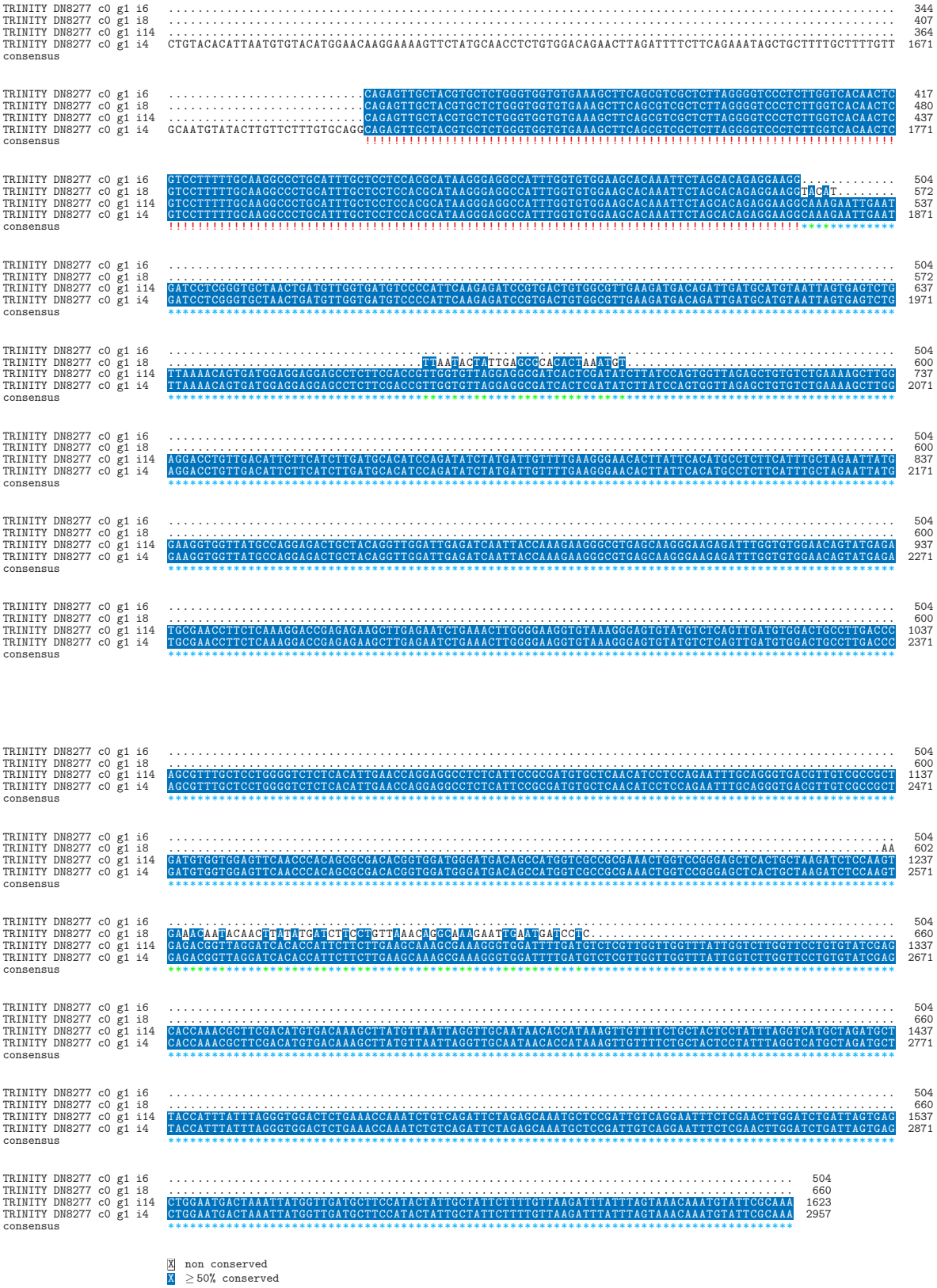

TRINITY DN8277 c0 g1 i6
TRINITY DN8277 c0 g1 i8
TRINITY DN8277 c0 g1 i14
TRINITY DN8277 c0 g1 i4
consensus
X non conserved
X ≥50% conserved

Figure 5: (a)

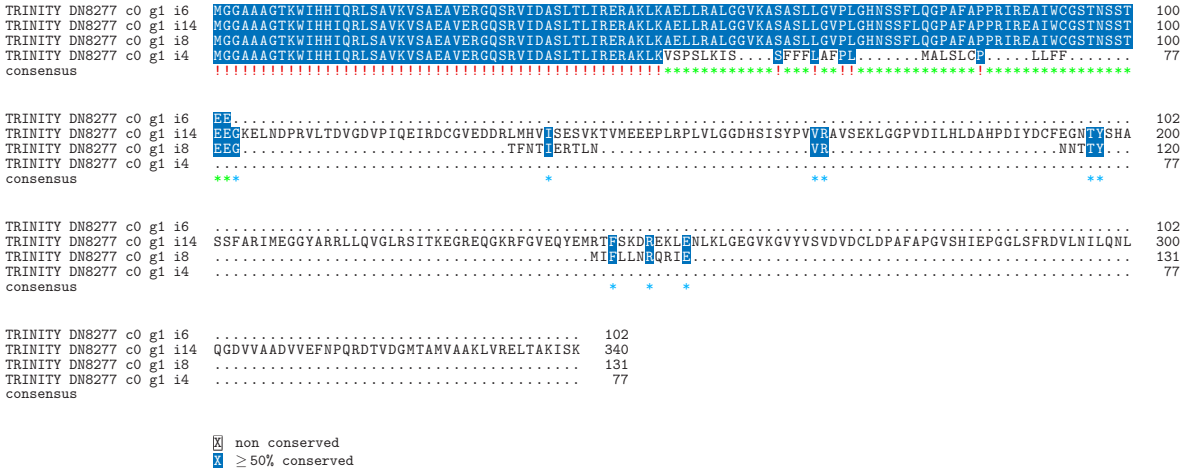
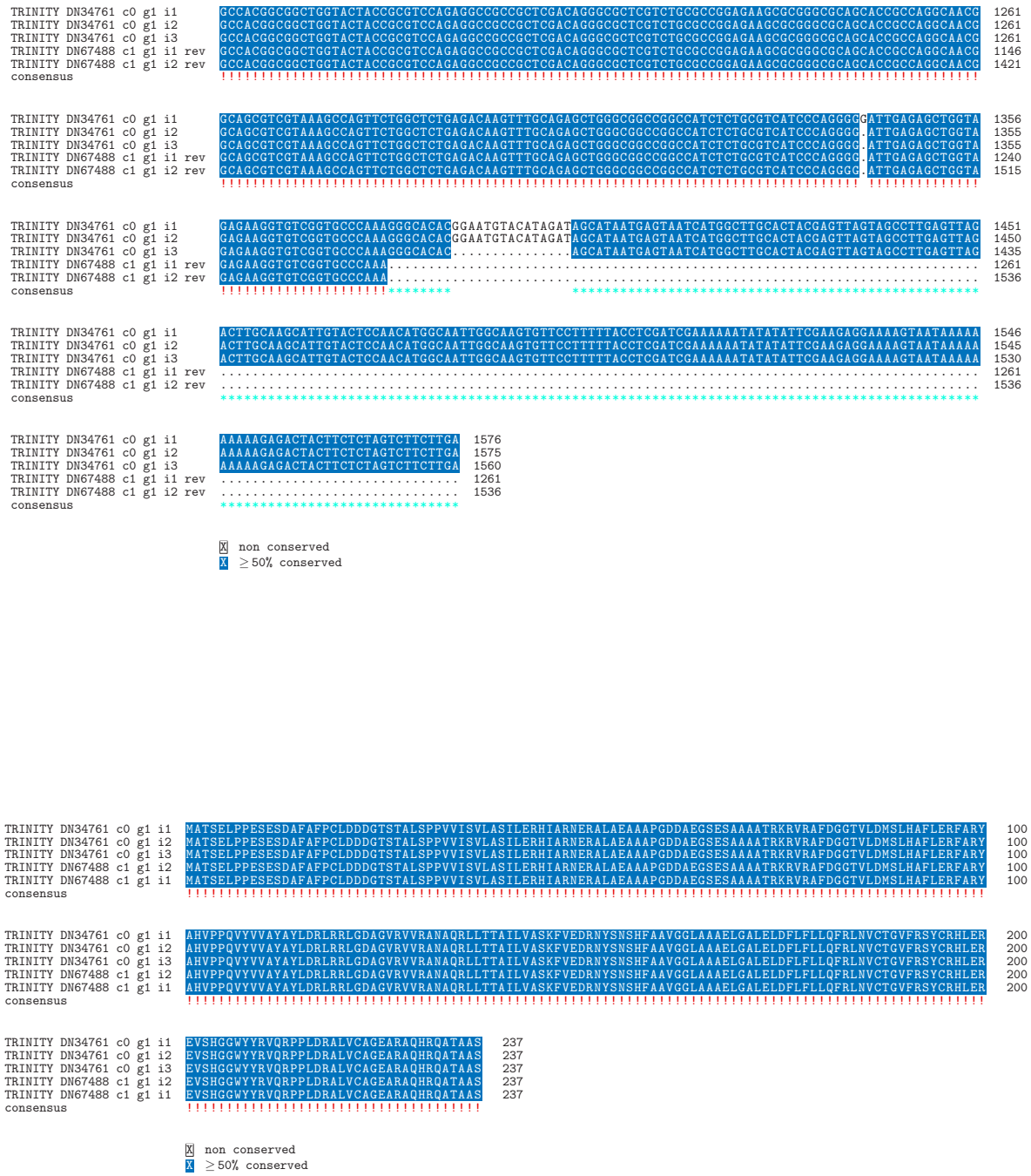

TRINITY DN8277 c0 g1 i6
TRINITY DN8277 c0 g1 i14
TRINITY DN8277 c0 g1 i8
TRINITY DN8277 c0 g1 i4
consensus
X non conserved
X ≥50% conserved

Figure 5: (a)

Figure 5: (b)


TRINITY DN34761 c0 g1 i1
TRINITY DN34761 c0 g1 i2
TRINITY DN34761 c0 g1 i3
TRINITY DN67488 c1 g1 i1 rev
TRINITY DN67488 c1 g1 i2 rev
consensus
X non conserved
X ≥50% conserved

TRINITY DN34761 c0 g1 i1
TRINITY DN34761 c0 g1 i2
TRINITY DN34761 c0 g1 i3
TRINITY DN67488 c1 g1 i2
TRINITY DN67488 c1 g1 i1
consensus
X non conserved
X ≥50% conserved

Figure 5: (b)

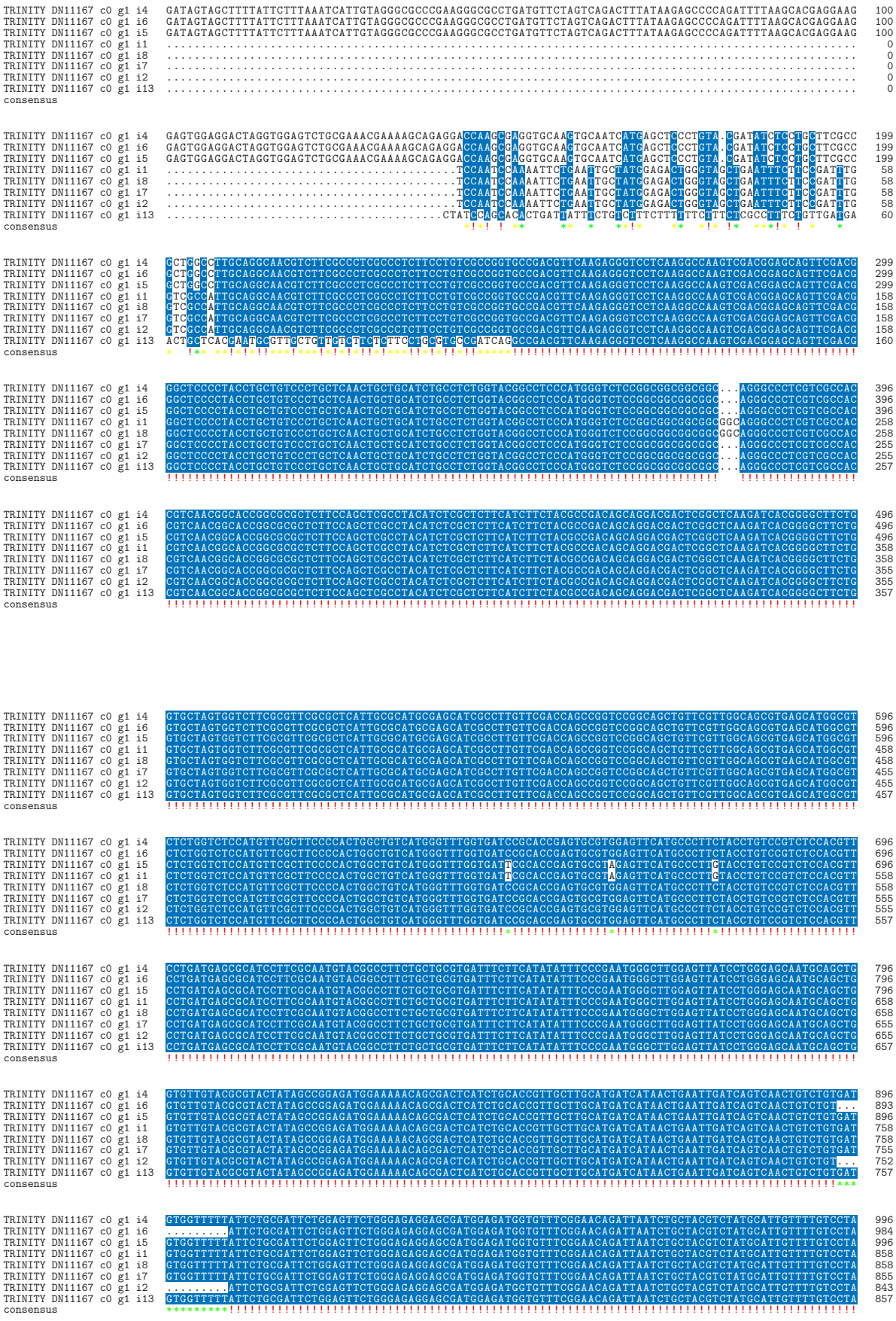


Figure 5: (c)

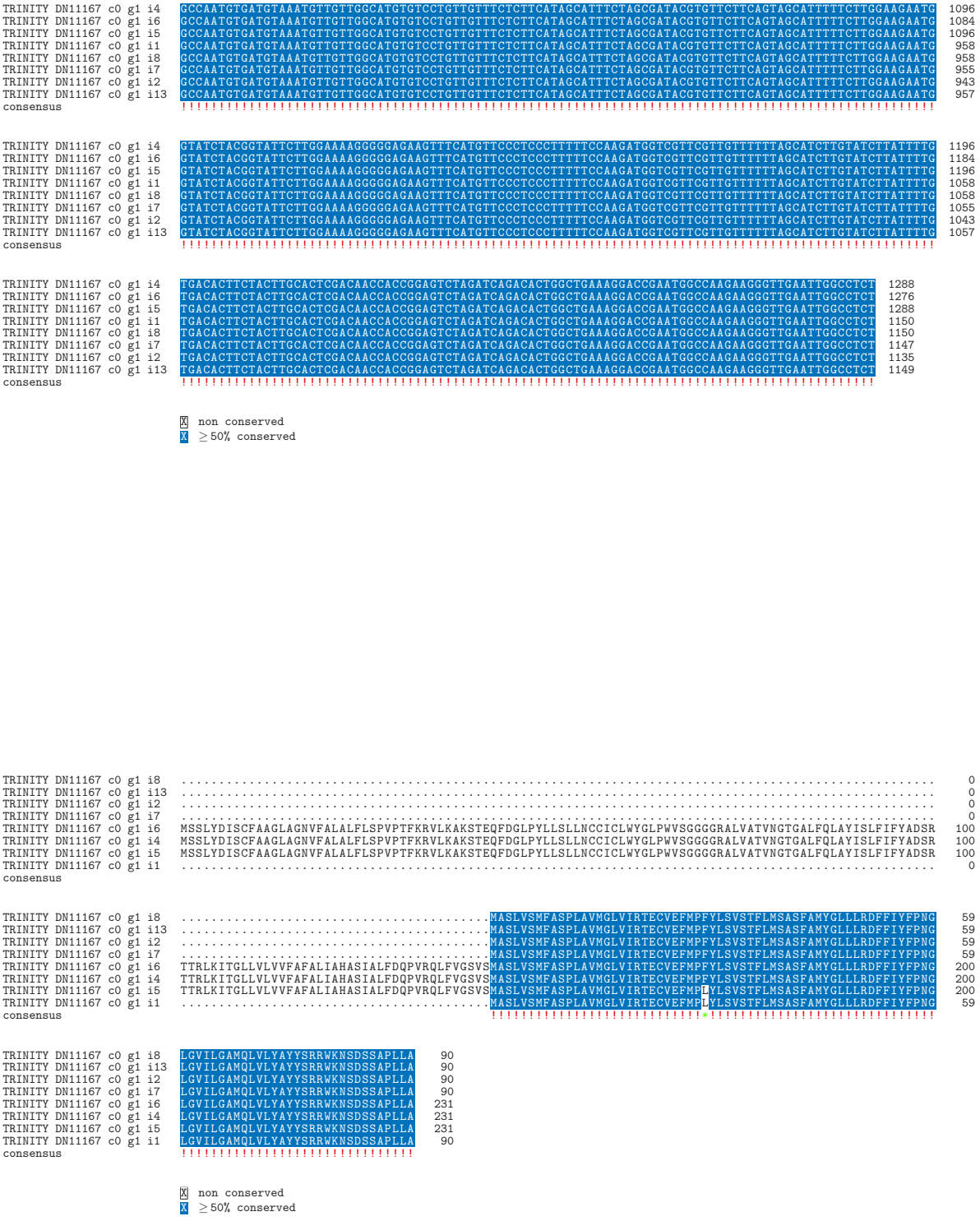

TRINITY DN11167 c0 g1 i4 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 1096
TRINITY DN11167 c0 g1 i6 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 1084
TRINITY DN11167 c0 g1 i5 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 1096
TRINITY DN11167 c0 g1 i1 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 958
TRINITY DN11167 c0 g1 i8 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 958
TRINITY DN11167 c0 g1 i7 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 955
TRINITY DN11167 c0 g1 i2 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 943
TRINITY DN11167 c0 g1 i13 GCCAATGTGATGTAAATGTTGTTGGCATGTGTCCTGTTGTTTCTCTTCATAGCATTTCTAGCGATACGTGTTCTTCAGTAGCATTTTTCTTGGAAGAATG 957
consensus
TRINITY DN11167 c0 g1 i4 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1196
TRINITY DN11167 c0 g1 i6 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1184
TRINITY DN11167 c0 g1 i5 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1196
TRINITY DN11167 c0 g1 i1 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1058
TRINITY DN11167 c0 g1 i8 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1058
TRINITY DN11167 c0 g1 i7 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1055
TRINITY DN11167 c0 g1 i2 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1043
TRINITY DN11167 c0 g1 i13 GTATCTACGGTATTCTTGGAAAAGGGGGAGAAGTTTCATGTTCCCTCCCTTTTTCCAAGATGGTCGTTCGTTGTTTTTTAGCATCTTGTATCTTATTTTG 1057
consensus
TRINITY DN11167 c0 g1 i4 TGACACTTCTACTTGCACTCGACAACCACCGCGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1288
TRINITY DN11167 c0 g1 i6 TGACACTTCTACTTGCACTCGACAACCACCGGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1276
TRINITY DN11167 c0 g1 i5 TGACACTTCTACTTGCACTCGACAACCACCGGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1288
TRINITY DN11167 c0 g1 i1 TGACACTTCTACTTGCACTCGACAACCACCGGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1150
TRINITY DN11167 c0 g1 i8 TGACACTTCTACTTGCACTCGACAACCACCGGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1150
TRINITY DN11167 c0 g1 i7 TGACACTTCTACTTGCACTCGACAACCACCGGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1147
TRINITY DN11167 c0 g1 i2 TGACACTTCTACTTGCACTCGACAACCACCGGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1135
TRINITY DN11167 c0 g1 i13 TGACACTTCTACTTGCACTCGACAACCACCGGAGTCTAGATCAGACACTGGCTGAAAGGACCGAATGGCCAAGAAGGGTTGAATTGGCCTCT 1149
consensus
X non conserved
X ≥ 50% conserved
TRINITY DN11167 c0 g1 i8 0
TRINITY DN11167 c0 g1 i13 0
TRINITY DN11167 c0 g1 i2 0
TRINITY DN11167 c0 g1 i7 0
TRINITY DN11167 c0 g1 i6 MSSLYDISCFAAGLAGNVFALALFLSPVPTFKRVLKAKSTEQFDGLPYLLSLLNCCICLWYGLPWVSGGGGRALVATVNGTGALFQLAYISLFIFYADSR 100
TRINITY DN11167 c0 g1 i4 MSSLYDISCFAAGLAGNVFALALFLSPVPTFKRVLKAKSTEQFDGLPYLLSLLNCCICLWYGLPWVSGGGGRALVATVNGTGALFQLAYISLFIFYADSR 100
TRINITY DN11167 c0 g1 i5 MSSLYDISCFAAGLAGNVFALALFLSPVPTFKRVLKAKSTEQFDGLPYLLSLLNCCICLWYGLPWVSGGGGRALVATVNGTGALFQLAYISLFIFYADSR 100
TRINITY DN11167 c0 g1 i1 0
consensus
TRINITY DN11167 c0 g1 i8 .....MASLVSMFASPLAVMGLVIRTECVEFMPFYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 59
TRINITY DN11167 c0 g1 i13 .....MASLVSMFASPLAVMGLVIRTECVEFMPFYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 59
TRINITY DN11167 c0 g1 i2 .....MASLVSMFASPLAVMGLVIRTECVEFMPFYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 59
TRINITY DN11167 c0 g1 i7 .....MASLVSMFASPLAVMGLVIRTECVEFMPFYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 59
TRINITY DN11167 c0 g1 i6 TTRLKITGLLVLVVFAFALIAHASIALFDQPVRQLFVGSVSMASLVSMFASPLAVMGLVIRTECVEFMPFYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 200
TRINITY DN11167 c0 g1 i4 TTRLKITGLLVLVVFAFALIAHASIALFDQPVRQLFVGSVSMASLVSMFASPLAVMGLVIRTECVEFMPFYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 200
TRINITY DN11167 c0 g1 i5 TTRLKITGLLVLVVFAFALIAHASIALFDQPVRQLFVGSVSMASLVSMFASPLAVMGLVIRTECVEFMPLYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 200
TRINITY DN11167 c0 g1 i1 .....MASLVSMFASPLAVMGLVIRTECVEFMPLYLSVSTFLMSASFAMYGLLLRDFFIYFPNG 59
consensus
TRINITY DN11167 c0 g1 i8 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 90
TRINITY DN11167 c0 g1 i13 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 90
TRINITY DN11167 c0 g1 i2 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 90
TRINITY DN11167 c0 g1 i7 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 90
TRINITY DN11167 c0 g1 i6 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 231
TRINITY DN11167 c0 g1 i4 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 231
TRINITY DN11167 c0 g1 i5 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 231
TRINITY DN11167 c0 g1 i1 LGVILGAMQLVLYAYYSRRWKNSDSSAPLLA 90
consensus
X non conserved
X ≥ 50% conserved

Figure 5: (c)

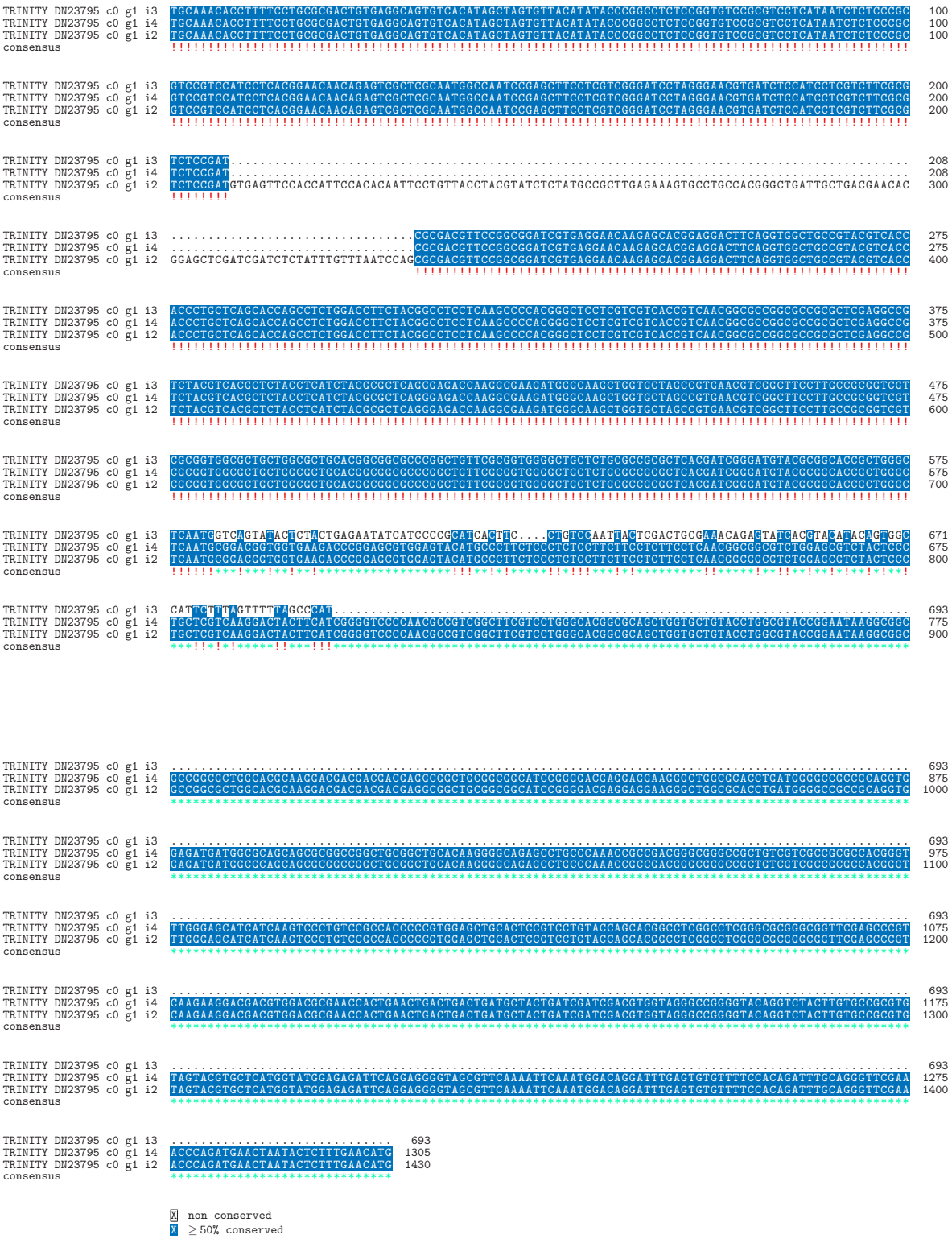

TRINITY DN23795 c0 g1 i3
TRINITY DN23795 c0 g1 i4
TRINITY DN23795 c0 g1 i2
consensus
X non conserved
X ≥50% conserved

Figure 5: (d)

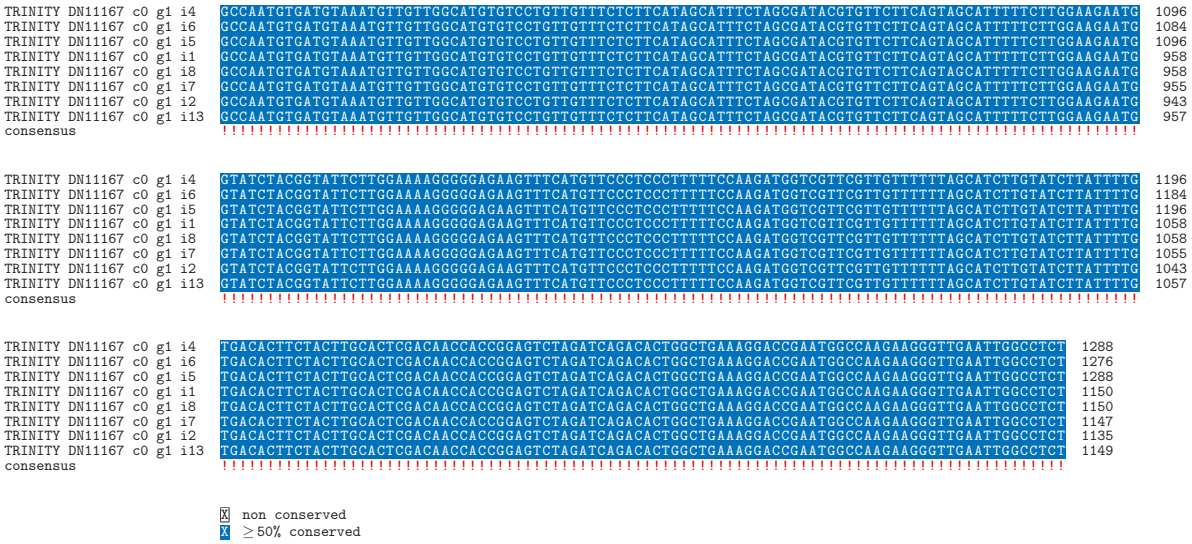


Figure 5: Multiple alignments with the cDNA and predicted amino acid sequences from the genes Arginase 1, mitochondrial (a); Cyclin P2-1 (b); SWEET2b (c); and SWEET16(d). For the gene TRINITY_DN67488_c1_g1, it was used the reverse complement from the sequence found in the transcriptome.

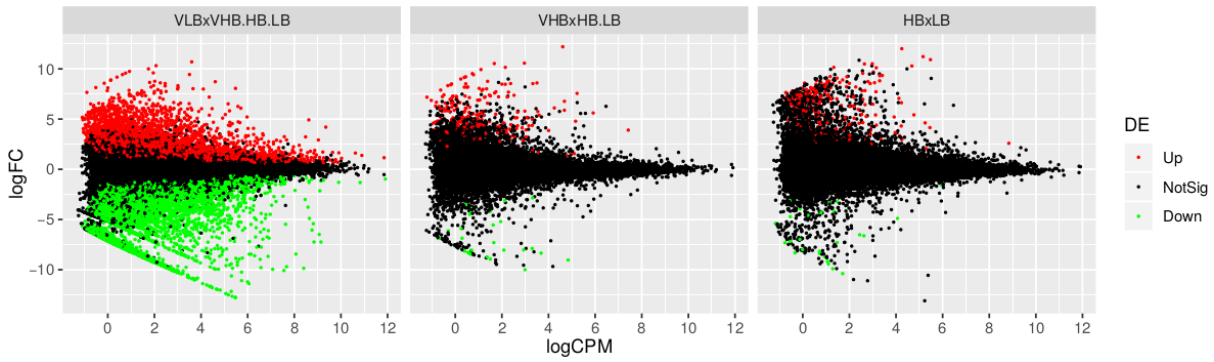


Figure 6: Mean-difference plot with the differentially expressed genes for all contrasts. The axes represent the fold changes (log_2FC) and the average expression levels in counts per million (log_2CPM). Colors represent the result of differential expression tests ($p < 0.05$, after FDR correction for multiple tests).

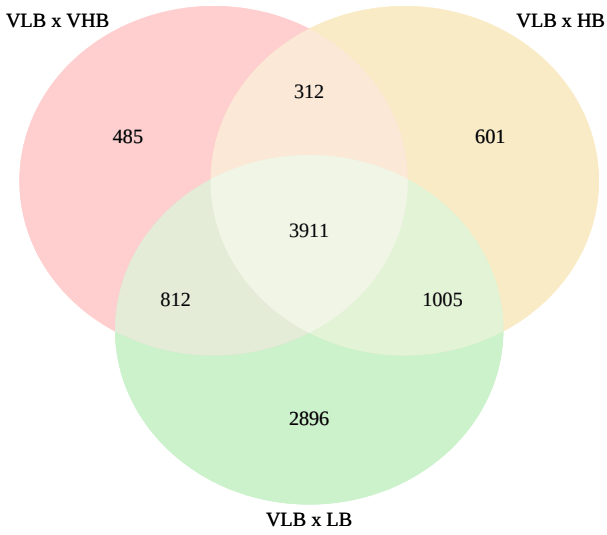


Figure 7: Sets of shared differentially expressed genes in comparisons of high-sugar groups against genotypes with Very Low ºBrix.

Sample	Mapping rate
SBC	
SP80-3280 R1	75.12%
SP80-3280 R2	75.46%
SP80-3280 R3	75.03%
R570 R1	84.22%
R570 R2	75.20%
R570 R3	79.19%
F36-819 R1	76.32%
F36-819 R2	76.61%
F36-819 R3	78.64%
IN84-58 R1	75.04%
IN84-58 R2	79.26%
IN84-58 R3	81.70%
Mean	76.47%
SBDG	
SP80-3280	75.12%
White Mauritius	75.86%
RB835486	77.62%
R570	84.27%
RB92579	76.33%
White Transparent	76.15%
F36-819	76.62%
Criolla Rayada	76.30%
IJ76-317	76.71%
IN84-58	81.70%
Krakatau	76.92%
SES205A	76.54%
Mean	76.58%

Table 1: Mapping rates for transcripts by sample and strategy using the *de novo* assembled transcriptome. SBC - Strategy based on clones; SBDG - Strategy based on diverse genotypes.

Alias	Accession ID	Gene name
NI	AHK06420.1	alkaline/neutral invertase protein [Saccharum hybrid cultivar GT28]
NI	AFV94466.1	alkaline/neutral invertase protein [Saccharum hybrid cultivar GT28]
CWI	AFV09275.1	cell wall invertase [Saccharum hybrid cultivar GT28]
CWI	AFV09274.1	cell wall invertase [Saccharum hybrid cultivar GT28]
SAI	AFV94475.1	soluble acid invertase 12 [Saccharum hybrid cultivar GT28]
SAI	AFV94474.1	soluble acid invertase 11 [Saccharum hybrid cultivar GT28]
SAI	AFV94473.1	soluble acid invertase 10 [Saccharum hybrid cultivar GT28]
SAI	AFV94472.1	soluble acid invertase 9 [Saccharum hybrid cultivar GT28]
SAI	AFV94471.1	soluble acid invertase 8 [Saccharum hybrid cultivar GT28]
SAI	AFV94470.1	soluble acid invertase 7 [Saccharum hybrid cultivar GT28]
SAI	AFV94469.1	soluble acid invertase 6 [Saccharum hybrid cultivar GT28]
SAI	AFV94467.1	soluble acid invertase 5 [Saccharum hybrid cultivar GT28]
SAI	AFV94468.1	soluble acid invertase 4 [Saccharum hybrid cultivar GT28]
SAI	AFV94414.1	soluble acid invertase 3 [Saccharum hybrid cultivar GT28]
SAI	AFV94413.1	soluble acid invertase 2 [Saccharum hybrid cultivar GT28]
SAI	AFV94412.1	soluble acid invertase 1 [Saccharum hybrid cultivar GT28]
SAI	AFN66440.1	soluble acid invertase [Saccharum hybrid cultivar]
SAI	AGT16261.1	beta-fructofuranosidase [Saccharum hybrid cultivar R570]
VGT1	NP_186959.2	vacuolar glucose transporter 1 [Arabidopsis thaliana]
TST1	ADG21982.1	tonoplast monosaccharide transporter 1 [Oryza sativa Japonica Group]
TST2	ADG21983.1	tonoplast monosaccharide transporter 2 [Oryza sativa Japonica Group]

Table 2: Manually selected sugarcane genes from the databases GenBank and RefSeq. The listed genes were compiled into one blast database and used to identify invertases, tonoplast monosaccharide transporter, and vacuolar glucose transporter families in the transcriptome. These genes were not available in Swiss-Prot in release 2019-1.

Genotype	Number of reads	QC cleaning rate	Mapping rate
Criolla Rayada	43,224,854	92.76%	76.20%
F36-819	45,598,224	92.71%	76.62%
IJ76-317	40,096,612	92.74%	76.71%
IN84-58	43,422,964	92.73%	81.70%
Krakatau	40,349,586	92.62%	76.92%
R570	45,443,304	92.99%	84.27%
RB835486	38,829,088	90.78%	77.62%
RB92579	45,191,582	92.83%	76.33%
SES205A	43,589,964	92.64%	76.54%
SP80-3280	50,481,854	91.79%	75.12%
White Mauritius	41,749,608	92.56%	75.86%
White Transparent	37,865,194	92.67%	76.15%

Table 3: RNA-Seq datasets statistics. Length of the original datasets by the number of individual reads (counting both R1 and R2 paired reads), quality control outcome after filtering FASTQ raw files, and transcripts mapping rates with salmon by genotype.