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

Assessing differential expression profiles and modeling allele-specific expression in leaves of *Saccharum* accessions contrasting in biomass production

Fernando Henrique Correr

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2021

Fernando Henrique Correr
Bachelor in Biotechnology

Assessing differential expression profiles and modeling allele-specific expression in leaves of *Saccharum* accessions contrasting in biomass production

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

Thesis presented to obtain the degree of Doctor in Science.
Area: Genetics and Plant Breeding

Piracicaba
2021

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

Correr, Fernando Henrique

Assessing differential expression profiles and modeling allele-specific expression in leaves of *Saccharum* accessions contrasting in biomass production / Fernando Henrique Correr. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2021 .

161 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Cana-de-açúcar 2. Transcriptomas 3. Similaridade fenotípica 4.Desbalanço alélico I. Título

DEDICATORY

To my parents, Claudia and Valdecir,
and my girlfriend Maria Clara

ACKNOWLEDGEMENTS

I thank the University of São Paulo for my professional qualification. To the "Luiz de Queiroz" College of Agriculture (ESALQ), especially to the Genetics Department for supporting me during the graduate courses.

To the Brazilian National Council for Scientific and Technological Development (CNPq) and to the Coordination for the Improvement of Higher Education Personnel (CAPES) for providing the financial support to conduct my research.

Special acknowledgments go to my advisor Prof. Gabriel Rodrigues Alves Margarido for his guidance and support during these years. Thank you for teaching me not only about genetics, statistics and bioinformatics, but for showing me how to be an ethical professional. I deeply appreciate the effort you spent for keeping me engaged. Thank you for being an example of professor and scientist! I also extend my gratitude to his family, Maidy and Beatriz, for all attention you gave me in Brisbane.

My special thanks go to my friends from the Laboratory of Bioinformatics Applied to Bioenergy for the scientific discussion, friendship, talks and funny moments: Amanda Avelar de Oliveira, Amanda Ghelfi Dumit, Ana Letycia Basso Garcia, André Luís Patrício, Diane de Fátima Sgariboldi, Guilherme Bovi Ambrosano, Guilherme Kenichi Hosaka, Leonardo Sartori Menegatto, Lorena Guimarães Batista and Victor Hugo de Mello Pessoa. It is a pleasure to be your friend!

Special acknowledgement to Prof. Monalisa Sampaio Carneiro, who contributed to all the biological material, laboratory infrastructure and insights. Thank you for showing the beautiful of sugarcane genetics!

Additionally, I express my gratitude to the University of Queensland and the Queensland Alliance for Agriculture and Food Innovation (QAAFI), for all the support provided during my internship. Special acknowledgment goes to Prof. Robert Henry for the opportunity to work in his group. Thank you for your kindness in helping me and for the valuable discussions we had. I also express my gratitude to Dr. Agnelo Furtado, for his help and friendship. I extend my gratitude to all my friends in Robert's group: Adhini, Angela, Ardy, Bader, Joseph, Kalpani, Katrina, Onkar, Othman, Patrick, Prameela, Priyanka, Sharmin, Vallari and Virginie. I also thank Prof. David Braun for sharing his knowledge on sucrose transporters and for his friendship during his stay in Brisbane. My gratitude to all QAAFI staff, especially to Annie Morley, Annie Cox, and Maria Caldeira. I am also thankful to Tarnya, Dave and Restu for all the good moments at the "Trenton Trio". My gratitude to Julia, Elliot, Carolina, Diana, Fernanda, Wei-An Tsai and William (Bill), who were incredible friends during this period in Australia. I also wish to acknowledge The University of Queensland's Research Computing Centre (RCC) for its support in this research.

My sincere thanks go to all professors of the Department of Genetics, especially to Claudia Barros Monteiro Vitorello, Gerhard Bandel, Giancarlo Conde Xavier Oliveira, Margarida Lopes Rodrigues de Aguiar-Perecin and Maria Carolina Quecine Verdi. Thank you for all the lectures and for the important talks regarding not only science, but also for showing the history of genetics in Brazil. My special gratitude to Prof. Augusto Franco Garcia, who contributed enormously to this work and with my academic development. Sincere thanks to the Department's staff: Antonio de Padua Gorga (Berdan), Cândida Vanderléia de Oliveira, Carlos Roberto Macedonio, Fernando Leopoldino, Maídia Maria Thomaziello, Silvia Zanatta and Valdir Próspero.

Additionally, I would like to thank all members of GVENCK and GENt for providing an incredible and creative environment within the department. To my friends Maria Bonatelli, Júlia Morosini and Fernando Espolador. To Cristiane Taniguti, for your long-term friendship.

I am also grateful to all the collaborators during my PhD: Prof. Anete Pereira, Carla Silva, Danilo Sforça and Mariana Cia. To Eliana Garcia and Rodrigo Pessanha for all help with the thesis

submission. I would like also to thank Prof. Clícia Grativol Gaspar de Matos, Prof. Marcelo Carazzolle, Dr. Augusto Lima Diniz, Prof. Paulo José Pereira Lima Teixeira and Prof. Tiago Antônio de Oliveira Mendes for accepting the invitation to participate in my defense. And to all the friends, teachers and professors who have contributed to my academic and personal growth.

Thanks to my family, especially to my parents, Claudia and Valdecir and my grandparents, Maria Cecília and Orlando. To my girlfriend Maria Clara. Thank you for your PATIENCE, encouragement and unconditional love! I love you!

Finally, I thank God for the gift of life, for never letting me weaken or giving up during this journey.

If a thing be ordained to another as to its end, its last end cannot consist in the preservation of its being. Hence a captain does not intend as a last end, the preservation of the ship entrusted to him, since a ship is ordained to something else as its end, viz. to navigation.

St. Thomas Aquinas

SUMMARY

RESUMO

Avaliação de perfis de expressão diferencial e modelagem da expressão alelo-específica em folhas de acessos de *Saccharum* contrastantes na produção de biomassa

A cana-de-açúcar é uma das mais importantes culturas agrícolas mundiais devido a seus principais produtos - açúcar e álcool -, o reuso de seus subprodutos e a capacidade de inovação de sua agroindústria. Apresenta um potencial para uma produção mais rentável e sustentável, que pode ser obtida pelo desenvolvimento de cultivares de alta produtividade. Por esse motivo, características além do teor de sacarose nos colmos devem ser exploradas. Recentemente, a chamada cana-energia fez com que os programas de melhoramento contemplassem características relacionadas à biomassa, como o conteúdo de fibra e a capacidade de perfilhamento. A variação genética associada a essas características pode ser maximizada pela inclusão de outros acessos de *Saccharum*, os quais ainda não foram explorados pelos melhoristas. Além disso, os estudos sobre os perfis de expressão gênica em folhas de diferentes grupos de genótipos ainda são limitados na literatura. Portanto, objetivou-se a avaliação dos transcriptomas das folhas de dois grupos de genótipos - alta e baixa biomassa - a fim de identificar genes ou alelos potencialmente envolvidos com o controle do conteúdo de biomassa. Para esse objetivo, genótipos foram selecionados com base em sua similaridade fenotípica, independentemente de suas classificações como cultivados ou selvagens. O estudo foi dividido em dois capítulos. No primeiro, os objetivos foram a identificação de genes diferencialmente expressos entre os grupos de biomassa e a investigação dos perfis de expressão de genes coexpressos. Os resultados mostraram que o estudo da expressão gênica permitiu não só caracterizar a variabilidade entre os grupos, como também a variabilidade dentro de cada grupo. Apesar da similaridade fenotípica, o grupo de alta biomassa mostrou uma alta variabilidade entre seus acessos, o que resultou em número expressivo de genes diferencialmente expressos, muito maior do que a comparação intergrupo. Genes que codificam a sacarose sintase e proteínas relacionadas à síntese de sacarose foram ligeiramente mais expressas no grupo de baixa biomassa, enquanto que aqueles envolvidos com a síntese de compostos da parede celular foram significativamente menos expressos. Curiosamente, a análise de coexpressão revelou que a expressão de genes relacionados com a fotossíntese foi maior em todos os genótipos híbridos e em *Saccharum officinarum*. Mostrou-se, também, que a quantificação da expressão em diferentes níveis tem influência nas considerações biológicas desse tipo de estudo. No segundo capítulo, testou-se a expressão alelo-específica (*allele-specific expression*, ou ASE) em um subconjunto de amostras de *Saccharum*. Esses acessos - três híbridos, uma *S. officinarum* e duas *S. spontaneum* - foram genotipados através da técnica de genotipagem-por-sequenciamento, o que permitiu a estimação da ploidia e das dosagens alélicas de sítios variantes. Modelou-se, para cada polimorfismo, a probabilidade da expressão do alelo de referência por um modelo Beta-Binomial hierárquico, no qual as dosagens alélicas genômicas serviram de informação *a priori*. Os resultados revelaram que a ASE afeta parte dos *loci* avaliados em *Saccharum*. Entretanto, nenhum termo funcional foi enriquecido entre os genes que exibiram ASE. Este estudo foi a primeira visão global da ocorrência de expressão alelo-específica em múltiplos genótipos de cana-de-açúcar. Ademais, o modelo hierárquico pode ser usado para avaliar ASE em outros organismos de ploidia mista.

Palavras-chave: Cana-de-açúcar, Transcriptomas, Similaridade fenotípica, Desbalanço alélico

ABSTRACT

Assessing differential expression profiles and modeling allele-specific expression in leaves of *Saccharum* accessions contrasting in biomass production

Sugarcane is one of the most important crops worldwide due to its main products - sugar and ethanol -, the reuse of byproducts and the innovation capability of the agroindustry. It offers the potential for a more profitable and sustainable production, which can be accomplished by developing high-yielding cultivars. For that reason, traits other than the sucrose content in culms should also be explored. The so-called energy cane has recently moved the attention of breeding programs towards biomass-related traits such as fiber content and tillering capacity. The genetic variation associated with these traits can be enhanced with other *Saccharum* accessions that have not yet been explored by breeders. In addition, studies regarding gene expression profiles in diverse groups of genotypes are still limited in the literature. Therefore, we aimed to assess the transcriptomes from leaves of two groups of genotypes - high and low biomass - to identify genes or alleles potentially involved with biomass content. To achieve such goal, genotypes were selected based on their similar phenotypes, regardless of their classification as cultivated or wild. We divided this study into two chapters. In the first chapter, our aim was to identify differentially expressed genes between the biomass groups and to investigate the expression profiles of coexpressed genes. Our results showed that gene expression allowed the study not only of the variability between the contrasting groups, but also the variation within each group. Despite the phenotypic similarity, the high biomass group showed a large variability among its accessions, resulting in many differentially expressed genes (DEGs), many more than in the intergroup comparison. Genes coding for sucrose synthase and proteins related to sucrose synthesis were slightly more expressed in the low biomass group, whereas genes involved with the synthesis of cell wall compounds were significantly less expressed. Interestingly, the coexpression analysis revealed that the expression of genes related to photosynthesis was higher in all hybrids and *Saccharum officinarum* genotypes. We also showed that different quantification levels have certain influence on the biological insights provided by this kind of study. In the second chapter, we tested for allele-specific expression (ASE) in a subset of the *Saccharum* samples. These accessions - three hybrids, a *S. officinarum* and two *S. spontaneum* - were genotyped via genotyping-by-sequencing, followed by the estimation of ploidy and allelic dosages. We then modeled, for each polymorphism, the probability of expressing the reference allele using a hierarchical Beta-Binomial model, where allelic dosages served as prior information. Results revealed that ASE affects part of the loci assessed in *Saccharum*. However, no functional term was enriched among genes showing ASE. This study provides the first global view of allele-specific expression in multiple genotypes of sugarcane. Furthermore, the hierarchical model can be used to evaluate ASE in other mixed-ploidy organisms.

Keywords: *Saccharum*, Transcriptomes, Phenotypic similarity, Allelic imbalance

1 INTRODUCTION

Sugarcane is an important crop in Brazil since the Portuguese colonization to produce sugar and, for approximately 50 years, to produce ethanol. Recently, data from the Brazilian Sugarcane Industry Association (UNICA) shows that sugarcane is planted in more than 5.5 million hectares in the State of São Paulo (http://www.unicadata.com.br/ - year 2018). According to the National Supply Company (CONAB), the Brazilian sugarcane production in the 2020/21 harvest is expected to increase in comparison with the previous year, reaching roughly 665.1 million tonnes [5]. While the total ethanol production will be reduced by 7.9%, sugar production is estimated to increase by 40.4% (41.8 million tonnes). Progress in the sugarcane industry was partially achieved through breeding high-performance cultivars. Briefly, the sugarcane breeding process relies on crossing parental genotypes, selecting superior genotypes for traits with high variability, then evaluating clones in proper experimental designs for lower heritability traits and, finally, assessing the genotype-environment interactions in competition trials [20]. Breeders have focused on increasing plant productivity to supply the industrial needs of raw material. At the same time, a more effective production in the same cultivable area is desired for a more sustainable agriculture. Scortecci and colleagues [49] stress the importance of leveraging the genetic potential of cultivars to achieve high yields and reduce the natural resources consumed by the plant. Moreover, we should explore not only the variability of sugarcane cultivars, but also from other *Saccharum* species.

Sugarcane is taxonomically classified as belonging to the genus *Saccharum*, subtribe *Saccharinae*, of the Poaceae family. Six species have been studied for understanding the evolution in the genus. Among them, four can be classified as cultivable: *Saccharum officinarum* L., *S. barberi* Jeswiet, *S. sinense* Roxb. and *S. edule* Hassk [40, 55, 39]. The same authors classify the two remaining species as wild: *S. spontaneum* L. and *S. robustum* Brandes & Jeswiet ex Grassl. Due to their proximity and the possibility of intergeneric crossings, *Erianthus*, *Miscanthus*, *Narenga*, *Saccharum* and *Sclerostachya* form the *Saccharum* complex [40, 55, 39]. Historically, the main objective of sugarcane breeding was sucrose accumulation in culms using mostly *S. officinarum* accessions. Later, crossings with *S. spontaneum* were performed to introgress traits related to stress tolerance [53]. The recent development of a group of high-productivity cultivars - energy canes - directed the breeders' attention to biomass [22, 13, 21]. As stated in studies dating from the 80s [8, 31], energy canes should achieve high yields of both sugar and biomass. The development of such new genotypes demands genetic resources in terms of biomass-related traits, such as fiber content in culms and tillering capacity. Breeding programs can thus benefit from enhanced knowledge about the molecular basis of desired traits, obtained via molecular markers and genomic sequences [7].

The association between genotypic and phenotypic data is not trivial in sugarcane. All *Saccharum* are polyploids showing a large number of chromosomes, which is variable in different accessions of the same species [55, 51]. As a consequence of the interspecific hybridization and successive backcrosses with *S. officinarum*, the modern cultivars have a very complex genome. Most of the basic chromosome architectures ($x = 10$) are represented by approximately eight *S. officinarum* homologs, *S. spontaneum* chromosomes and a small proportion of recombinants between the two species [51]. During sugarcane breeding, other *Saccharum* species - *S. barberi*, *S. sinense* and *S. robustum* - had a minimum contribution [38, 52]. Multiple strategies were used to unravel its genome sequence [50, 58, 33]. Recently, Garsmeur and colleagues [16] published a mosaic genome assembly of a commercial hybrid; Zhang and colleagues [26] published the sequence of a tetraploid *S. spontaneum* genome; and Souza and colleagues [4] published the gene space assembly of a Brazilian hybrid. However, analyzing the sugarcane genome is still a difficult task when different *Saccharum* accessions are being studied. Approaches using transcriptomes are useful to investigate likely cellular functions of putative genes, aiming to obtain molecular markers from functional genomic regions. Pioneering initiatives paved the way for functional genomics in sugar-

cane. First, Carson and colleagues [43] assessed gene expression in sugarcane leaf rolls using expressed sequence tags (ESTs). Two years later, after assessing the transcriptome of sugarcane leaves, they found genes functionally associated with the control and maintenance of cellular metabolism, transport and response to stresses [41]. Afterwards, researchers in the SUCEST project obtained more than 200 thousand ESTs from different samples [57]. Differentially expressed genes related to cell wall, cellulose and lignin biosynthesis were identified among different stages of culm development via transcriptome profiling [9].

These functional genomics and physiological studies in sugarcane provided evidence of important genes related to sucrose accumulation and synthesis of structural compounds. Along with advances described in the literature for other plants, efforts have also been made to connect genes in pathways to understand carbon partitioning in sugarcane. Wang and colleagues [48] showed the main steps for this process, from sucrose synthesis to its distribution to the sink cells. They showed that after photosynthesis on sugarcane leaves, sucrose is translocated in the phloem and reaches the stem parenchyma cells through both symplast and apoplast. These authors also reported key enzymes for sucrose accumulation: i) sucrose phosphate synthase (SPS) synthesizing sucrose-P from fructose-6-P and UDP-glucose; ii) sucrose phosphate phosphatase (SPP) producing sucrose from sucrose-P; iii) sucrose synthase (SuSy) being responsible for a reversible reaction converting fructose and UDP-glucose to sucrose; iv) cell wall invertase hydrolyzing sucrose into hexoses in the apoplast. There are also other classes of invertases and transporters that participate in transferring hexoses and sucrose into the cellular compartments. In addition to the transport via symplast, hexoses are transported by carriers and resynthesized into sucrose in the cytoplasm. Curiously, *Saccharum* species accumulate similar levels of symplastic and apoplastic solutes [2]. However, in general, high fiber species - *S. robustum* and *S. spontaneum* - show higher percentages of insoluble solids than sucrose-rich *Saccharum*, which in turn present a higher content of soluble solids [46, 2]. It is worth mentioning that *S. spontaneum* has a higher content of starch in mature culms to probably meet metabolic demands, serving as a resource for tillering and when the plant is submitted to stress [46].

Attention has been devoted to understand the synthesis of cell wall compounds, as the fibrous part can now be used as raw material by the sugarcane industry. The cell wall can be used in diverse manners, such as a prime source of energy, as feedstock and to develop cellulose-based materials [1]. Regarding the structure, primary and secondary walls of grasses are formed mostly by cellulose, followed by hemicellulose - arabinan- and xylan-derived compounds -, phenolic compounds, pectins, proteins and silica [1, 47]. The composition varies in different developmental stages of the culm. While the hemicellulose content is higher in younger internodes, cellulose is higher in mature internodes [46]. The synthesis of these elements requires the action of enzymes coordinated in different molecular pathways. More than a hundred candidate genes were found to be significantly associated with different fiber composition traits [47]. For cellulose, UDP-glucose from SuSy reaction is used by a complex set of cellulose synthase proteins to synthesize the glucan chain [45, 48]. This is corroborated by the significant association of both SuSy and UDP-glucosyl transferase with cellulose [47]. The biosynthesis of lignin is carried out by many enzymes of the phenylpropanoid pathway. In this pathway, Jardim-Messeder and colleagues [44] defined a core set of genes involved in lignin biosynthesis from the following families: phenylalanine/tyrosine ammonia-lyase, 4-(hydroxy) cinnamoyl CoA ligase, cinnamate 4-hydroxylase, hydroxycinnamoyl CoA shikimate:quinate hydroxycinnamoyltransferase, ρ-coumaroyl shikimate:quinate 3´-hydroxylase, caffeoyl CoA O-methyltransferase, caffeic acid/5-hydroxyferulic acid O-methyltransferase, ferulic acid/coniferaldehyde/coniferyl alcohol 5-hydroxylase, (hydroxy)cinnamoyl CoA reductase and (hydroxy)cinnamyl alcohol dehydrogenase. Authors reported that the expression of genes of the biosynthesis of monolignols have both genotype- and tissue-specificity [46]. High-fiber *Saccharum* species - *S. robustum* and *S. spontaneum* - show more diverse lignin oligomers [46]. The set of 15 phenylpropanoid core genes showed increased expression levels according to culm development [44].

These authors also analyzed the haplotypes of these genes, revealing an uneven distribution in the *S. spontaneum* genome. However, they could identify similar distribution of *cis*-elements in the upstream region of different haplotypes of a gene. Transcription factors can bind to such regions an regulate the expression of members of the phenylpropanoid pathway. In fact, biosynthesis of secondary cell wall can be regulated by myeloblastosis (MYB) and NAC transcription factors, as they are correlated to genes acting on the synthesis of lignin, tricin and hemicellulose [45].

New sequencing technologies, the possibility of assembling transcriptomes *de novo* and the development of statistical methods led to a revolution in the analysis of transcriptomes. The so-called RNA-Sequencing [30] has allowed an increase in the number of characterized sugarcane transcripts, as well as the comparison between contrasting conditions. In 2014, Cardoso-Silva and colleagues [56] assembled the transcriptomes of six cultivars, discovering 5,272 new putative genes not found in the SUCEST database. These authors found genes related to sucrose accumulation and responses to diseases. In the same year, the transcriptomes of the cultivar SP80-3280, accessions of *S. officinarum* and *S. spontaneum* were investigated [10]. These authors showed a high number of *S. spontaneum*-specific transcripts related to stress, signal transduction and transcription factors in sugarcane leaves. They also found that 78.28% of the transcripts were expressed in all genotypes and suggested that major phenotypic differences may be due to reasons other than expression variation at the gene level, such as isoforms, allelic variation and polymorphisms. Later, more than 500 transcripts associated to carbohydrate metabolism and transport were identified in the transcriptome of a high-sucrose cultivar [17].

The advance of sequencing methods has allowed the identification of isoforms, their occurrence in different tissues, development stages or growth conditions. In sugarcane, libraries from different organs were combined: i) first, second and third visible dewlap leaves; ii) immature and mature roots; and iii) the third internode from the top and the third internode from the base [59]. They generated a *de novo* transcriptome using Illumina sequencing on samples of (iii) and the isoform sequencing (Iso-Seq) from Pacific Biosciences on (i), (ii) and (iii) to identify isoforms. The *de novo* assembled transcriptome had a higher percentage of read alignment, more predicted proteins with homology to Viridiplantae and allowed the discovery of a larger number of KEGG pathways. Iso-Seq, on the other hand, recovered more complete transcripts, which aligned better to the *Sorghum bicolor* genome [59]. These results indicate the potential of Iso-Seq for comparative analyses.

Gene expression can be quantified after mapping reads to the transcripts from which they were originated. RNA-Sequencing has the potential to capture the dynamism of expressed genes from a population of cells, in a given experimental condition, creating the base for differential expression studies [28, 19, 30]. Gene expression data were also used to compare genotypes with different biomass content, aiming to identify transcripts related to carbon partitioning and to precursors of fiber components. Vicentini and collaborators [54] compared two cultivars showing 4% difference in lignin content. They identified more than 2,000 differentially expressed genes (DEGs), with four main distinct expression profiles and more than 100 groups of genes with similar expression. Among the DEGs, authors reported enrichment of the phenylpropanoid pathway, glutathione-S-transferases, trehalose metabolism, cell-wall proteins, response to biotic stresses and plant hormones. Instead of using clonal replicates of single genotypes to represent a given phenotypic group, Kasirajan and colleagues [32] compared two groups of genotypes with contrasting lignin content. They found DEGs more expressed in the high-fiber genotypes that were present in the phenylpropanoid pathway - lignin precursors - and associated with carbohydrate metabolism. However, by only using elite germplasm these articles exploit little existing variability for fiber content and, consequently, for biomass yield.

There are also other approaches to use the expression data provided by these high-throughput methods. One strategy is not to focus on expression at the gene level, but to look for differentially expressed transcripts and characterize splicing events. A second procedure is to assess the variation in

expression levels among the alleles of a gene. In that case, differences in the expression magnitude of two alleles can indicate allele-specific expression (ASE). This phenomenon can be explained by *cis*-regulation on promoter regions, frameshift mutations and epigenetic modifications that result on higher expression of one allele [24]. To evaluate ASE, polymorphisms have to be detected and allelic quantification should be obtained from RNA-Seq reads [14, 35, 23]. Then, for each polymorphism, a statistical test can be performed to detect allelic imbalance, by checking for deviations from equivalent expression between the alleles [12, 36]. ASE has been commonly assessed in large scale projects, mostly in human genetics [27, 37, 36, 35]. For example, a higher genic dosage caused by structural variations resulting from tumors was directly associated to increased allelic imbalance [36]. Recently, Lee and collaborators [37] found genes with allele-specific expression related to autism spectrum disorder risk. This approach has been used also in plants [14, 18, 42] and can be explored in other species.

As stated previously, ASE studies jointly use genotypic and expression data, which is feasible for sugarcane. Mancini and collaborators [7] discuss the main advances in sugarcane genetics and genomics. One of the most important is the use of SNPs to estimate the doses of the sugarcane alleles [11]. The high abundance of such markers is important for detecting a large number of polymorphisms, which are used to build genetic maps, discover QTLs and genomic regions associated with a given trait. It also opens the possibility for integration with expression data. A diverse set of *Saccharum* accessions was established in the Federal University of São Carlos (UFSCar), where researchers of the sugarcane breeding program laid out the Brazilian Panel of Sugarcane Genotypes [29, 3]. It is composed by 254 genotypes, representing wild species, cultivars with historic relevance and more recent cultivars. Some authors have already benefited from the genotyping of the panel [15, 29, 11, 3]. Using quantitative genotyping pipelines to obtain SNPs, the relative allelic proportions can be also estimated in this complex crop [11, 25, 34]. Then, the combination of such data with RNA-Sequencing provides enough information to evaluate ASE in sugarcane. However, a careful examination of the data is needed, as biases in the procedures - mapping and genotyping - can result in false ASE [24].

In this context, we point that it is feasible to understand, at the transcript level, differences between groups of accessions contrasting in their biomass content. In addition, investigating allelic imbalance can provide complementary results to the conventional analyses of gene profiles [36]. We explored gene expression data from leaves of twelve *Saccharum* accessions, phenotypically clustered in high- and low-biomass groups. First, we aimed to explore the variation between and within the groups in terms of differential gene expression. Next, we investigated the extent to which ASE occurred in both wild and cultivated accessions. We present and discuss our main findings regarding these objectives in two thesis chapters. The first chapter contains the investigation of differential gene expression, which was published in BMC Genomics [6]. We kept the integrity of all sections of this manuscript, including all the main and supplementary information. The second chapter focuses on the development of a model to test for ASE in complex polyploids such as sugarcane. It is also organized as a manuscript to be submitted.

References

[1] de Oliveira Buanafina MM, Cosgrove DJ. Cell Walls: Structure and Biogenesis. In: Sugarcane: Physiology, Biochemistry, and Functional Biology. Chichester, UK: John Wiley & Sons Ltd; 2013. p. 307–329. Available from: http://doi.wiley.com/10.1002/9781118771280.ch13.

[2] Welbaum GE. Water Relations and Cell Expansion of Storage Tissue. In: Sugarcane: Physiology, Biochemistry, and Functional Biology. Chichester, UK: John Wiley & Sons Ltd; 2013. p. 197–220. Available from: http://doi.wiley.com/10.1002/9781118771280.ch9.

[3] Medeiros C, Balsalobre TWA, Carneiro MS. Molecular diversity and genetic structure of Saccharum complex accessions. PLOS ONE. 2020 may;15(5):e0233211. Available from: http://dx.doi.org/10.1371/journal.pone.0233211https://dx.plos.org/10.1371/journal.pone.0233211.

[4] Souza GM, Van Sluys MA, Lembke CG, Lee H, Margarido GRA, Hotta CT, et al. Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. GigaScience. 2019 dec;8(12):1–18. Available from: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giz129/5647371.

[5] Companhia Nacional de Abastecimento (CONAB). Acompanhamento da Safra Brasileira de Cana-de-Açúcar – Terceiro Levantamento da safra 2020/21. Monitoramento agrícola – Cana-de-açúcar. 2020;7(3):1–62.

[6] Correr FH, Hosaka GK, Barreto FZ, Valadão IB, Balsalobre TWA, Furtado A, et al. Differential expression in leaves of Saccharum genotypes contrasting in biomass production provides evidence of genes involved in carbon partitioning. BMC Genomics. 2020 dec;21(1):673. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-07091-y.

[7] Mancini MC, Cardoso-Silva CB, Costa EA, Marconi TG, Garcia AAF, De Souza AP. New Developments in Sugarcane Genetics and Genomics. In: Buckeridge MS, De Souza AP, editors. Advances of Basic Science for Second Generation Bioethanol from Sugarcane. Cham: Springer International Publishing; 2017. p. 159–174. Available from: http://link.springer.com/10.1007/978-3-319-49826-3http://link.springer.com/10.1007/978-3-319-49826-3{_}9.

[8] Alexander AG. The energy cane alternative. Amsterdam, Netherlands: Elsevier Science Publishers B.V.; 1985.

[9] Casu RE, Jarmey JM, Bonnett GD, Manners JM. 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. 2007 feb;7(2):153–167. Available from: http://link.springer.com/10.1007/s10142-006-0038-z.

[10] Nishiyama MY, Ferreira SS, Tang PZ, Becker S, Pörtner-Taliana A, Souza GM. Full-length enriched cDNA libraries and ORFeome analysis of sugarcane hybrid and ancestor genotypes. PLoS ONE. 2014 sep;9(9):e107351. Available from: http://dx.plos.org/10.1371/journal.pone.0107351.

[11] Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, et al. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Scientific Reports. 2013 dec;3(1):3399. Available from: http://www.nature.com/articles/srep03399.

[12] Wood DLA, Nones K, Steptoe A, Christ A, Harliwong I, Newell F, et al. Recommendations for accurate resolution of Gene and isoform allele-specific expression in RNA-seq data. PLoS ONE. 2015;10(5):1–27.

[13] Jackson PA. Breeding for improved sugar content in sugarcane. Field Crops Research. 2005 jun;92(2-3):277–290. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0378429005000365.

[14] Hu X, Wang H, Diao X, Liu Z, Li K, Wu Y, et al. Transcriptome profiling and comparison of maize ear heterosis during the spikelet and floret differentiation stages. BMC Genomics. 2016;17(1):1–18. Available from: http://dx.doi.org/10.1186/s12864-016-3296-8.

[15] Balsalobre TWA, da Silva Pereira G, Margarido GRA, Gazaffi R, Barreto FZ, Anoni CO, et al. GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics. 2017 dec;18(1):72. Available from: http://dx.doi.org/10.1186/s12864-016-3383-xhttp://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3383-x.

[16] Garsmeur O, Droc G, Antonise R, Grimwood J, Potier B, Aitken K, et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nature Communications. 2018;9(1). Available from: http://dx.doi.org/10.1038/s41467-018-05051-5.

[17] Huang DL, Gao YJ, Gui YY, Chen ZL, Qin CX, Wang M, et al. Transcriptome of High-Sucrose Sugarcane Variety GT35. Sugar Tech. 2016;18(5):520–528.

[18] Ereful NC, Liu LY, Tsai E, Kao SM, Dixit S, Mauleon R, et al. Analysis of Allelic Imbalance in Rice Hybrids Under Water Stress and Association of Asymmetrically Expressed Genes with Drought-Response QTLs. Rice. 2016;9(1). Available from: http://dx.doi.org/10.1186/s12284-016-0123-4.

[19] Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, et al. Characterization of the Yeast Transcriptome. Cell. 1997 jan;88(2):243–251. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867400818450.

[20] Gazaffi R, Oliveira KM, de Souza AP, Garcia AAF. Melhoramento genético e mapeamento da cana-de-açúcar. In: Cortez LAB, editor. Bioetanol de Cana-de-Açúcar: P&D para produtividade e sustentabilidade; 2010. p. 333–344.

[21] Cavalett O, Chagas MF, Junqueira TL, Watanabe MDB, Bonomi A. Environmental impacts of technology learning curve for cellulosic ethanol in Brazil. Industrial Crops and Products. 2017 nov;106:31–39. Available from: http://dx.doi.org/10.1016/j.indcrop.2016.11.025http://linkinghub.elsevier.com/retrieve/pii/S0926669016307695.

[22] Creste S, Xavier MA, Landell MGA. Importância do germoplasma no desenvolvimento de cultivares de cana-de-açúcar com perfil agroenergético. In: Cortez LAB, editor. Bioetanol de Cana-de-Açúcar: P&D para produtividade e sustentabilidade; 2010. p. 313–317.

[23] Romanel A, Lago S, Prandi D, Sboner A, Demichelis F. ASEQ: Fast allele-specific studies from next-generation sequencing data. BMC Medical Genomics. 2015;8(1):1–12.

[24] Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biology. 2015;16(1):1–12. Available from: http://dx.doi.org/10.1186/s13059-015-0762-6.

[25] Pereira GS, Garcia AAF, Margarido GRA. A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC Bioinformatics. 2018;19(1):1–10.

[26] Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nature Genetics. 2018;50(11):1565–1573.

[27] Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009;25(24):3207–3212.

[28] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14(1):91. Available from: http://www.biomedcentral.com/1471-2105/14/91/abstract{\%}5Cnhttp://www.biomedcentral.com/1471-2105/14/91{\%}5Cnhttp://www.biomedcentral.com/content/pdf/1471-2105-14-91.pdfhttp://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-91.

[29] Barreto FZ, Rosa JRBF, Balsalobre TWA, Pastina MM, Silva RR, Hoffmann HP, et al. A genome-wide association study identified loci for yield component traits in sugarcane (Saccharum spp.). PLOS ONE. 2019 jul;14(7):e0219843. Available from: http://dx.plos.org/10.1371/journal.pone.0219843.

[30] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57–63.

[31] Matsuoka S, Kennedy AJ, dos Santos EGD, Tomazela AL, Rubio LCS. Energy Cane: Its Concept, Development, Characteristics, and Prospects. Advances in Botany. 2014;2014:1–13. Available from: http://www.hindawi.com/archive/2014/597275/.

[32] Kasirajan L, Hoang NV, Furtado A, Botha FC, Henry RJ. Transcriptome analysis highlights key differentially expressed genes involved in cellulose and lignin biosynthesis of sugarcane genotypes varying in fiber content. Scientific Reports. 2018;8(1):1–16.

[33] Riaño-Pachón DM, Mattiello L. Draft genome sequencing of the sugarcane hybrid SP80-3280. F1000Research. 2017;6(0):861. Available from: https://f1000research.com/articles/6-861/v1.

[34] Serang O, Mollinari M, Garcia AAF. Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS ONE. 2012;7(2):1–13.

[35] Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015 may;348(6235):666–669. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1261877.

[36] Tuch BB, Laborde RR, Xu X, Gu J, Chung CB, Monighetti CK, et al. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. PLoS ONE. 2010;5(2).

[37] Lee C, Kang EY, Gandal MJ, Eskin E, Geschwind DH. Profiling allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage. Nature Neuroscience. 2019;22(9):1521–1532. Available from: http://dx.doi.org/10.1038/s41593-019-0461-9.

[38] Grivet L, Arruda P. Sugarcane genomics: Depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology. 2002;5(2):122–127.

[39] Moore PH, Paterson AH, Tew T. Sugarcane: The Crop, the Plant, and Domestication. In: Sugarcane: Physiology, Biochemistry, and Functional Biology. Chichester, UK: John Wiley & Sons Ltd; 2013. p. 1–17. Available from: http://doi.wiley.com/10.1002/9781118771280.ch1.

[40] Daniels J, Roach BT. Taxonomy and evolution. In: Heinz DJ, editor. Developments in Crop Science. vol. 11 of Developments in Crop Science. Elsevier; 1987. p. 7–84. Available from: http://www.sciencedirect.com/science/article/pii/B9780444427694500072.

[41] Carson DL, Huckett BI, Botha FC. Differential gene expression in sugarcane leaf and internodal tissues of varying maturity. South African Journal of Botany. 2002 dec;68(4):434–442. Available from: c:{\%}5CDocumentsandSettings{\%}5Ccas128{\%}5CMyDocuments{\%}5CDownloadedpapers{\%}5CSAJB-2002-68-434-442.pdf{\%}5Cnhttp://linkinghub.elsevier.com/retrieve/pii/S0254629915303707https://linkinghub.elsevier.com/retrieve/pii/S0254629915303707.

[42] Pham GM, Newton L, Wiegert-Rininger K, Vaillancourt B, Douches DS, Buell CR. Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. Plant Journal. 2017;92(4):624–637.

[43] Carson DL, Botha FC. Preliminary Analysis of Expressed Sequence Tags for Sugarcane. Crop Science. 2000;40(6):1769. Available from: https://www.crops.org/publications/cs/abstracts/40/6/1769.

[44] Jardim-Messeder D, Felix-Cordeiro T, Barzilai L, de Souza-Vieira Y, Galhego V, Bastos GA, et al. Genome-wide analysis of general phenylpropanoid and monolignol-specific metabolism genes in sugarcane. Functional & Integrative Genomics. 2021 jan;21(1):73–99. Available from: http://link.springer.com/10.1007/s10142-020-00762-9.

[45] Simões MS, Ferreira SS, Grandis A, Rencoret J, Persson S, Floh EIS, et al. Differentiation of Tracheary Elements in Sugarcane Suspension Cells Involves Changes in Secondary Wall Deposition and Extensive Transcriptional Reprogramming. Frontiers in Plant Science. 2020 dec;11(December):1–19. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2020.617020/full.

[46] Llerena JPP, Figueiredo R, Brito MdS, Kiyota E, Mayer JLS, Araujo P, et al. Deposition of lignin in four species of Saccharum. Scientific Reports. 2019 dec;9(1):5877. Available from: http://dx.doi.org/10.1038/s41598-019-42350-3http://www.nature.com/articles/s41598-019-42350-3.

[47] Yang X, Todd J, Arundale R, Binder JB, Luo Z, Islam MS, et al. Identifying loci controlling fiber composition in polyploid sugarcane (Saccharum spp.) through genome-wide association study. Industrial Crops and Products. 2019 apr;130(January):598–605. Available from: https://doi.org/10.1016/j.indcrop.2019.01.023https://linkinghub.elsevier.com/retrieve/pii/S0926669019300305.

[48] Wang J, Nayak S, Koch K, Ming R. Carbon partitioning in sugarcane (Saccharum species). Frontiers in Plant Science. 2013;4(June):2005–2010. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2013.00201/abstract.

[49] Scortecci KC, Creste S, Jr TC, Xavier MA, Landell MGA, Figueira A, et al. Challenges, Opportunities and Recent Advances in Sugarcane Breeding. In: Plant Breeding. InTech; 2012. p. 352. Available from: http://www.intechopen.com/books/plant-breeding/challenges-opportunities-and-recent-advances-in-sugarcane-breeding.

[50] Grativol C, Regulski M, Bertalan M, McCombie WR, Da Silva FR, Zerlotini Neto A, et al. Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus Saccharum. Plant Journal. 2014 jul;79(1):162–172. Available from: http://doi.wiley.com/10.1111/tpj.12539.

[51] Piperidis N, D'Hont A. Sugarcane genome architecture decrypted with chromosome-specific oligo probes. The Plant Journal. 2020 jul:tpj.14881. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.14881.

[52] Piperidis G, Piperidis N, D'Hont A. Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. Molecular Genetics and Genomics. 2010 jul;284(1):65–73. Available from: http://link.springer.com/10.1007/s00438-010-0546-3.

[53] Matsuoka S, Ferro J, Arruda P. The Brazilian experience of sugarcane ethanol industry. In Vitro Cellular & Developmental Biology - Plant. 2009 jun;45(3):372–381. Available from: http://link.springer.com/10.1007/978-1-4419-7145-6{_}9http://link.springer.com/10.1007/s11627-009-9220-z.

[54] Vicentini R, Bottcher A, Dos Santos Brito M, Dos Santos AB, Creste S, De Andrade Landell MG, et al. Large-scale transcriptome analysis of two sugarcane genotypes contrasting for lignin content. PLoS ONE. 2015 aug;10(8):e0134909. Available from: http://dx.plos.org/10.1371/journal.pone.0134909.

[55] Irvine JE. Saccharum species as horticultural classes. Theoretical and Applied Genetics. 1999 feb;98(2):186–194. Available from: http://dx.doi.org/10.1007/s001220051057http://link.springer.com/10.1007/s001220051057.

[56] Cardoso-Silva CB, Costa EA, Mancini MC, Balsalobre TWA, Costa Canesin LE, Pinto LR, et al. De novo assembly and transcriptome analysis of contrasting sugarcane varieties. PLoS ONE. 2014;9(2).

[57] Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MIT, et al. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Research. 2003 dec;13(12):2725–2735. Available from: http://www.genome.org/cgi/doi/10.1101/gr.1532103.

[58] Okura VK, de Souza RSC, de Siqueira Tada SF, Arruda P. BAC-Pool Sequencing and Assembly of 19 Mb of the Complex Sugarcane Genome. Frontiers in Plant Science. 2016 mar;7(March):342. Available from: http://journal.frontiersin.org/Article/10.3389/fpls.2016.00342/abstract.

[59] Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L, Thirugnanasambandam PP, et al. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genomics. 2017;18(1):395. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3757-8.

2 DIFFERENTIAL EXPRESSION IN LEAVES OF *Saccharum* GENOTYPES CONTRASTING IN BIOMASS PRODUCTION PROVIDES EVIDENCE OF GENES INVOLVED IN CARBON PARTITIONING

Abstract

Background: The development of biomass crops aims to meet industrial yield demands, in order to optimize profitability and sustainability. Achieving these goals in an energy crop like sugarcane relies on breeding for sucrose accumulation, fiber content and stalk number. To expand the understanding of the biological pathways related to these traits, we evaluated gene expression of two groups of genotypes contrasting in biomass composition.

Results: First visible dewlap leaves were collected from 12 genotypes, six per group, to perform RNA-Seq. We found a high number of differentially expressed genes, showing how hybridization in a complex polyploid system caused extensive modifications in genome functioning. We found evidence that differences in transposition and defense related genes may arise due to the complex nature of the polyploid *Saccharum* genomes. Genotypes within both biomass groups showed substantial variability in genes involved in photosynthesis. However, most genes coding for photosystem components or those coding for *phosphoenolpyruvate carboxylases* (PEPCs) were upregulated in the high biomass group. *Sucrose synthase* (SuSy) coding genes were upregulated in the low biomass group, showing that this enzyme class can be involved with sucrose synthesis in leaves, similarly to *sucrose phosphate synthase* (SPS) and *sucrose phosphate phosphatase* (SPP). Genes in pathways related to biosynthesis of cell wall components and *expansins* coding genes showed low average expression levels and were mostly upregulated in the high biomass group.

Conclusions: Together, these results show differences in carbohydrate synthesis and carbon partitioning in the source tissue of distinct phenotypic groups. Our data from sugarcane leaves revealed how hybridization in a complex polyploid system resulted in noticeably different transcriptomic profiles between contrasting genotypes.

Keywords: Sugarcane; Gene expression; Transcriptomics; RNA-Seq; Polyploid.

2.1 Background

Bioenergy crops are cultivable species with favorable traits as feedstocks for the production of energy [39]. One such biofuel is ethanol, which is produced from the conversion of plant carbohydrates. The disaccharide sucrose is easily converted into ethanol by fermentation, but starch and lignocellulosic polymers have to be converted into monosaccharides prior to fermentation [39, 26]. Lignocellulosic biomass must be disrupted with enzymatic or physical methods as a pretreatment to form a hydrolysable material [26]. Sugarcane culms have been used to produce ethanol from sugar juice fermentation and bagasse, which is also burned to generate electricity. As a result, sugarcane leaves form part of the straw remaining in the field after harvesting. This residual can be used as a biomass source in mills or deposited on the soil to form organic matter. Thus, leaves are a potential biomass supplement to increase the energy supply [42, 15].

Sugarcane species are members of the genus *Saccharum*, of the Poaceae family. There are two ancestral species, *S. robustum* and *S. spontaneum*. The former was the ancestor of the cultivated *S. officinarum* and *S. edule* [5, 38]. Other two cultivated species, *S. barberi* and *S. sinense*, are derived from crosses between *S. officinarum* and *S. spontaneum* [5, 38]. Genotypes of *S. officinarum* were used for cultivation due to their high capacity to produce and store sucrose. Sugarcane stalks are the primary

source of sucrose for industrial purposes and have historically been the main target of breeding efforts [62]. Later, crosses of *S. officinarum* with *S. spontaneum* were proposed to avoid abiotic and biotic stresses. Recently, breeding programs have directed efforts to obtain more fibrous genotypes - the so-called energy canes. Because wild genotypes show substantial variability [48, 17], they can be used as a source to introgress traits such as fiber content and stalk number, increasing total biomass yield [59].

Modern sugarcane breeding can benefit from a molecular framework to unravel the underlying genetic basis of important traits. Polyploidy is an inherent characteristic of the *Saccharum* genomes, with *S. officinarum* presenting 80 chromosomes (2n = 8x = 80) and ancient genotypes with a large chromosome number variation [32]. More than 80% of the chromosomes of modern hybrids come from *S. officinarum*, 10-20% from *S. spontaneum* and the remaining are recombinants. There is also aneuploidy in the homeologous groups [34]. The high ploidy in cultivars results in a complex genome of 10 Gbp, that can be represented by an x = 10 monoploid genome [38]. Despite this genomic complexity, progress has been achieved in understanding the role of proteins in carbon partitioning to sucrose or cell wall. Several studies have investigated gene expression to improve understanding of changes in pathways among different plant parts. This has identified the expression of enzymes involved in sucrose metabolism [6, 33], like *sucrose synthase*, that can show organ-specific expression patterns [50, 45]. The expression of genes coding proteins related to cellulose, hemicellulose and lignin metabolism was explored by comparing genotypes contrasting in biomass or in cell wall-related traits [36, 24]. Genes coding for enzymes of the lignin pathway were stimulated in a high-biomass genotype [24], and their expression levels were higher in bottom rather than top internodes [36]. Singh and colleagues [13] found that high-biomass genotypes of an F2 population were more photosynthetically active, as a result of the upregulation of genes coding for photorespiration, Calvin cycle and light reaction proteins.

A wide range of functional categories have been found in studies of gene expression in sugarcane leaves including transporter activity, regulation, response to stimulus and to stress [33, 10]. In addition to their direct use as a biomass source, leaves are the source tissue with which plants produce photoassimilates used to maintain leaf activities and for cell wall synthesis or sucrose accumulation in vacuoles of the stalks and sink organs [37]. Determining the regulation of genes functionally related to biomass-associated traits has value for potential biotechnological applications [39]. To achieve this, we must enhance our knowledge about genes involved in processes of carbohydrate metabolism, especially those related to production of sucrose and lignocellulosic components. To that end, we evaluated the transcriptomes of twelve diverse sugarcane genotypes divided into two contrasting biomass groups. The broad diversity of these genotypes is reflected by the presence of four *S. spontaneum*, a *S. robustum*, two *S. officinarum* representatives and five hybrid cultivars. The five hybrid cultivars come from different genetic backgrounds, from breeding programs in Argentina, Brazil and the United States. In addition to investigating differential gene expression between the two groups, we aimed to identify biological processes that differed between the genotypes within each group.

2.2 Results

2.2.1 Data summary

Leaf samples were collected from field-grown plants with six months of age, from twelve different genotypes assigned to two groups contrasting in sucrose-associated traits - soluble solids content, sucrose and purity - and biomass-associated traits - fiber content and number of stalks (Fig. 1 and Additional file 1 - Figure 1). These figures show a group with four *S. spontaneum* representatives - IN84–58, IN84–88, Krakatau and SES205A -, the *S. robustum* genotype IJ76–318 and the hybrid US85-1008. The second group was formed by genotypes that have higher sucrose levels in culms: two *S. officinarum* genotypes - White Transparent and Criolla Rayada -, the hybrid TUC71–7 and more modern hybrids - RB72454,

SP80–3280, and RB855156. For simplicity, we will refer to the main difference between the two groups in terms of biomass. Therefore, these genotypes were chosen to include accessions of different *Saccharum* species to form two groups contrasting in biomass content. Although cytogenetic information is limited for sugarcane genotypes, we do expect differences in chromosome numbers and ploidy level among them. Most hybrids, with the exception of US85–1008, have a larger number of *S. officinarum* chromosomes and a minor and variable contribution of *S. spontaneum*, likely with a basic chromosome number of x = 10 [60]. The basic chromosome number of *S. officinarum* is also x = 10, but different numbers have been verified in *S. spontaneum* [60]. Ploidy levels and interspecific hybridization have the potential to affect gene expression patterns, in addition to mechanisms of transcriptional control and epigenetic factors [44, 1]. Nevertheless, our study aimed to find direct associations between transcript abundance and phenotypic traits, without trying to identify the upstream causes of differences in gene expression levels. Our analyses do not depend on prior knowledge about the ploidy of each accession, but we note that variation in chromosome copy counts are possible causes for similarities or differences between particular genotypes.

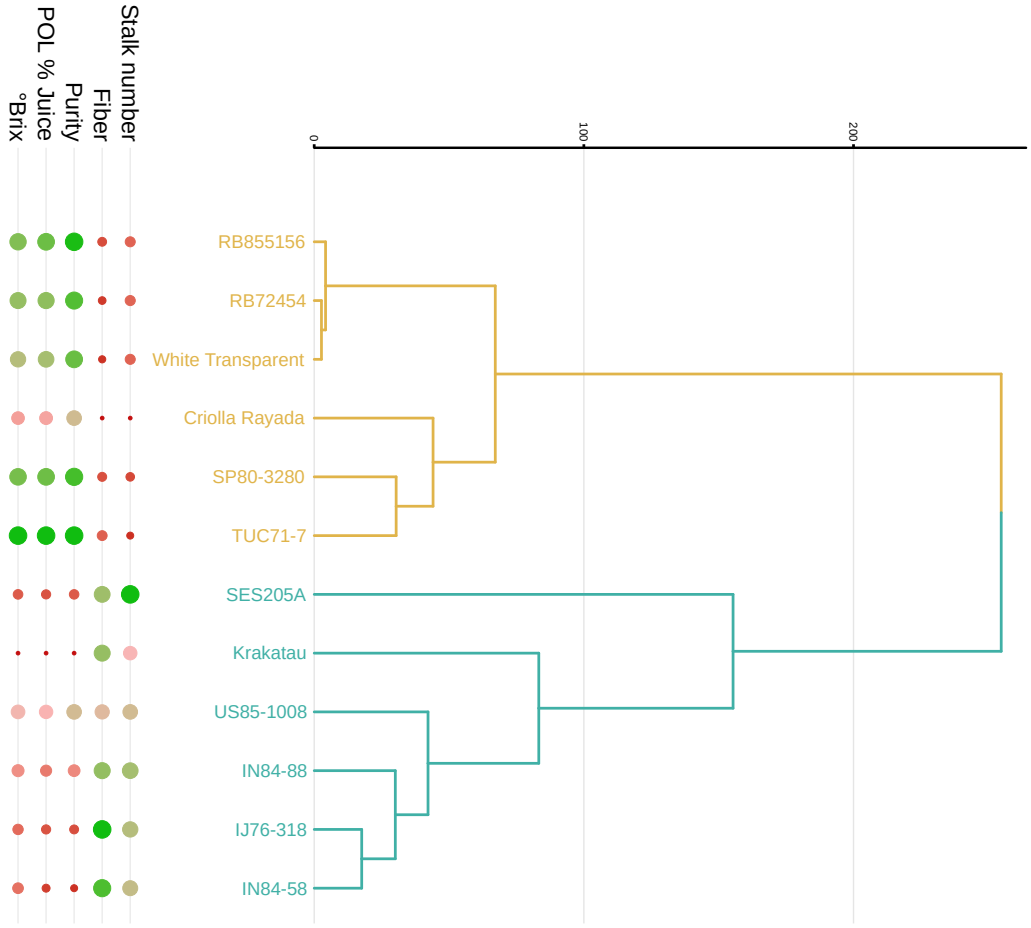

Figure 1: Dendrogram of the twelve sugarcane genotypes based on phenotypic traits. We performed a hierarchical clustering of the genotypes based on Euclidean distances calculated for all evaluated traits. Points at the bottom represent the gradient of the scaled phenotypic measures of each accession, where larger green points represent higher phenotypic values. The measured phenotypic traits include: content of soluble solids in the cane juice (°Brix); polarization or sucrose percentage in the juice (POL % Juice); percentage of sucrose in the total solids of the juice (Purity); percentage of fiber in the bagasse (Fiber); and the number of stalks in each plot

The mapping rate of sequenced libraries ranged from 80.52 to 85.37% (Table 1 in Additional file 3). To characterize the variability in the expression profiles, we initially assessed the distances between

samples based on gene expression levels, using the multidimensional scaling plot to identify clusters. We noted that clonal genotype replicates were close to each other, as expected (Fig. 2). As was the case for phenotypic traits (Fig. 1 in Additional file 1), the first dimension basically separated the high and low biomass groups, and genotypes of the former were farther from each other, revealing higher gene expression variability within the high biomass group. US85–1008 samples clustered between the two groups, apparently reflecting the origin of this genotype in a breeding program. Investigation of the low biomass group (Fig. 2) showed that RB855156 was close to TUC71–7, most likely because it was originated as a hybrid between RB72454 and TUC71–7. In fact, the Brazilian hybrids are closely related, because RB72454 is the offspring of CP53–76 (used as the maternal parent), which is also the maternal grandfather of SP80–3280. The second dimension separated the high biomass genotypes in three sets: i) SES205A at the top; ii) Krakatau, IN84–88 and US85–1008 in the middle; and iii) IN84–58 and IJ76–318. Curiously, in the latter group, an accession classified as *S. robustum* (IJ76–318) grouped closely with a *S. spontaneum* genotype. Variability within the low biomass group is clearly verified if a third dimension is added (Fig. 1 in Additional file 3), in which the most extreme genotypes were RB72454 and SP80–3280 - phenotypically close to each other (Figure 1 in Additional file 1). This result indicates that distances among the low biomass genotypes are smaller than among the high biomass accessions.

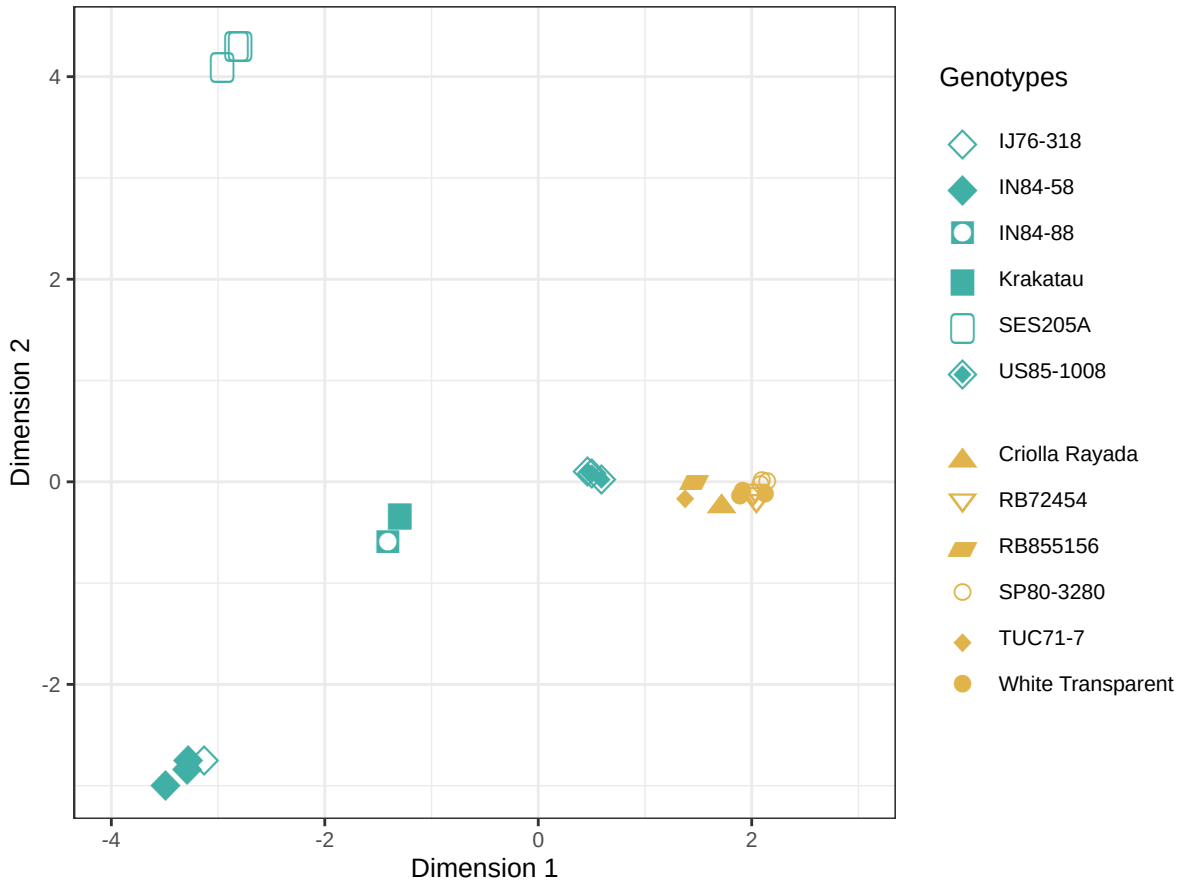

Figure 2: Multidimensional scaling plot to assess dissimilarities between samples. Points in blue represent the high biomass genotypes, while the ones within the low biomass group members are tagged in orange. Different shapes represent different genotypes within each group. Note that three genotypes in each group are represented by three clonal replicates

We first tested for differences in gene expression levels between the two biomass groups, taking the high biomass group as reference. This resulted in 10,903 downregulated and 10,171 upregulated genes in the low biomass group. In this model, the dispersion estimate includes biological variation between

all samples in both groups. This resulted in a biological coefficient of variation (BCV) of 0.86. Although the test within the high biomass group resulted in a BCV of 0.31, more genes were deemed differentially expressed than comparing the groups (Table 2 in Additional file 3). In accordance to the similarity among genotypes, the test within the low biomass group had a similar BCV (0.27) and the lowest number of differentially expressed genes (DEGs) among the three contrasts. Assessing the overlap between these lists of genes, the higher number of unique DEGs occurred when testing for differences among the high biomass genotypes (Figure 2 in Additional file 3), which is consistent with the higher variability among them.

Enrichment analysis was used to assess if functional categories are overrepresented among DEGs, giving evidence of widespread changes in the transcriptional landscape of biological pathways. Functional enrichment analysis with DEGs from the comparison between biomass groups revealed changes in translation and DNA integration - which is a parent term of transposon integration in the Gene Ontology (GO) hierarchy (Figure 3 in Additional file 3). The tests comparing genotypes within the two groups showed many enriched GO terms related to transposition, defense-related and carbohydrate-related (Figs. 3 and 4). Differential expression of transposition-associated genes was more marked when contrasting the two biomass groups and within the high biomass genotypes (Figure 4 in Additional file 3). Also, the high biomass genotypes showed significant differences in the expression level of genes related to cell division, replication and post-replication repair terms. On the other hand, in addition to DEGs related to replication, transcription and kinases, the test within the low biomass group revealed differences in *O-methyltransferase activity* (Figure 4). The molecular function *glutathione transferase activity* was enriched in both within-group contrasts (Figs. 3 and 4). We also found changes in genes coding for proteins involved in the response to salicylic acid in both tests.

A functional enrichment test performed with the common DEGs detected in the three contrasts corroborates defense response and transposition, as well as gives evidence of a possible genomic stress (Figure 5 in Additional file 3). Using the 7350 DEGs in the pairwise intersection of within-groups contrasts, enrichment analysis revealed changes in the synthesis of cell wall (Figure 6 in Additional file 3).

2.2.2 Co-expressed genes and metabolic pathways

We identified 16 modules with co-expressed genes, with the number of genes in each module ranging from 514 to 7814. Functional analyses among annotated co-expressed genes in each set revealed enriched GO terms in eleven of these modules (Table 3 in Additional file 3). We identified an overlap of translation- and transcription-related terms predominantly in modules one and seven, such as those involved in the assembly of ribosomal subunits, protein processing, protein degradation and processing of RNAs (Table 3 and Figure 7 in Additional file 3).

Cellular components of chloroplasts were found in five modules of the network: three, seven, eight, eleven and sixteen (Table 3 in Additional file 3). Module 16 was mostly formed by genes related to chloroplast, photosystem and photosynthesis (Figure 7 in Additional file 3). This was the only module to show enrichment of responses to hormones (abscisic acid, cytokinin, ethylene and gibberellin) and these DEGs were mainly repressed in high biomass genotypes (Figure 8 in Additional file 3). We noticed that many genes in module 16 showed high absolute log fold change (LFC) values in all three contrasts, but to a lesser extent in the comparison between *S. officinarum* and the low biomass hybrids (Figure 9 in Additional file 3). This is explained by the expression profile of the genes present in this module, for which the expression level in the low biomass group was higher and similar among the samples (Figure 10 in Additional file 3).

The results of the comparison between the main groups identified up and downregulated DEGs

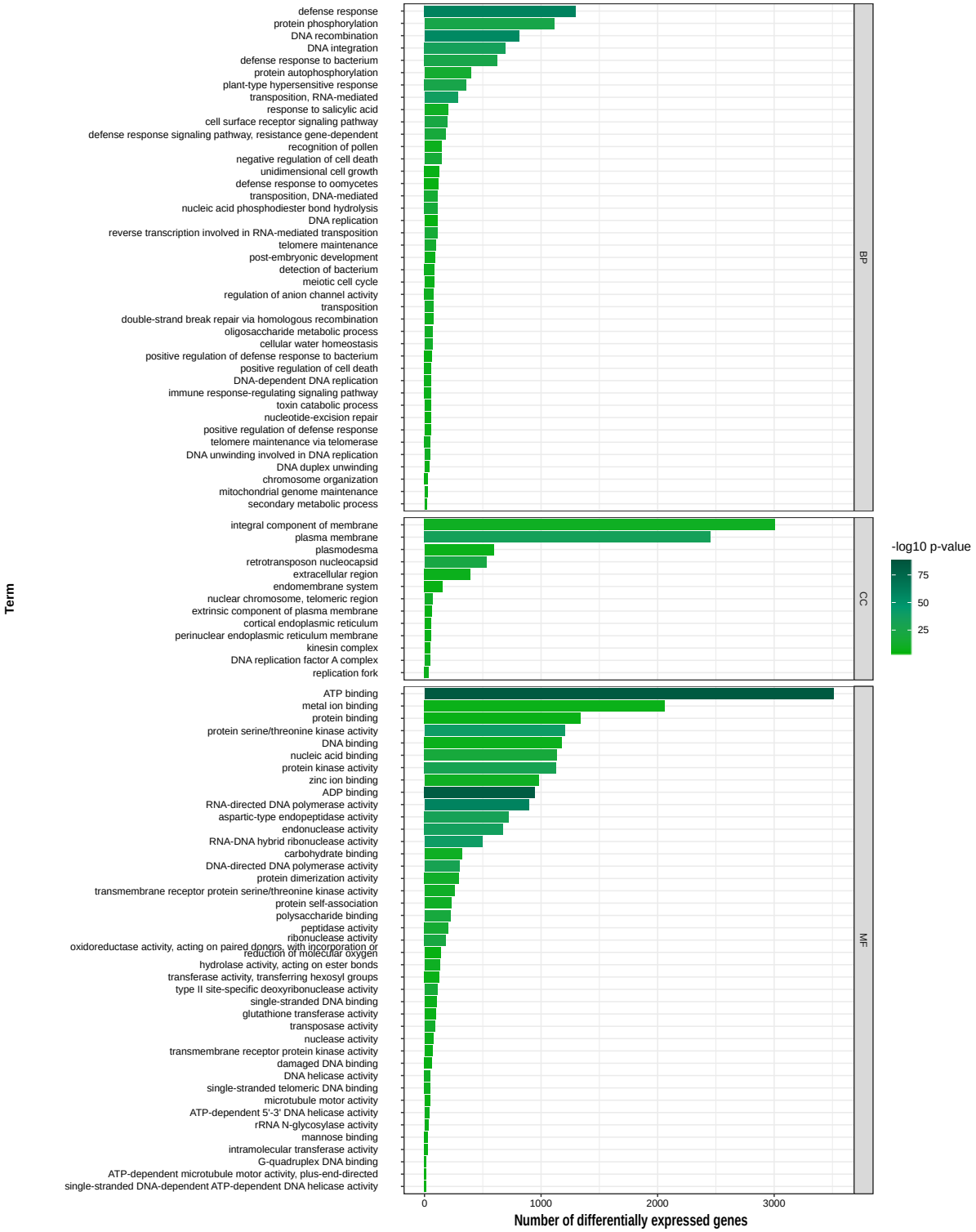

Figure 3: Bar chart of the number of DEGs in each enriched functional class for the differences within the high biomass group. Bars show the number of differentially expressed genes in each Gene Ontology term. Smaller p-values are shown by darker green colors. Terms were grouped by the categories BP (Biological Process), CC (Cellular Component) and MF (Molecular Function)

in all metabolic processes provided by the MapMan4 functional BINs (Figure 11 in Additional file 3). Many genes involved in photophosphorylation were downregulated in the low biomass group, annotated as components of the *photosystem II (Psb) proteins*, *photosystem I (Psa)* and *cytochrome (Pet) subunits* and *photosystem I assembly* (YCF3 and YCF4) (Figure 12 in Additional file 3). Other genes of the pho-

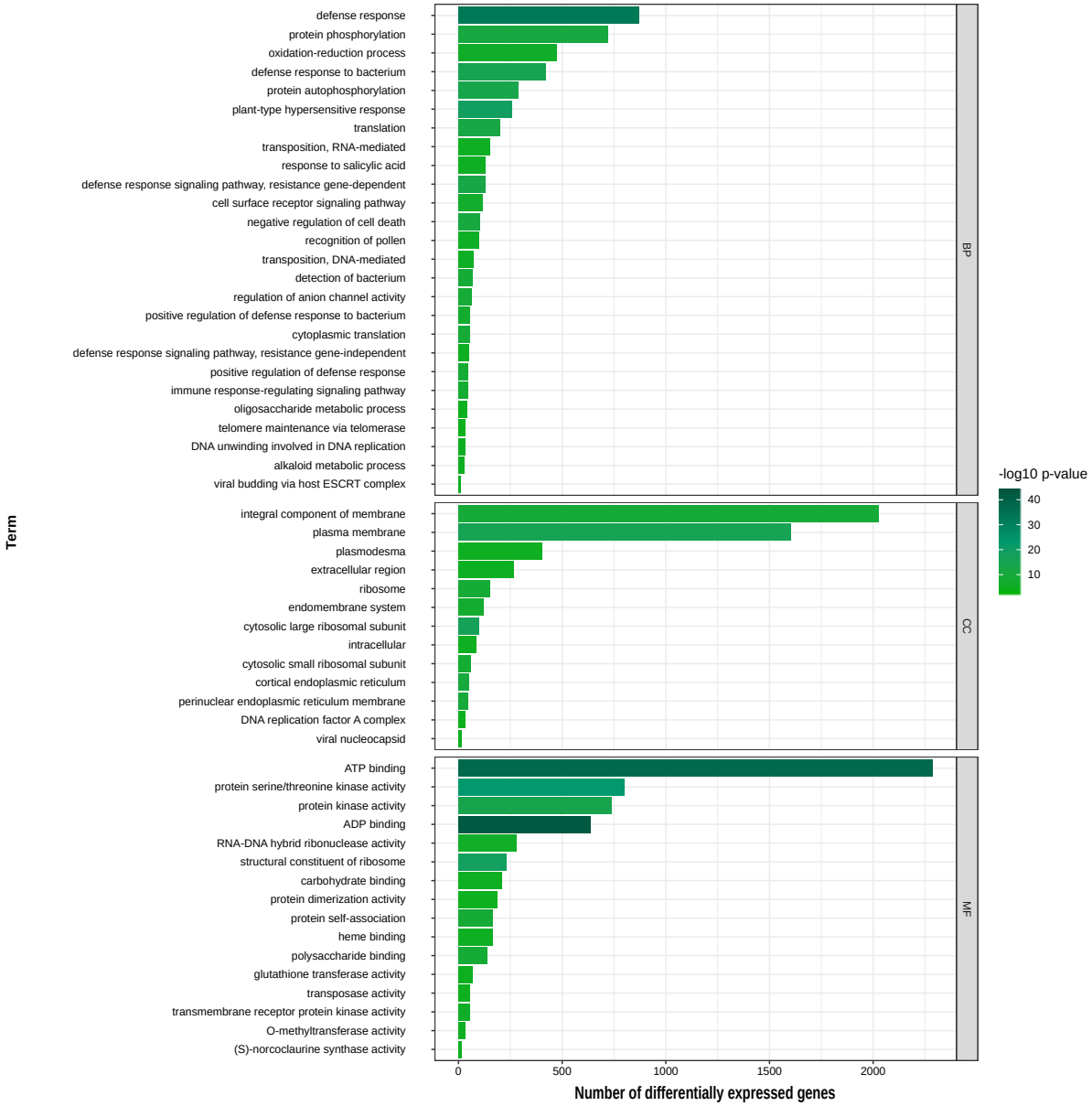

Figure 4: Bar chart of the number of DEGs in each enriched functional class for the differences within the low biomass group. Bars show the number of differentially expressed genes in each Gene Ontology term. Smaller p-values are shown by darker green colors. Terms were grouped by the categories BP (Biological Process), CC (Cellular Component) and MF (Molecular Function)

tosynthesis light reactions were differentially expressed within the two groups, in both cases consistently upregulated in the genotypes with the lowest fiber content (Figure 13 and Figure 14 in Additional file 3). However, genes coding for proteins acting on C4/CAM photosynthesis were downregulated in White Transparent (Figure 14 in Additional file 3). This is in accordance with our co-expression analysis, where many photosynthesis genes with high LFC were present in low biomass genotypes and in US85–1008, but were non-DE when White Transparent was compared to low biomass hybrids (Figure 9 in Additional file 3). DEGs coding for *phosphoenolpyruvate carboxylase* (PEPC) were repressed in low biomass genotypes, being expressed at similar levels in the high biomass accessions (Figure 15 in Additional file 3).

Compared to the high biomass group, low biomass genotypes showed lower expression of genes related to secondary metabolism, such as those annotated to the monolignol synthesis (Figure 16 in Additional file 3). However, the MapMan4 lignin pathway revealed upregulation of certain enzymes

in the low biomass genotypes: *phenylalanine ammonia lyase* (PAL), *caffeic acid O-methyltransferase* (COMT), *4-coumarate: CoA ligase* (4CL), *cinnamyl-alcohol dehydrogenase* (CAD) and a *β-glucosidase* (Figure 17 in Additional file 3). US85–1008 and the wild *S. spontaneum* genotypes were similar in the expression of genes coding for enzymes of the lignin metabolism, with significant differences for five genes - a 4CL, a *β-glucosidase*, *a Caffeoyl-CoA O-methyltransferase* and two *cinnamoyl-Coa reductases* (CCR) (Figure 18 in Additional file 3).

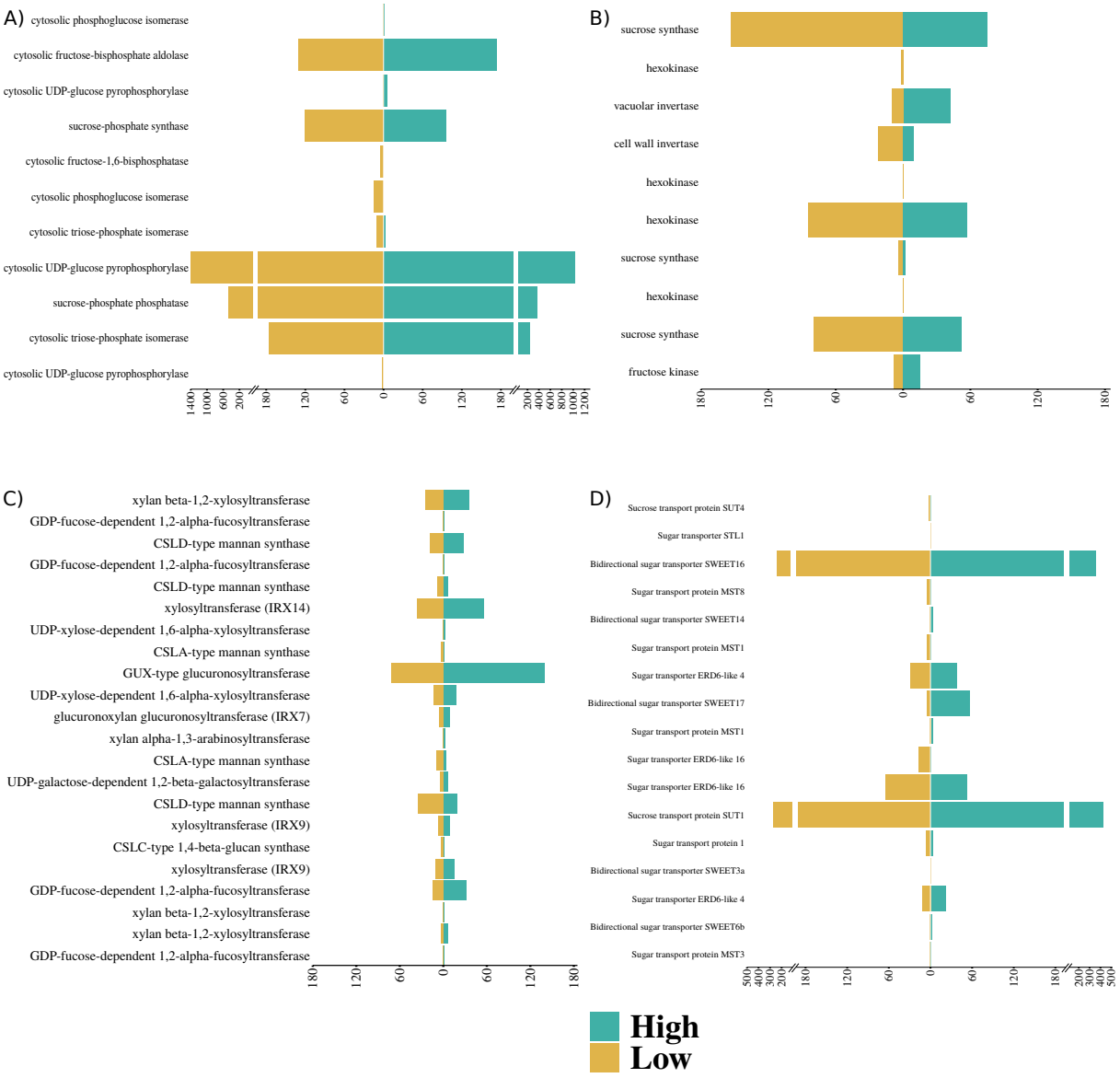

Figure 5: Expression of DEGs involved with sucrose metabolism: synthesis (**a**); degradation (**b**); synthesis of cell wall compounds (**c**); and sucrose and sugar transporters (**d**). Gene expression in each biomass group was calculated using the mean of the normalized counts per million. Note that the scale is different among plots. The high biomass group is colored in blue (right side) and the low biomass group in orange (left side)

We observed that many genes coding for enzymes acting on xylan were upregulated in high biomass genotypes, even in the within-group comparisons (Fig. 5c and Additional file 3 - Figure 19). Regarding cell modification and degradation, a *1,6-alpha-xylosidase* was highly expressed in the low biomass group (Figure 19-B in Additional file 3). Genes annotated with *xylosyltransferase activity* were co-expressed with those involved with the Golgi apparatus, membrane components and endocytosis, being

more highly expressed in high biomass genotypes (Table 3 - Module 10 and Figure 10 in Additional file 3). This is expected given that the Golgi apparatus synthesizes most polysaccharides of the cell wall, where transferases catalyze the synthesis of the xyloglucan backbone and side branches [51]. We also found significant differences in the expression levels of genes associated with cell wall flexibility. In particular, DEGs coding for expansins of the β subfamily were more highly expressed in *S. spontaneum* and *S. robustum* (Figure 20 in Additional file 3).

The biomass groups revealed different expression levels of genes coding for enzymes of sucrose metabolism. *Sucrose-phosphate synthase* (SPS) and *sucrose-phosphate phosphatase* (SPP) genes were upregulated in low biomass genotypes (Fig. 5a). Curiously, genes coding for *sucrose synthase* (SuSy) - an enzyme family mainly involved with sucrose degradation - were upregulated in the low biomass group and in US85–1008 (Fig. 5b and Additional file 3 - Figure 21). The comparison between groups also showed different expression levels of genes coding for sucrose transport proteins SUT1 and SUT4. Although SUT4 was strongly upregulated in the low biomass group (Figure 22 in Additional file 3), SUT1 was highly expressed in the high biomass genotypes (Fig. 5d). We found different expression profiles of genes coding for sugar transporters of the same family. Genes coding for SWEETs (Sugars will eventually be exported transporters) were downregulated in the low biomass group, while within the groups these DEGs showed a genotype-specific expression (Figure 22-B in Additional file 3).

2.2.3 Assessing gene expression at different levels

We evaluated how processes are functionally enriched according to the quantification method grouping counts at the gene or transcript level, considering only the contrast between the two main biomass groups. For both approaches, around 30% of each reference set (transcripts or genes) passed the minimum expression threshold (Table 1 in Additional file 4). For 5886 DEGs, none of their corresponding individual transcripts showed statistically significant evidence of differential expression. On the other hand, 8693 genes showed at least one DET, but were not differentially expressed when read counts were gathered at the gene level (Figure 1 in Additional file 4). In addition to the six functional terms enriched among DEGs, analysis of differentially expressed transcripts (DETs) revealed enrichment of another 44 terms (Table 3 in Additional file 4). *Geranylgeranyl-Diphosphate Geranylgeranyltransferase* enrichment indicates changes in the synthesis of geranylgeranyl, a precursor of chlorophyll, carotenoids and gibberellins via the 2-C-methyl-D-erythritol 4-phosphate pathway. This is reinforced by the enrichment of *phytoene synthase*, acting on geranylgeranyl diphosphate in the carotenoid synthesis pathway. We also found enrichment of enzymes acting on precursors of sterols, in the isoprenoid biosynthesis pathway: *farnesyl-diphosphate farnesyltransferase activity* and *squalene synthase activity*. Two non-DEGs coding for *glyceraldehyde-3-phosphate dehydrogenases* (GAPDH) showed five DETs, and the DET with the higher expression level was upregulated in high biomass genotypes (Figure 4 in Additional file 4). Enrichment of GAPDH activity can likely be associated to the photosynthetic carbon reduction promoted by this enzyme, because we found DETs annotated as chloroplastic GAPDHs (Figure 4 in Additional file 4).

Combining the expression levels of DETs to obtain gene-level quantifications can result failure to detect DEGs, masking important functional changes. As an example, we considered the annotated genes of the *photosynthesis* biological process. We found five DEGs without any corresponding DETs – in fact, individual transcripts for three of these genes did not pass the expression filter, due to their low expression level (Figure 3 in Additional file 4). At the same time, 47 non-DE genes revealed at least one DET (Fig. 6). Lowly expressed isoforms did show significant differential expression when the fold changes were very high, i.e., when expression occurred almost entirely in one of the biomass groups (Fig. 6).

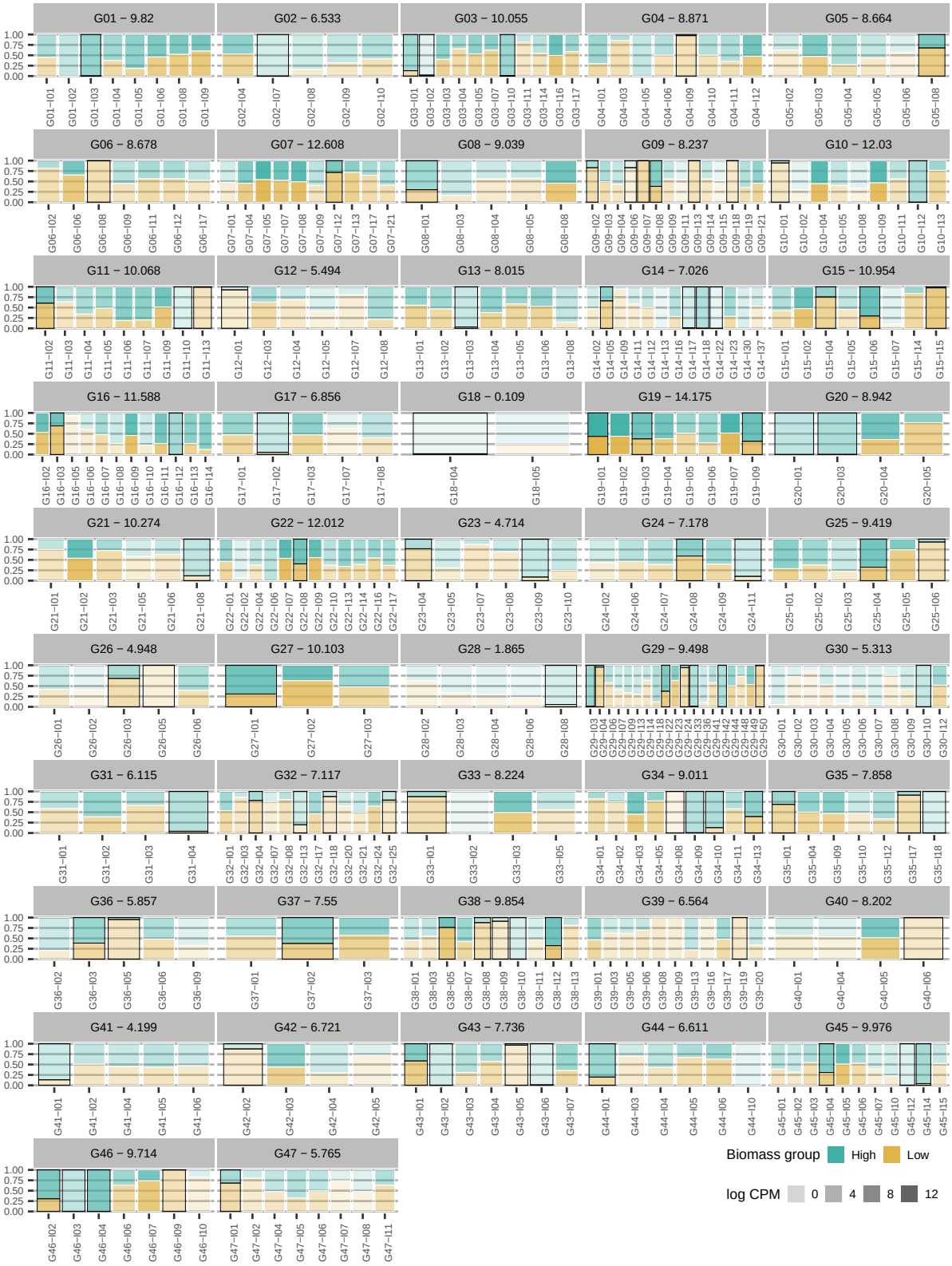

Figure 6: Expression profiles of differentially expressed transcripts of photosynthesis-related genes. Differential expression at the gene level was not significant for the corresponding genes. For each isoform, bar lengths correspond to the relative expression levels in each biomass group. Color intensity represents the logarithm of the counts per million (cpm) of the corresponding transcript. For each gene identifier we also show the log2 of the average counts per million. Differentially expressed transcripts are indicated by black edges

2.3 Discussion

Clustering based on gene expression profiles grouped samples in accordance to their phenotypic measures, but also revealed differences within the groups. A higher BCV when contrasting groups was expected because we used different genotypes as replicates of the same group (Fig. 2). The two within-group contrasts are relevant to capture differences between hybrids and wild genotypes that present similar phenotypes. Previously, using SSR genotyping of a subset of the Brazilian Panel of Sugarcane Genotypes, TUC71–7 and SP80–3280 were assigned to the same subpopulation, RB72454 and RB855156 to another and, separately, White Transparent and IN84–58 to the two remaining subpopulations [56]. Indeed, the third dimension of the multidimensional scaling based on gene expression showed that SP80-3280 clustered apart from RB72454 (Fig. 1 in Additional file 3). We hypothesize that the lower number of DEGs in the low biomass group reflects sugarcane breeding, because the hybrids in this group have a higher genomic contribution from *S. officinarum*. Hence, they are not only phenotypically more similar to Criolla Rayada and White Transparent than the high biomass accessions, but also share similar gene expression profiles.

The position of accession US85–1008 between the biomass groups also seemingly reflects the sugarcane breeding history, because this hybrid diverged from the high biomass genotypes more than *S. officinarum* (Criolla Rayada and White Transparent) did from commercial hybrids. Furthermore, the high biomass group included US85–1008 and accessions of two ancestral species – *S. spontaneum* and *S. robustum*. Samples of the *S. spontaneum* SES205A were grouped apart, possibly reflecting the diversity within the subpopulations of this species [48]. The wild sugarcane genotypes of the high biomass group showed substantial differences in their expression profiles and we did not find any evidence of kinship among them in the scientific literature. Wild genotypes, particularly those of *S. spontaneum*, have specific alleles that make them a source of variability for sugarcane breeding. Based on SSR markers, IN84–58 showed more species-specific fragments than Badila and Ganda Cheni - *S. officinarum* and *S. barberi* genotypes, respectively [56]. Also, IN84–58 showed a similar expression profile to IJ76–318, a *S. robustum* accession. In fact, Ferreira and colleagues [58] concluded that *S. spontaneum* and *S. robustum* can have similar expression patterns and group together, separately from *S. officinarum* or a hybrid accession.

Transposition-associated terms were enriched among DEGs both for between- and within-group comparisons. Phylogenetically close species have different transposable elements (TEs) families and differ in the number of TEs in the genome [18]. *Saccharum* species have a high number of TEs, mainly Long Terminal Repeat (LTR) retrotransposons [61, 52]. We suggest that the differential expression of TEs was likely due to the genome differences among the genotypes compared in each contrast. *S. officinarum* showed less differential expression of transposition-related genes in comparison to hybrids relative to that found in the comparisons between groups or between US85–1008 and the other high biomass genotypes (Figure 4 in Additional file 3). This may partly be explained by the higher contribution of the *S. officinarum* genome in hybrids and by large differences between the genomes of the wild canes. This is reinforced by the observation that the divergence between *S. officinarum* and *S. spontaneum* is partially due to the expansion of two TE families in *S. officinarum* [47]. TEs may demonstrate restricted expansion in specific genomes, such as certain families of miniature inverted-repeat transposable elements (MITE) with proliferation-specificity to the *T. aestivum* subgenomes [54]. Moreover, the activity of TEs resulting from polyploidization is analogous to the induction of TEs promoted by stresses [18], a form of genomic shock [22, 57], which is a well described phenomenon in allopolyploids [30]. We can conclude that differences in transposition found within the low biomass group were largely due to variation between commercial hybrids and White Transparent, similar to the observation when contrasting *S. officinarum* to the cultivar RB867515 [58].

Polyploidy creates an imbalance in the nucleotide pool, causing genomic stress in the cell and triggering non-additive expression of genotype-specific responsive genes and other stochastic differences [19, 7]. In addition to polyploidy, hybridization is also a potential cause of genetic variation leading to changes in gene expression between hybrids and parental genotypes. In Asteraceae, Qi and colleagues identified hybridization as the main cause for non-additive expression after comparing gene expression levels of parents (*Chrysanthemum nankingense* and *Tanacetum vulgare*), the interspecific hybrid and three derived allopolyploids [1]. Along with transposition, we noted enriched defense-associated terms when comparing both biomass groups (Figs. 3 and 4). There is evidence that proteins involved in basal metabolism can be more active during stresses. For instance, Ferreira et al. [58] hypothesized that upregulation of histone genes in a hybrid genotype arose from changes in epigenetic control caused by the genomic stress of hybridization. Carson and colleagues [33] evaluated gene expression in sugarcane leaves and found, among many functions, genes coding for proteins responsible for the maintenance and control of cellular metabolism, as well as transport and stress responses. Not only does ploidy regulate these responses, but genes coding for resistance proteins were also upregulated in culms to protect against the stress caused by increased sugar levels in sucrose-rich genotypes [2]. Genotypes in the high biomass group differed in their response to oxidation-reduction, presenting changes in genes whose products are associated to detoxification. *Glutathione transferases*, involved in detoxification, display gene classes occurring in tandem on plant genomes, coding for enzymes acting over a wide range of substrates [49]. Previously, higher expression levels of transcripts related to glutathione-S-transferase were observed in a fiber-rich genotype [24].

The co-expression analysis complemented the enrichment tests based on sets of DEGs. Genes associated with transposition formed two clusters of co-expressed genes that showed similarities within the groups (Table 3 and Figure 10 in Additional file 3). The machineries of replication, transcription, translation and regulatory mechanisms were enriched with similarly expressed genes. Our differential expression analysis involved leaf samples, but no carbon assimilation terms were enriched among DEGs. Interestingly, genes whose products are involved with this process were grouped in a co-expressed module (Table 3 in Additional file 3). Depending on the contrast assessed, pathway analysis showed changes in specific photosynthesis processes, such as C4/CAM photosynthesis and photorespiration (Figures 11, 13 and 14 in Additional file 3). Recently, Singh and colleagues [13] detected upregulation of almost all photosynthesis-related coding genes in high biomass genotypes. As a C4 grass, sugarcane photosynthesis includes a pathway to obtain a four-carbon compound, a process that occurs in the mesophyll and is orchestrated by PEPC. In agreement with Verma and colleagues [46], we noted that high biomass genotypes may require a more intense expression of PEPC coding genes to support metabolic functions other than sucrose accumulation. Expression of PEPC genes was lower in young leaves associated with maturing culms but was practically invariable in leaves connected with more mature stalks [46]. In addition, a group of photophosphorylation genes coding for *Psa*, *Psb* and cytochrome proteins formed a downregulated cluster in low biomass genotypes (Figure 12 in Additional file 3). The module with photosynthesis co-expressed genes was also enriched with terms related to the responses to four hormones - abscisic acid, cytokinin, ethylene and gibberellin. DEGs annotated with hormone responses inside this co-expression module were downregulated in *S. spontaneum* (Figure 8 in Additional file 3). In fact, Singh and colleagues [13] noted that low fiber sugarcanes showed upregulation of genes involved with responses to auxin, jasmonic acid, salicylic acid, abscisic acid and ethylene [13].

Genes coding for enzymes involved in sucrose synthesis, breakdown and transport had been previously studied in different phenological stages of sugarcane culm development [3] and between varying (groups of) genotypes [36, 24, 2]. The pioneering transcriptome studies in sugarcane addressed gene expression in leaves or leaf rolls [6, 33]. Analysis of tissue-specific expression enabled the detection of functions in leaves and culms [33]. Synthesis of sucrose occurs in sugarcane leaves, followed by its

transport through phloem to be stored in stalk parenchyma cells [37]. Clearly, sucrose storage is higher in the hybrids and *S. officinarum* clones analyzed herein (Fig. 1 and Additional file 1 - Table 1). In leaves, higher expression of SPS and SPP coding genes in the low biomass group may indicate that the stalk of these genotypes requires more sucrose. They also showed an upregulated gene coding for *Cell Wall Invertase* (CWINV), an enzyme acting on sucrose hydrolysis and allowing the apoplastic entry of hexoses in the stem parenchyma cell [37]. However, CWINV overexpression can promote monomer accumulation in leaves, impairing carbohydrate storage and affecting growth, as described in cassava [55].

SPS and CWINV have been shown to be highly expressed in sugarcane before maturation of culms, precisely to allow the development of leaves and to compensate for sucrose storage requirements in sink tissue [46]. These authors also pointed out that genes coding for enzymes such as PEPC and SUT1 can show stable or increased expression levels in more mature leaves. Our data shows, that in + 1 leaves, genes coding for SUT4 were upregulated in hybrids and *S. officinarum*. However, the SUT1 coding gene was downregulated in the low biomass group but had a higher overall expression level that SUT4 (Fig. 5d), which makes it difficult to determine which SUTs are more relevant to sucrose accumulation. A gene coding for the SWEET14 protein was described as repressed in *S. officinarum* and *S. spontaneum* [58], but we found a SWEET14 gene repressed in the low biomass group, with no evidence of differential expression within this group. We believe that genes coding sugar transporter proteins or sucrose transporter families may be differentially expressed in a genotype-specific manner (Figure 22-B in Additional file 3).

Carbohydrate metabolism in culms also includes gene products from members of the SuSy family. When differentially expressed in a given contrast, SuSy coding genes were always upregulated in genotypes with the higher sucrose level (Fig. 5b). One DEG was also detected in the two other contrasts; other two DEG coding SuSy were upregulated in US85–1008 (Figure 21-B in Additional file 3). In contrast to its common role in stems, SuSy can synthesize sucrose from the reducing sugars present in leaves. Hoffmann-Thoma and colleagues [28] found a higher SuSy activity than SPS in 60 and 90-day expanded leaves. In the same experiment, they found that the content of hexoses was higher than sucrose and that SPS was more active than SuSy in older leaves (2 through 7). In leaf rolls, a low sucrose breakdown/synthesis ratio indicates that SuSy contributes to sucrose synthesis in young sugarcane tissues [50]. Immature leaf rolls, internodes one to six and roots showed higher expression of SuSy1 than leaves [27]. The same study, however, revealed a highly expressed SuSy2 gene in immature and mature leaf lamina. The five DEGs coding for SuSy identified with MERCATOR showed low average expression levels in our study (Table 4 in Additional file 3), three of them being upregulated in low biomass genotypes. Thirugnanasambandam and colleagues [45] noted that the expression levels of four SuSy genes in leaves were lower than in other tissues, regardless of genotype. Although SuSy is possibly synthesizing sucrose, we also stress the importance of SPS for sucrose synthesis in the low biomass group (Figure 21-A in Additional file 3).

Genes coding for proteins of the lignocellulose pathways were upregulated in high biomass genotypes. *Expansins* are a class of enzymes that can modify the structure of the cell wall, promoting its expansion [20]. The sugarcane genome has roughly ninety *expansin*-coding genes, mostly from the families α and β [14]. In Poaceae, β-expansin members act over the matrix polysaccharides, loosening the cell wall [20]. In our study, the high biomass group showed higher expression of *expansin* genes, possibly promoting the development of the leaf. Because structures of the sugarcane top are relevant as biomass sources for energy cane, leaf growth is a desirable trait. Moreover, wild high biomass canes displayed higher expression of expansins α - 2, β - 11 and β - 3, which can be explored as candidate genes in other functional genomic studies. More directly related to the cell wall, many genes coding enzymes that assemble polysaccharides were upregulated in the high biomass genotypes. We identified genes coding for *xylosyltransferases*, *arabinosyltransferases* and *fucosyltransferases* (Fig. 5c and Additional file 3 – Figure

10 and Figure 19), which are glucosyltransferases involved in the biosynthesis of xyloglucan in the Golgi stacks [51]. Loss of function in a *xylosyltransferase* coding gene led to higher saccharification in mutant rice plants, facilitating xylan extraction [16].

Sugarcane genotypes rich in biomass have a higher content of cellulose, hemicellulose and lignin, in detriment to the sucrose content [25]. Clustering of sugarcane genotypes based on similar biomass and sucrose accumulation traits (Figure 1 in Additional file 1) was confirmed by gene expression (Fig. 2). The high biomass group contained mainly wild genotypes, while the low biomass group was represented by *S. officinarum* and hybrids. The high biomass hybrid US85–1008 is the offspring of a wild female parent - an unknown *S. spontaneum* -, while the low biomass hybrids have other hybrids as female parents [56, 8, 63]. Moreover, the low biomass hybrids we studied are all genetically related, with varying degrees of relatedness. This distinct variability within each of the two groups reflects the genomic differences of the accessions (Figure 1 in Additional file 3). Leveraging wild genotypes in sugarcane breeding can be useful to expand the narrow genetic basis of this crop [8, 35], making it possible to develop cultivars with adequate biomass-associated traits, addressing the current limitations in the field and industry. There are also obstacles in sucrose accumulation, which also have to be taken into account because energy canes must be efficient both in biomass and sugar yields [42].

2.4 Conclusion

This work presented a broad view of the expression of many coding genes in sugarcane leaves of different genotypes. With regard to cell wall, most genes were upregulated in the high biomass group, but in general with low average expression levels. On the other hand, highly expressed genes involved in sucrose synthesis were upregulated in hybrids and *S. officinarum* genotypes. These results agree with current knowledge about the partitioning of carbohydrate to sucrose storage and maintenance of plant structure and metabolism in wild genotypes and modern cultivars. In addition, our research shows that investigating expression profiles in wild genotypes can enhance the understanding of genes selected through domestication and breeding. Expression profiles in other plant parts of wild and cultivated accessions are needed to provide knowledge about the action of the genes involved in carbohydrate metabolism and biomass production. Our data from sugarcane leaves revealed how hybridization in a complex polyploid system resulted in noticeably different transcriptomic profiles between contrasting genotypes.

2.5 Methods

2.5.1 Plant material

We collected leaves of genotypes from the Brazilian Panel of Sugarcane Genotypes [56], selected from groups contrasting in key biomass traits, as measured by fiber content and stalk number. This panel is managed by the sugarcane breeding program of the Inter-University Network for the Development of the Sugarcane Sector (RIDESA), at the Federal University of São Carlos (Araras, Brazil). No special permission was necessary to collect biological samples from these plants. Genotypes of the high biomass group were IN84–58, IN84–88, Krakatau, SES205A, IJ76–318 and US85–1008. In the low biomass group, we selected White Transparent, Criolla Rayada, TUC71–7, RB72454, SP80–3280 and RB855156. Their phenotypic means for soluble solids content (°Brix), percentage of apparent sucrose present in juice (POL % Juice), purity, fiber content (FIB%) and stalk number are summarized in Table 1 (Additional file 1). We performed a hierarchical clustering and a principal component analysis using these measures, and identified two main groups that reflect the separation of high and low fiber genotypes (Fig. 1 and Additional file 1 - Figure 1).

In the high biomass group, there were four *S. spontaneum* representatives (IN84–58, IN84–88, Krakatau and SES205A), a *S. robustum* (IJ76–318) and a hybrid (US85–1008). SES205A is a genotype from India, used in studies of hybrids generated by crosses with *S. officinarum* [48, 12]. Krakatau is an Indonesian *S. spontaneum* widely used in works about biological nitrogen fixation [48, 40, 21]. Genotypes IN84–88, IN84–58 and IJ76–318 are also from Indonesia, and US85–1008 is an accession originated by a cross between a *S. spontaneum* genotype and US60–313 [48, 63].

Samples of the low biomass group include four hybrid cultivars - TUC71–7, RB72454, SP80–3280 and RB855156 - and two *S. officinarum* genotypes - White Transparent and Criolla Rayada. White Transparent was used during the nobilization process [8, 31]. TUC71–7 is a cultivar from Tucumán-Argentina [56, 8], and RB72454, SP80–3280 and RB855156 are Brazilian commercial hybrids [56].

Replicates of each biomass group consisted in one leaf from each genotype. Additionally, we sampled clonal replicates by collecting three leaves from six genotypes (IN84–58, SES205A, US85–1008, White Transparent, RB72454 and SP80–3280). This resulted in a total of 24 samples – 12 genotypes, half of them with clonal replicates. By doing so, we aimed to sample biological variation at two levels: i) between biomass groups, replicates were composed of different genotypes; ii) clonal replicates of particular genotypes allowed for comparisons within each group. Our goal was to have clonal replicates of distant genotypes within each group.

Portions of the first visible dewlap leaves (+ 1) were collected from six-month-old sugarcane plants in April 2016, grown in the field in Araras, Brazil ($22°18'41.0''S$, $47°23'05.0''W$, at an altitude of 611 m). We collected the middle section of each leaf, removing the midrib. After cutting, they were placed in plastic tubes (50 mL), immediately frozen in liquid nitrogen and stored at -80 °C until RNA extraction. Figure 1 of Additional file 2 shows a summary of our laboratory and bioinformatics steps.

2.5.2 RNA extraction, sequencing and quality of the libraries

We used the RNeasy Plant Mini Kit (Qiagen, cat. no. *74904*) with roughly 50 mg of starting leaves to extract total RNA from each sample. RNA quality was evaluated by observing the 25S and 18S rRNAs bands via 1% agarose gel electrophoresis. We assessed RNA integrity via 2100 Bioanalyzer (Agilent Technologies) capillary electrophoresis and only kept samples with RNA Integrity Number (RIN) greater than 8. Libraries were prepared with the TruSeq Stranded kit and sequenced in an Illumina HiSeq 2500 platform. We pooled the 24 libraries and sequenced this pool in two lanes, in paired-end mode (2 $\times$ 100 bp).

2.5.3 Differential expression and functional enrichment analyses

We quantified expression levels of *de novo* assembled transcripts using Salmon [29] (see Additional file 2 for details about read filtering, *de novo* transcriptome assembly and functional annotation). Isoform expression information was aggregated to gene-count levels using the tximport R package [11]. Next, the data were filtered for genes with expression levels of at least one count per million (cpm) in at least three samples. We performed differential expression analyses with edgeR [4], using two different strategies. First, all samples were used to design a model with two groups contrasting in biomass content. Next, we fitted two separate models to contrast genotypes within each biomass group, including only the genotypes with clonal replicates of each group in an ANOVA-like test. Two contrasts were performed to obtain a Fold Change value within the groups, comparing US85–1008 with the mean of IN84–58 and SES205A, and White Transparent to the mean of SP80–3280 and RB72454. For each model, the DEGs were those with an FDR-adjusted p-value less than 5% [41].

Functional enrichment analyses were performed with the goseq R package [9], separately for each differential expression model. The background set was composed of the expressed genes passing the

cpm filter. A GO term was considered enriched among DEGs if its overrepresentation adjusted p-value was less than 5%.

Additionally, we carried out tests at the transcript level to find differentially expressed transcripts between the same biomass groups. We then compared the two approaches by measuring the overlap between the lists of DETs and DEGs.

2.5.4 Co-expression network and gene set enrichment analysis

A co-expression network was built with WGCNA [43], using as input the logarithm of the normalized cpm matrix of the expressed genes. We chose a soft-thresholding power of nine, reaching a correlation coefficient of approximately 0.8 for the scale-free topology fit. Our choice was to build an adjacency matrix preserving the sign of the connection. After hierarchical clustering of genes based on their dissimilarity, modules that were composed of at least 300 genes were considered. We grouped modules that had highly co-expressed genes, using a correlation threshold of 0.75 for the module eigengenes. The sets of genes defined by each module, were used to evaluate the presence of enriched Gene Ontology terms with GOSEQ , again considering an overrepresented adjusted p-value less than 5%.

Next, we checked the enrichment of the gene set formed by each co-expression module by ranking genes based on their absolute LFC for each contrast. This analysis was conducted with the GSEAPRERANKED tool in the GSEA software [23].

2.5.5 Pathway analysis

The MAPMAN4 pipeline [53] was used to functionally assign genes to land plant protein categories. The full transcriptome was annotated using the MERCATOR4 tool. Because the expression quantification was done at the gene level, the transcript identifiers of the MERCATOR4 mapping file were changed to gene identifiers. Thus, the functional annotation attributed to isoforms of a gene were also combined. Genes in the MAPMAN4 pathways were tagged and colored based on the LFC from the differential expression tests.

References

[1] Qi X, Wang H, Song A, Jiang J, Chen S, Chen F. Genomic and transcriptomic alterations following intergeneric hybridization and polyploidization in the Chrysanthemum nankingense×Tanacetum vulgare hybrid and allopolyploid (Asteraceae). Horticulture Research. 2018 dec;5(1):5. Available from: http://dx.doi.org/10.1038/s41438-017-0003-0http://www.nature.com/articles/s41438-017-0003-0.

[2] Thirugnanasambandam PP, Hoang NV, Furtado A, Botha FC, Henry RJ. Association of variation in the sugarcane transcriptome with sugar content. BMC Genomics. 2017;18(1):1–22.

[3] Casu RE, Jarmey JM, Bonnett GD, Manners JM. 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. 2007 feb;7(2):153–167. Available from: http://link.springer.com/10.1007/s10142-006-0038-z.

[4] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 jan;26(1):139–140. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp616.

[5] Grivet L, Daniels C, Glaszmann JCC, Hont aD, D'Hont A. A review of recent molecular genetics evidence for sugarcane evolution and domestication. Ethnobotany Research & Applications. 2004;2(0):9–17.

[6] Carson DL, Botha FC. Preliminary Analysis of Expressed Sequence Tags for Sugarcane. Crop Science. 2000;40(6):1769. Available from: https://www.crops.org/publications/cs/abstracts/40/6/1769.

[7] Jackson S, Chen ZJ. Genomic and expression plasticity of polyploidy. Current Opinion in Plant Biology. 2010 apr;13(2):153–159. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369526609001757.

[8] Acevedo A, Tejedor MT, Erazzú LE, Cabada S, Sopena R. Pedigree comparison highlights genetic similarities and potential industrial values of sugarcane cultivars. Euphytica. 2017;213(6).

[9] Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology. 2010;11(2):R14. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-2-r14.

[10] Cardoso-Silva CB, Costa EA, Mancini MC, Balsalobre TWA, Costa Canesin LE, Pinto LR, et al. De novo assembly and transcriptome analysis of contrasting sugarcane varieties. PLoS ONE. 2014;9(2).

[11] Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2016 feb;4:1521. Available from: https://f1000research.com/articles/4-1521/v2.

[12] Pan YB, Burner DM, Legendre BL, Grisham MP, White WH. An assessment of the genetic diversity within a collection of Saccharum spontaneum L. with RAPD-PCR. Genetic Resources and Crop Evolution. 2005;51(8):895–903.

[13] Singh R, Jones T, Wai CM, Jifon J, Nagai C, Ming R, et al. Transcriptomic analysis of transgressive segregants revealed the central role of photosynthetic capacity and efficiency in biomass accumulation in sugarcane. Scientific Reports. 2018;8(1):1–10. Available from: http://dx.doi.org/10.1038/s41598-018-22798-5.

[14] Santiago TR, Pereira VM, de Souza WR, Steindorff AS, Cunha BADB, Gaspar M, et al. Genome-wide identification, characterization and expression profile analysis of expansins gene family in sugarcane (Saccharum spp.). PLoS ONE. 2018;13(1):1–18.

[15] Leal MRLV, Galdos MV, Scarpare FV, Seabra JEA, Walter A, Oliveira COF. Sugarcane straw availability, quality, recovery and energy use: A literature review. Biomass and Bioenergy. 2013;53:11–19.

[16] Chiniquy D, Sharma V, Schultink A, Baidoo EE, Rautengarten C, Cheng K, et al. XAX1 from glycosyltransferase family 61 mediates xylosyltransfer to rice xylan. Proceedings of the National Academy of Sciences. 2012 oct;109(42):17117–17122. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1202079109.

[17] Swapna M, Sivaraju K, Sharma RK, Singh NK, Mohapatra T. Single-Strand Conformational Polymorphism of EST-SSRs: A Potential Tool for Diversity Analysis and Varietal Identification in Sugarcane. Plant Molecular Biology Reporter. 2011 sep;29(3):505–513. Available from: http://link.springer.com/10.1007/s11105-010-0254-5.

[18] Vicient CM, Casacuberta JM. Impact of transposable elements on polyploid plant genomes. Annals of Botany. 2017;120(2):195–207.

[19] Fasano C, Diretto G, Aversano R, D'Agostino N, Di Matteo A, Frusciante L, et al. Transcriptome and metabolome of synthetic Solanum autotetraploids reveal key genomic stress events following polyploidization. New Phytologist. 2016 jun;210(4):1382–1394. Available from: http://doi.wiley.com/10.1111/nph.13878.

[20] Sampedro J, Guttman M, Li LC, Cosgrove DJ. Evolutionary divergence of β-expansin structure and function in grasses parallels emergence of distinctive primary cell wall traits. Plant Journal. 2015;81(1):108–120.

[21] Urquiaga S, Xavier RP, de Morais RF, Batista RB, Schultz N, Leite JM, et al. Evidence from field nitrogen balance and 15N natural abundance data for the contribution of biological N2 fixation to Brazilian sugarcane varieties. Plant and Soil. 2012 jul;356(1-2):5–21. Available from: http://link.springer.com/10.1007/s11104-011-1016-3.

[22] Fedoroff NV, Bennetzen JL. Transposons, Genomic Shock, and Genome Evolution. In: Plant Transposons and Genome Dynamics in Evolution. Oxford, UK: Wiley-Blackwell; 2013. p. 181–201. Available from: http://doi.wiley.com/10.1002/9781118500156.ch10.

[23] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545–15550. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16199517{\%}0Ahttp://www.pnas.org/cgi/doi/10.1073/pnas.0506580102.

[24] Vicentini R, Bottcher A, Dos Santos Brito M, Dos Santos AB, Creste S, De Andrade Landell MG, et al. Large-scale transcriptome analysis of two sugarcane genotypes contrasting for lignin content. PLoS ONE. 2015 aug;10(8):e0134909. Available from: http://dx.plos.org/10.1371/journal.pone.0134909.

[25] Hoang NV, Furtado A, Donnan L, Keeffe EC, Botha FC, Henry RJ. High-Throughput Profiling of the Fiber and Sugar Composition of Sugarcane Biomass. Bioenergy Research. 2017;10(2):400–416.

[26] Rubin EM. Genomics of cellulosic biofuels. Nature. 2008;454(7206):841–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18704079.

[27] Lingle SE, Dyer JM. Cloning and expression of sucrose synthase-1 cDNA from sugarcane. Journal of Plant Physiology. 2001;158(1):129–131.

[28] Hoffmann-Thoma G, Hinkel K, Nicolay P, Willenbrink J. Sucrose accumulation in sweet sorghum stem internodes in relation to growth. Physiologia Plantarum. 1996;97(2):277–284.

[29] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nature Methods. 2017;14(4):417–419.

[30] Chen ZJ. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annual Review of Plant Biology. 2007;58(1):377–406.

[31] Creste S, Xavier MA, Landell MGA. Importância do germoplasma no desenvolvimento de cultivares de cana-de-açúcar com perfil agroenergético. In: Cortez LAB, editor. Bioetanol de Cana-de-Açúcar: P&D para produtividade e sustentabilidade; 2010. p. 313–317.

[32] Zhang J, Nagai C, Yu Q, Pan YB, Ayala-Silva T, Schnell RJ, et al. Genome size variation in three Saccharum species. Euphytica. 2012;185(3):511–519.

[33] Carson DL, Huckett BI, Botha FC. Differential gene expression in sugarcane leaf and internodal tissues of varying maturity. South African Journal of Botany. 2002 dec;68(4):434–442. Available from: c:{\%}5CDocumentsandSettings{\%}5Ccas128{\%}5CMyDocuments{\%}5CDownloadedpapers{\%}5CSAJB-2002-68-434-442.pdf{\%}5Cnhttp://linkinghub.elsevier.com/retrieve/pii/S0254629915303707https://linkinghub.elsevier.com/retrieve/pii/S0254629915303707.

[34] Grivet L, Arruda P. Sugarcane genomics: Depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology. 2002;5(2):122–127.

[35] Thirugnanasambandam PP, Hoang NV, Henry RJ. The Challenge of Analyzing the Sugarcane Genome. Frontiers in Plant Science. 2018;9(May):1–18. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2018.00616/full.

[36] Kasirajan L, Hoang NV, Furtado A, Botha FC, Henry RJ. Transcriptome analysis highlights key differentially expressed genes involved in cellulose and lignin biosynthesis of sugarcane genotypes varying in fiber content. Scientific Reports. 2018;8(1):1–16.

[37] Wang J, Nayak S, Koch K, Ming R. Carbon partitioning in sugarcane (Saccharum species). Frontiers in Plant Science. 2013;4(June):2005–2010. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2013.00201/abstract.

[38] Irvine JE. Saccharum species as horticultural classes. Theoretical and Applied Genetics. 1999 feb;98(2):186–194. Available from: http://dx.doi.org/10.1007/s001220051057http://link.springer.com/10.1007/s001220051057.

[39] Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN. Plants to power: bioenergy to fuel the future. Trends in Plant Science. 2008;13(8):421–429.

[40] Burbano CS, Liu Y, Rösner KL, Reis VM, Caballero-Mellado J, Reinhold-Hurek B, et al. Predominant nifH transcript phylotypes related to Rhizobium rosettiformans in field-grown sugarcane plants and in Norway spruce. Environmental Microbiology Reports. 2011;3(3):383–389.

[41] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995;57(1):289–300. Available from: http://www.jstor.org/stable/2346101{\%}5Cnhttp://about.jstor.org/terms.

[42] Alexander AG. The energy cane alternative. Amsterdam, Netherlands: Elsevier Science Publishers B.V.; 1985.

[43] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9:559. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19114008{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2631488.

[44] Osborn TC, Chris Pires J, Birchler JA, Auger DL, Chen ZJ, Lee HS, et al. Understanding mechanisms of novel gene expression in polyploids. Trends in Genetics. 2003;19(3):141–147.

[45] Thirugnanasambandam PP, Mason PJ, Hoang NV, Furtado A, Botha FC, Henry RJ. Analysis of the diversity and tissue specificity of sucrose synthase genes in the long read transcriptome of sugarcane. BMC Plant Biology. 2019;19(1):160. Available from: https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-019-1733-y.

[46] Verma I, Roopendra K, Sharma A, Chandra A, Kamal A. Expression analysis of genes associated with sucrose accumulation and its effect on source–sink relationship in high sucrose accumulating early maturing sugarcane variety. Physiology and Molecular Biology of Plants. 2019;25(1):207–220. Available from: https://doi.org/10.1007/s12298-018-0627-z.

[47] Garsmeur O, Droc G, Antonise R, Grimwood J, Potier B, Aitken K, et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nature Communications. 2018;9(1). Available from: http://dx.doi.org/10.1038/s41467-018-05051-5.

[48] Aitken K, Li J, Piperidis G, Qing C, Yuanhong F, Jackson P. Worldwide Genetic Diversity of the Wild Species and Level of Diversity Captured within Sugarcane Breeding Programs. Crop Science. 2018;58(1):218. Available from: https://dl.sciencesocieties.org/publications/cs/abstracts/58/1/218.

[49] Labrou NE, Papageorgiou AC, Pavli O, Flemetakis E. Plant GSTome: structure and functional role in xenome network and plant stress response. Current Opinion in Biotechnology. 2015 apr;32:186–194. Available from: http://dx.doi.org/10.1016/j.copbio.2014.12.024https://linkinghub.elsevier.com/retrieve/pii/S0958166914002390.

[50] Schäfer WE, Rohwer JM, Botha FC. Partial purification and characterisation of sucrose synthase in sugarcane. Journal of Plant Physiology. 2005;162(1):11–20.

[51] Driouich A, Follet-Gueye ML, Bernard S, Kousar S, Chevalier L, Vicré-Gibouin M, et al. Golgi-Mediated Synthesis and Secretion of Matrix Polysaccharides of the Primary Cell Wall of Higher Plants. Frontiers in Plant Science. 2012;3. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2012.00079/abstract.

[52] Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nature Genetics. 2018;50(11):1565–1573.

[53] Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, Hallab A, et al. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Molecular Plant. 2019 jun;12(6):879–892. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1674205219300085.

[54] Keidar-Friedman D, Bariah I, Kashkush K. Genome-wide analyses of miniature inverted-repeat transposable elements reveals new insights into the evolution of the triticum-Aegilops group. PLoS ONE. 2018;13(10):1–23.

[55] Yan W, Wu X, Li Y, Liu G, Cui Z, Jiang T, et al. Cell Wall Invertase 3 Affects Cassava Productivity via Regulating Sugar Allocation From Source to Sink. Frontiers in Plant Science. 2019;10(April):1–16.

[56] Barreto FZ, Rosa JRBF, Balsalobre TWA, Pastina MM, Silva RR, Hoffmann HP, et al. A genome-wide association study identified loci for yield component traits in sugarcane (Saccharum spp.). PLOS ONE. 2019 jul;14(7):e0219843. Available from: http://dx.plos.org/10.1371/journal.pone.0219843.

[57] McClintock B. The significance of responses of the genome to challenge. Science. 1984;226(4676):792–801.

[58] Ferreira SS, Hotta CT, Poelking VGdC, Leite DCC, Buckeridge MS, Loureiro ME, et al. Co-expression network analysis reveals transcription factors associated to cell wall biosynthesis in sugarcane. Plant Molecular Biology. 2016 may;91(1-2):15–35. Available from: http://link.springer.com/10.1007/s11103-016-0434-2.

[59] Diniz AL, Ferreira SS, Ten-Caten F, Margarido GRA, dos Santos JM, Barbosa GVdS, et al. Genomic resources for energy cane breeding in the post genomics era. Computational and Structural Biotechnology Journal. 2019;17:1404–1414. Available from: https://doi.org/10.1016/j.csbj.2019.10.006.

[60] Piperidis N, D'Hont A. Sugarcane genome architecture decrypted with chromosome-specific oligo probes. The Plant Journal. 2020 jul:tpj.14881. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.14881.

[61] de Setta N, Monteiro-Vitorello C, Metcalfe C, Cruz GM, Del Bem L, Vicentini R, et al. Building the sugarcane genome for biotechnology and identifying evolutionary trends. BMC Genomics. 2014;15(1):540. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-540.

[62] Matsuoka S, Ferro J, Arruda P. The Brazilian experience of sugarcane ethanol industry. In Vitro Cellular & Developmental Biology - Plant. 2009 jun;45(3):372–381. Available from: http://link.springer.com/10.1007/978-1-4419-7145-6{_}9http://link.springer.com/10.1007/s11627-009-9220-z.

[63] da Silveira LCI, Brasileiro BP, Kist V, Daros E, Peternelli LA. Genetic diversity and coefficient of kinship among potential genitors for obtaining cultivars of energy cane. Revista Ciencia Agronomica. 2015;46(2):358–368.

Additional file 1 - Phenotypic characterization

We obtained measures of °Brix, POL % Juice, purity, fiber and stalk number of four replicates per genotype during two harvest years - 2015 and 2016. In 2015 the phenotypic data was collected from plant cane, while ratoon cane was used for the phenotypic measures in 2016. The mean of the measures for each genotype are shown in Table 1. Using the dissimilarity of the genotypes based on euclidean distances of the phenotypic measures, we performed a hierarchical clustering to obtain a dendrogram. Samples clustered in two main groups, reflecting the separation of high and low fiber genotypes (Figure 1). We observed the same clustering pattern through a principal component analysis (Figure 1), where the first dimension explained more than 90% of the variation. These groups differed in their biomass content, in terms of fiber and stalk number, and by their sucrose accumulation in culms, which are related with POL % Juice, °Brix and purity.

Table 1: Phenotypic measures of the genotypes. °Brix corresponds to the content of soluble solids in the cane juice. POL % Juice is the polarization measurement of the sucrose percentage in the juice. Purity indicates the percentage of sucrose in the total solids of the juice. Fiber is the percentage of fiber in the bagasse and stalk number represents the number of stalks in each plot.

Genotype	°Brix	POL % Juice	Purity	Fiber	Stalk number
Criolla Rayada	16.53 ± 1.37	12.79 ± 1.82	77.05 ± 4.76	9.83 ± 1.21	22.71 ± 14.44
White Transparent	19.6 ± 2.02	17.15 ± 2.54	87.12 ± 4.31	10.82 ± 1.54	115.88 ± 24.64
RB72454	20.52 ± 1.66	18.03 ± 2.38	88.69 ± 3.15	11.14 ± 2.08	117.62 ± 13.03
RB855156	21.01 ± 1.19	19.15 ± 1.51	91.07 ± 3.17	12 ± 1.07	115.38 ± 20.85
SP80-3280	21.29 ± 1.88	19.1 ± 2.3	89.5 ± 3.02	12.14 ± 0.68	78.12 ± 23.17
TUC71-7	22.76 ± 0.8	20.77 ± 0.77	91.25 ± 0.39	12.97 ± 1.27	47.88 ± 14.96
US85-1008	17.61 ± 1.11	13.55 ± 1.54	76.81 ± 6.17	18.96 ± 1.82	298.71 ± 72.87
SES205A	13.99 ± 3.14	8.77 ± 3.78	59.89 ± 16.62	21.86 ± 3.6	464.62 ± 91.42
Krakatau	12.03 ± 2.02	6.53 ± 2.09	53.08 ± 9.52	22.22 ± 3.5	244.38 ± 61.53
IN84-88	15.91 ± 2.43	10.64 ± 4.24	66.03 ± 22.63	22.35 ± 2.26	353.86 ± 62.8
IN84-58	14.78 ± 1.89	7.88 ± 1.5	55.28 ± 4.86	24.64 ± 1.7	316.88 ± 105.6
IJ76-318	14.54 ± 3.24	8.78 ± 3.49	58.52 ± 11.12	25.55 ± 4.03	334.14 ± 117.65

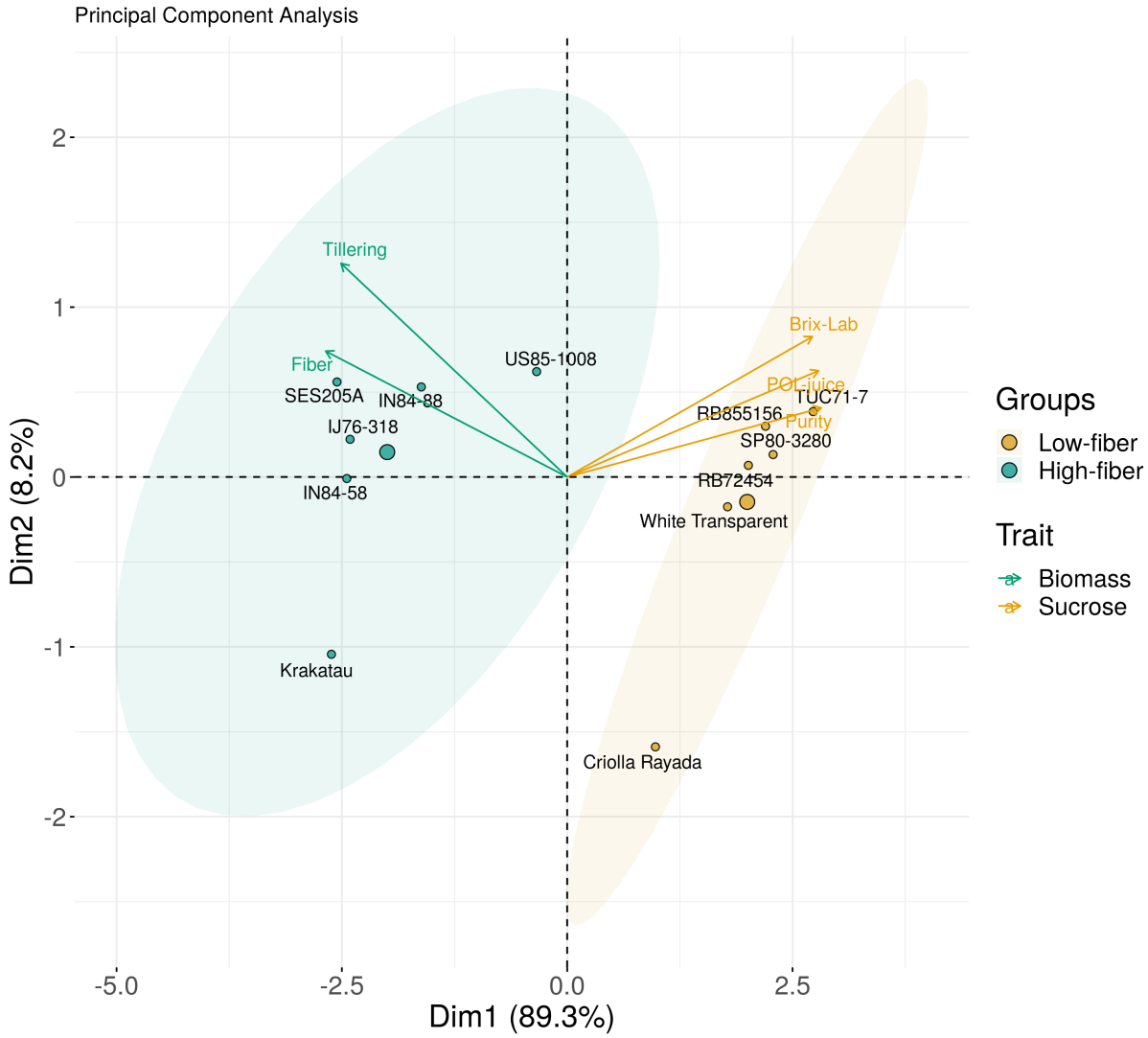

Figure 1: Principal component analysis of biomass and sucrose traits. Yellow projections reflect traits related to higher biomass, while green projetions indicate traits associated to sucrose accumulation. The two biomass groups (red and blue) were found based on the k-means algorithm.

Additional file 2 - Supporting information for methods

Methods summary

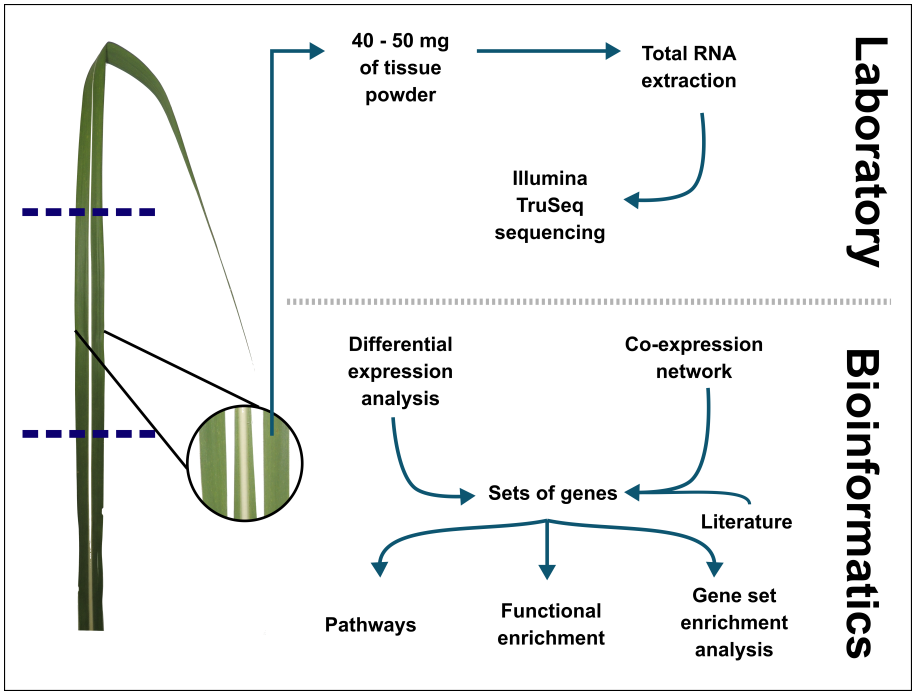

Figure 1: =Summary of the main methodological steps. Workflow is divided in laboratory and bioinformatic steps. The laboratory section includes steps from the collection of leaves to the Illumina TruSeq sequencing. The bioinformatics section indicates the analyses to find gene sets, differential expression and co-expression networks, including the search for genes in literature. From the sets of genes, we searched for pathway mappings in MapMan and functional enrichment of Gene Ontology terms.

Preprocessing of raw reads and *de novo* transcriptome assembly

To evaluate the quality of sequencing runs, we used the diagnosis tool FASTQC [7]. Removal of adapters and low quality bases was performed with Trimmomatic [1], using windows with a minimum average Phred quality score of 20. We also trimmed the first 12 bases and kept reads with at least 75 bases.

We performed transcriptome *de novo* assemblies with Trinity (v.2.8.4) [2], using as parameters the k-mer size of 25, normalization of FASTQ pairs (*normalize_by_read_set*) and minimum contig length (*min_contig_length*) of 300. In addition to these these parameters, in the second assembly we set k-mer coverage (*min_kmer_cov*) to two. In the third assembly we set the maximum number of reads to combine into a single path (*max_reads_per_graph*), minimum percent identity (*min_per_id_same_path*) and maximum differences between two paths (*max_diffs_same_path*) to 3,000,000, 90 and 10, respectively. This means that we increased the number of reads anchored within a graph, reduced the identity for the paths be combined into a single one and allowed more differences to combine two paths. A fourth *de novo* transcriptome was built combining parameters of the two previous assemblies.

Assembly statistics, such as the number of unigenes and number of transcripts, are in Table 1. The completeness of the *de novo* assemblies was evaluated with BUSCO [6] using the set of longest isoforms of the assembly and datasets of conserved orthologs from *Viridiplantae* and *Liliopsida*. To assess RNA-Seq read representation, we mapped the preprocessed reads to each transcriptome using HISAT [3]. This mapping was used only as a metric to assess the assembly with the best read representation.

Table 1: *De novo* transcriptome assembly statistics. Assembly 1 was *de novo* assembled with the common set of parameters. A k-mer coverage of at least two was used for the Assembly 2. Assembly 3 had as parameters regarding a same path: a maximum of 3,000,000 reads to be combined, a minimum identity of 90% and up to 10 differences. All the parameters of the assemblies 2 and 3 were combined to generate Assembly 4.

	Assembly 1	**Assembly 2**	**Assembly 3**	**Assembly 4**
Total trinity genes	174,755	111,670	166,517	106,010
Total trinity transcripts	437,123	331,229	373,896	279,896
Percent GC	48.94	49.29	48.63	49.03
'Genes' N50	1,734	1,779	1,926	1,902
Longest isoform N50	1,123	1,325	1,192	1,396

Mapping of reads to the longest isoform was higher in both the first and second assemblies (Table 2). The representation of complete conserved orthologs (Table 3) was higher in the first assembly, particularly for the full *Viridiplantae* gene set.

Table 2: Number of input reads and overall alignment rate. Assembly 1 was *de novo* assembled with the common set of parameters. A k-mer coverage of at least two was used for the Assembly 2. Assembly 3 had as parameters regarding a same path: a maximum of 3,000,000 reads to be combined, a minimum identity of 90% and up to 10 differences. All the parameters of the assemblies 2 and 3 were combined to generate Assembly 4.

Sample	Input fragments	Assembly 1	Assembly 2	Assembly 3	Assembly 4
Criolla Rayada	17,449,229	74.80	74.43	73.08	72.72
IJ76-318	20,673,607	74.27	73.84	72.75	72.44
IN84-58	26,880,400	73.53	73.49	72.05	71.58
IN84-58	17,388,508	73.50	73.34	71.94	71.75
IN84-58	19,386,319	74.41	74.02	72.62	72.35
IN84-88	16,343,061	73.47	72.86	72.02	71.68
Krakatau	18,943,919	73.52	73.14	72.41	71.98
RB72454	15,828,991	73.77	73.31	72.31	72.31
RB72454	16,474,182	74.00	73.62	72.69	72.44
RB72454	20,096,595	74.59	74.25	73.30	73.03
RB855156	19,062,736	74.66	74.19	73.43	73.03
SES205A	17,771,779	74.55	74.66	73.47	73.22
SES205A	19,574,309	73.78	73.57	72.45	72.21
SES205A	19,234,085	73.14	72.91	71.77	71.56
SP80-3280	15,332,802	74.40	74.11	72.75	72.57
SP80-3280	16,877,418	74.16	73.78	72.44	72.51
SP80-3280	22,863,504	75.06	74.66	73.48	73.18
TUC71-7	18,759,428	75.32	75.07	73.96	73.61
US85-1008	22,047,957	73.88	73.78	72.74	72.31
US85-1008	21,531,366	71.69	71.56	70.70	70.74
US85-1008	16,146,634	74.94	74.94	74.16	73.82
White Transparent	16,157,056	74.13	73.88	72.14	72.43
White Transparent	18,175,403	74.20	73.64	71.91	72.48
White Transparent	17,786,061	75.07	74.75	73.60	73.30

Because the first and second assemblies showed the best results for these two criteria, we evaluated their DETONATE RSEM-EVAL Score. This model-based score is based on support of RNA-Seq reads and other factors, such as assembly compactness [8]. The first assembly score (-4.42×10^9) was higher than that of the second assembly (-16.91×10^9).

Finally, we examined the full-length transcript counting using the script *analyze_blastPlus_topHit_coverage.pl* comparing our two assemblies with UniProt. After aligning the

Table 3: Percentage of conserved orthologs from *Viridiplantae* and *Liliopsida* present in the longest isoform assemblies. Assembly 1 was *de novo* assembled with the common set of parameters. A k-mer coverage of at least two was used for the Assembly 2. Assembly 3 had as parameters regarding a same path: a maximum of 3,000,000 reads to be combined, a minimum identity of 90% and up to 10 differences. All the parameters of the assemblies 2 and 3 were combined to generate Assembly 4.

	Assembly 1	**Assembly 2**	**Assembly 3**	**Assembly 4**
Viridiplantae				
Complete and single-copy BUSCOs (S)	74.4	69.1	63.3	68.8
Complete and duplicated BUSCOs (D)	1.2	1.4	1.9	0.9
Fragmented BUSCOs (F)	17.4	21.9	24.9	22.8
Missing BUSCOs	7.0	7.6	9.9	7.5
Liliopsida				
Complete and single-copy BUSCOs (S)	68.4	67.7	62.0	65.2
Complete and duplicated BUSCOs (D)	2.2	2.0	2.0	1.6
Fragmented BUSCOs (F)	17.4	17.5	21.0	18.8
Missing BUSCOs	12.0	12.8	15.0	14.4

transcripts of each assembly with UniProt proteins by Blastx-, we grouped Blast hits using the script *blast_outfmt6_group_segments.tophit_coverage.pl*. For all protein coverage thresholds, the number of proteins was higher in the first assembly (Figure 2). This analysis also indicates that the first assembly was more appropriate for the subsequent steps of the analysis.

Using the complete transcriptome obtained with the first assembly, 97.4% of conserved eukaryotic orthologs were found as complete (Table 4). The assembled transcriptome proves to be a suitable sugarcane reference, representing the eukaryotic orthologs as well as other sugarcane transcriptomes used as references [4].

	Complete Transcriptome	**Longest Isoforms**
Complete and single-copy BUSCOs (S)	17.2	69.0
Complete and duplicated BUSCOs (D)	80.2	22.1
Fragmented BUSCOs (F)	2.0	7.9
Missing BUSCOs	0.6	1.0

Table 4: Percentage of conserved orthologs from *Eukaryota* present in the complete transcriptome and in the longest isoforms.

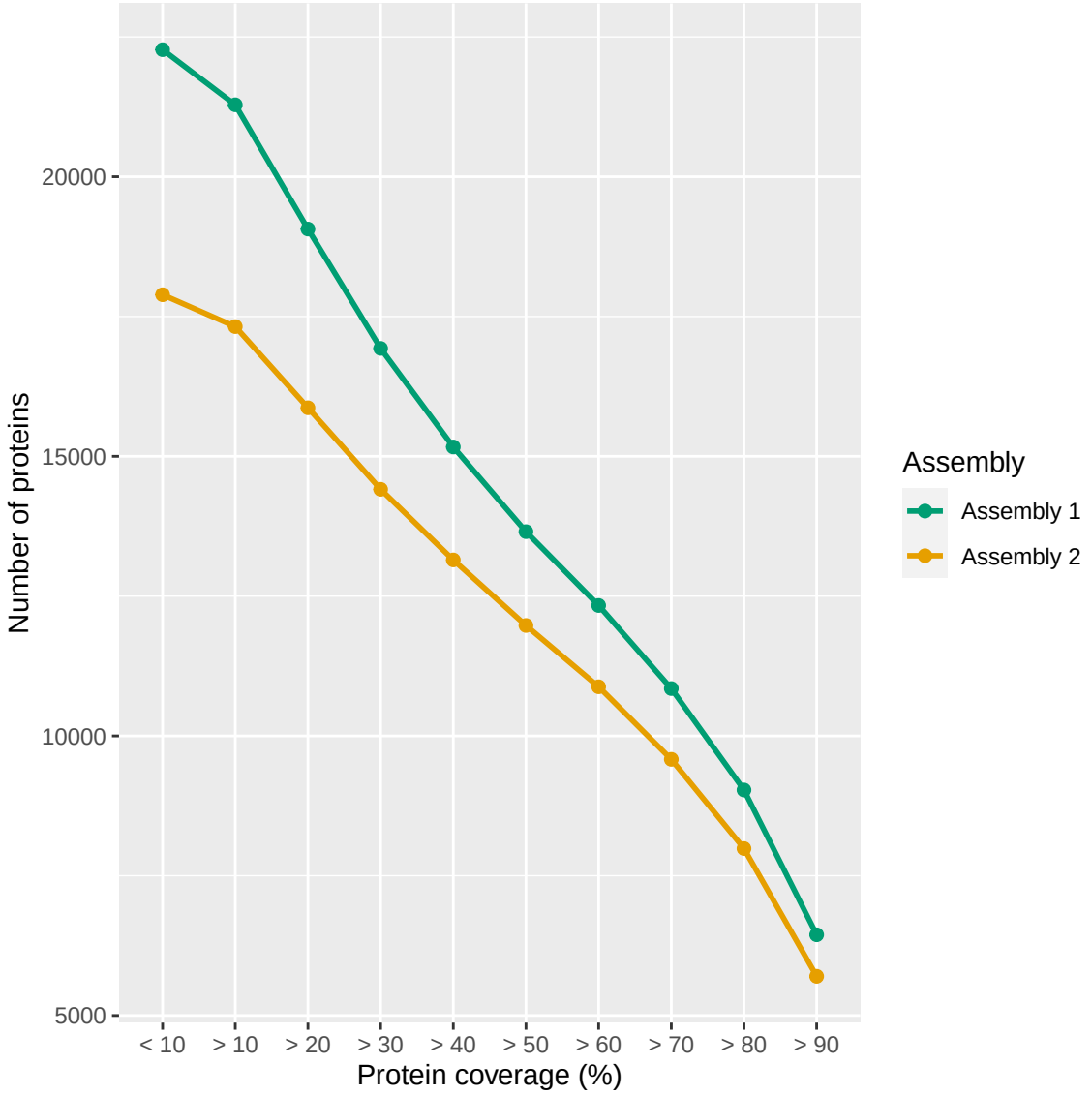

Figure 2: Counts of full-length transcripts for varying thresholds of protein coverage.

Transcriptome annotation

We performed annotation with TRINOTATE [5], using: i) homology search of our sequences to the UniProt database; ii) protein domain identification from Pfam; iii) prediction of protein signal peptides and transmembrane domains. This approach can recover information from the databases of Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and eggNOG.

References

[1] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–2120. Available from: `+http://dx.doi.org/10.1093/bioinformatics/btu170`.

[2] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011 jul;29(7):644–652. Available from: http://dx.doi.org/10.1038/nbt.1883http://www.nature.com/nbt/journal/v29/n7/abs/nbt.1883.html{\#}supplementary-informationhttp://www.nature.com/articles/nbt.1883.

[3] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015 mar;12(4):357–360. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3317http://www.ncbi.nlm.nih.gov/pubmed/25751142{\%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4655817.

[4] Hoang NV, Furtado A, Perlo V, Botha FC, Henry RJ. The Impact of cDNA Normalization on Long-Read Sequencing of a Complex Transcriptome. Frontiers in Genetics. 2019 jul;10. Available from: https://www.frontiersin.org/article/10.3389/fgene.2019.00654/full.

[5] Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Reports. 2017 jan;18(3):762–776. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2211124716317703.

[6] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 oct;31(19):3210–3212. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv351.

[7] Andrews S. FastQC: a quality control tool for high throughput sequence data. Available in: http://wwwbioinformaticsbabrahamacuk/projects/fastqc. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[8] Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology. 2014;15(12):1–21.

Additional file 3 - Supporting information for results

Mapping and quantification

Mapping rates of preprocessed reads obtained with SALMON [1] are shown in Table 1.

Table 1: SALMON mapping rates of the preprocessed reads. Table contains the percentage of mapping rate for samples in both high and low biomass groups.

Sample	Group	Mapping rate (%)
IJ76-318	High biomass	83.91
IN84-58	High biomass	82.27
IN84-58	High biomass	83.53
IN84-58	High biomass	82.76
IN84-88	High biomass	83.59
Krakatau	High biomass	82.03
SES205A	High biomass	83.00
SES205A	High biomass	84.71
SES205A	High biomass	80.52
US85-1008	High biomass	83.56
US85-1008	High biomass	80.84
US85-1008	High biomass	84.22
Criolla Rayada	Low biomass	84.36
RB72454	Low biomass	83.41
RB72454	Low biomass	82.99
RB72454	Low biomass	83.40
RB855156	Low biomass	84.93
SP80-3280	Low biomass	82.62
SP80-3280	Low biomass	85.37
SP80-3280	Low biomass	84.27
TUC71-7	Low biomass	84.98
White Transparent	Low biomass	83.93
White Transparent	Low biomass	83.65
White Transparent	Low biomass	83.69

Differential expression and functional enrichment analyses

Sample clustering based on expression

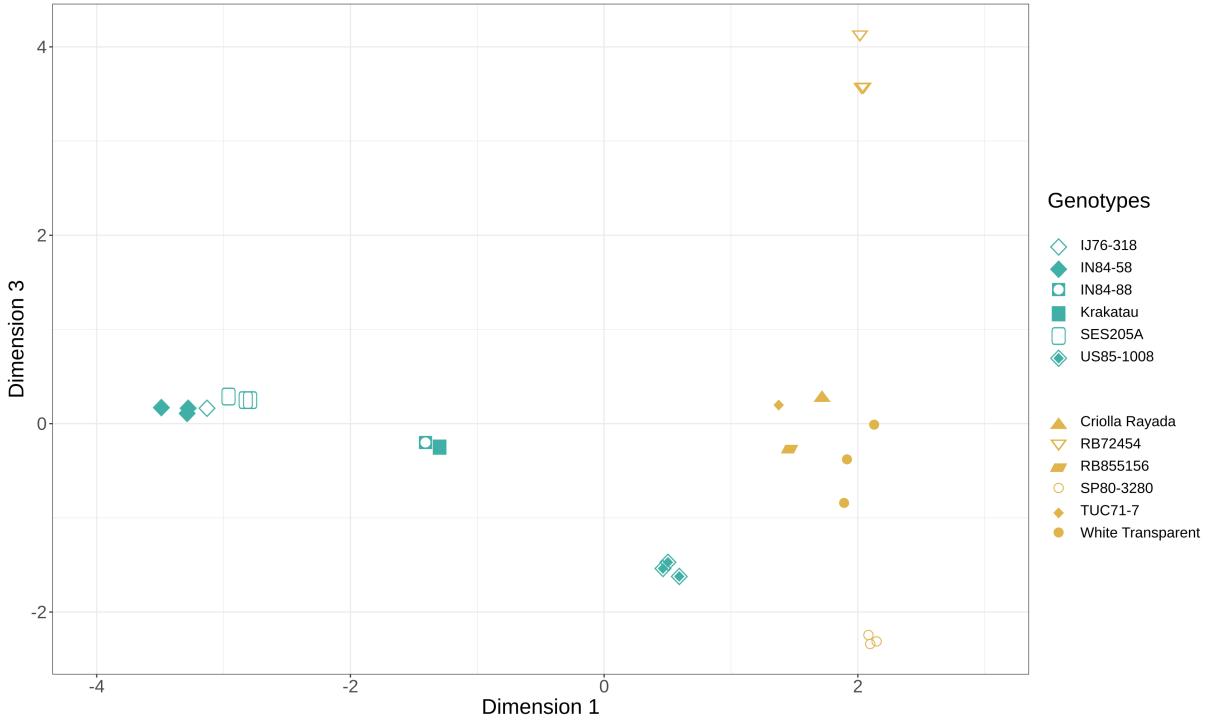

Figure 1: Multidimensional scaling to assess dissimilarities between samples. Blue symbols indicate high biomass genotypes, and low biomass genotypes are colored in orange.

Tests

Table 2: Results of the differential expression analysis in the three proposed tests: i) Low biomass group compared to high biomass group; ii) ANOVA-like test using genotypes within the high biomass group; iii) ANOVA-like test using genotypes within the low biomass group.

	Low biomass vs High biomass	**ANOVA-like high biomass**	**ANOVA-like low biomass**
Differentially expressed	21074	27981	17099
Not significantly regulated	26602	19695	30577

We evaluated Gene Ontology enriched terms in each of these tests, in the following order:

- Low biomass genotypes compared to the high biomass genotypes (Figure 3)
- Differences within the high biomass group (main document)
- Differences within the low biomass group (main document)

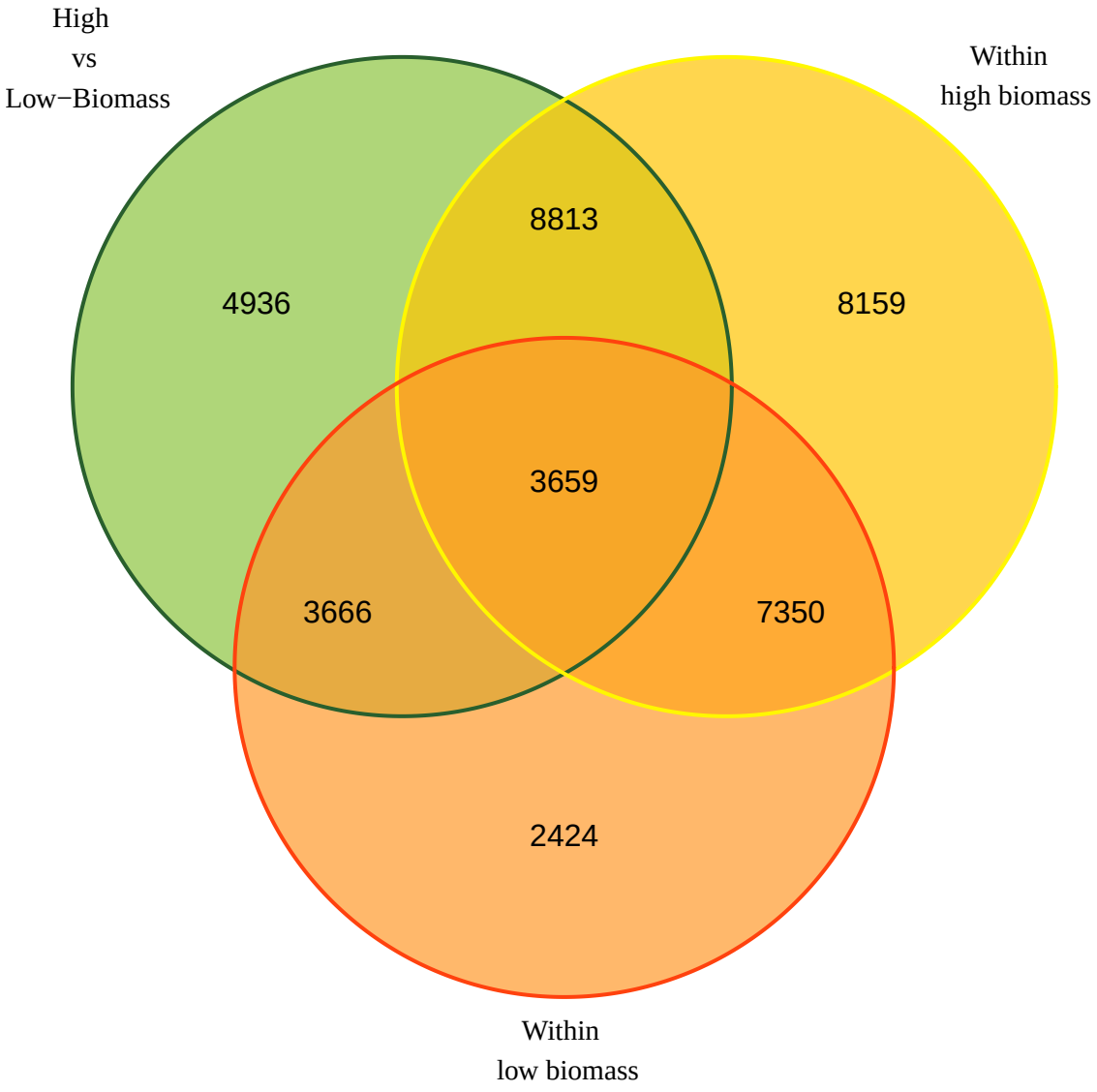

Figure 2: Venn diagram of the overlap between lists of differentially expressed genes in the three tests.

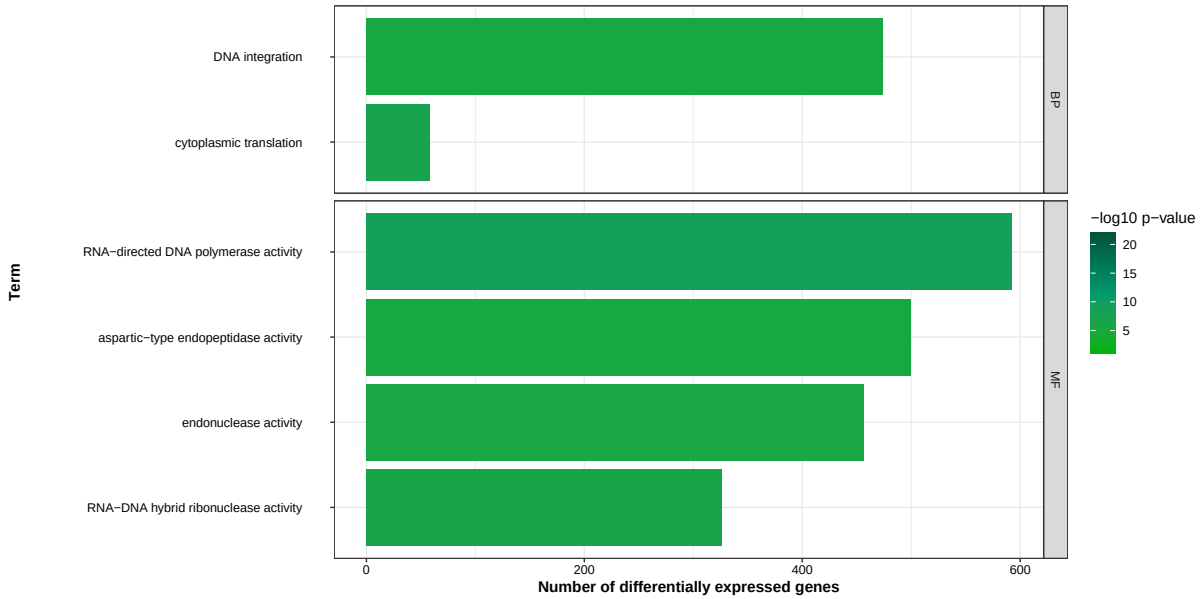

Figure 3: Bar chart of the number of DEGs in each enriched functional class for the biomass group contrast. Gene ontology categories are indicated by BP (Biological Process), CC (Cellular Component) and MF (Molecular Function).

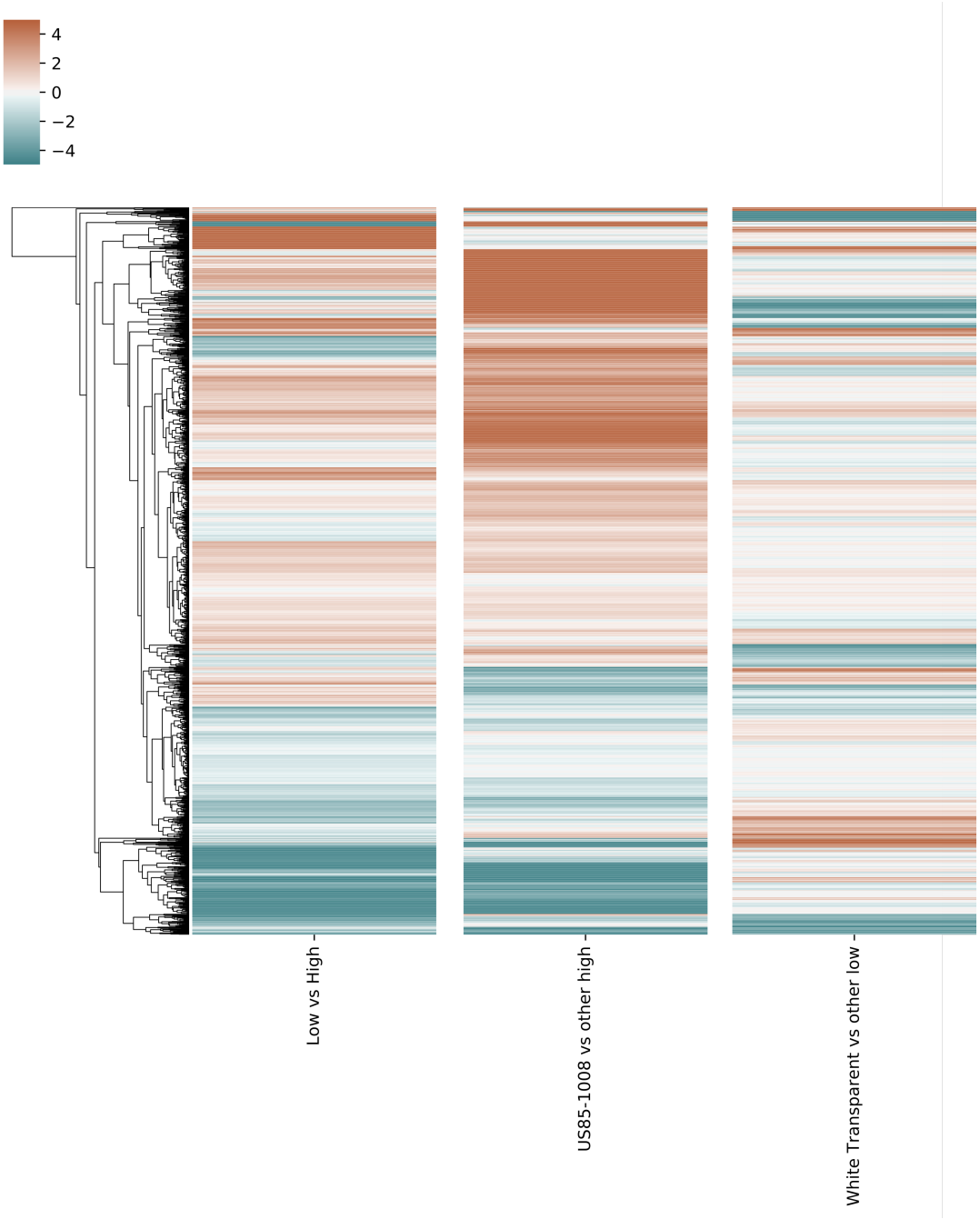

Figure 4: Heatmap of differential expression for genes associated with transposition. Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids.

Using the common DEGs among the three contrast, the enrichment corroborates differences in stress response and transposition (Figure 5). To avoid the enrichment of these often apparent terms, we performed a functional enrichment analyses using the common genes between the high and low groups contrasts, removing those common to the fiber contrast. With that we verified that sugarcane genotypes, even in a same phenotypic group, have differences in the cell wall biogenesis (Figure 6).

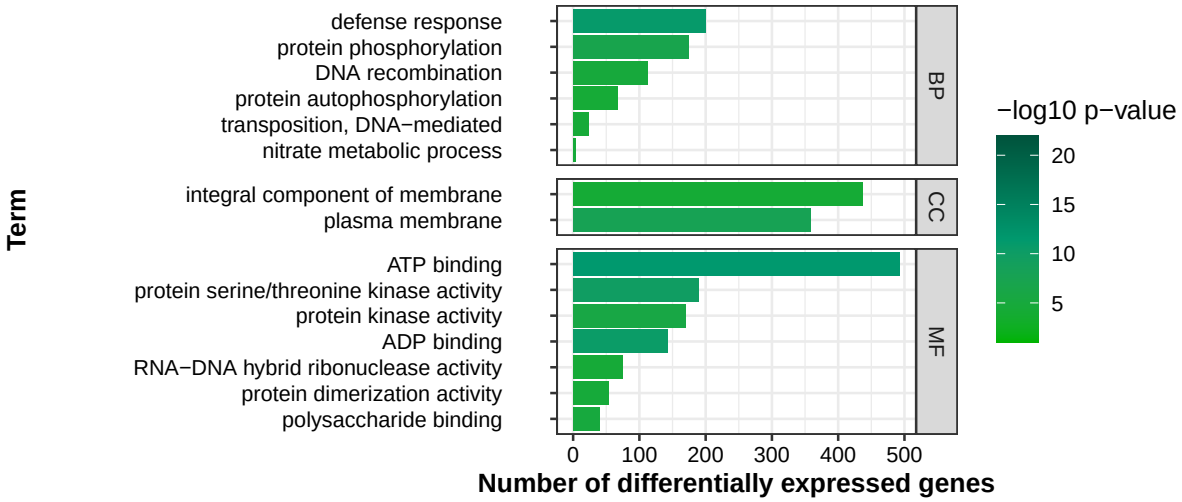

Figure 5: Bar chart of the number of DEGs in each category enriched with common DEGs between the three contrasts. Gene ontology categories are indicated by BP (Biological Process), CC (Cellular Component) and MF (Molecular Function)

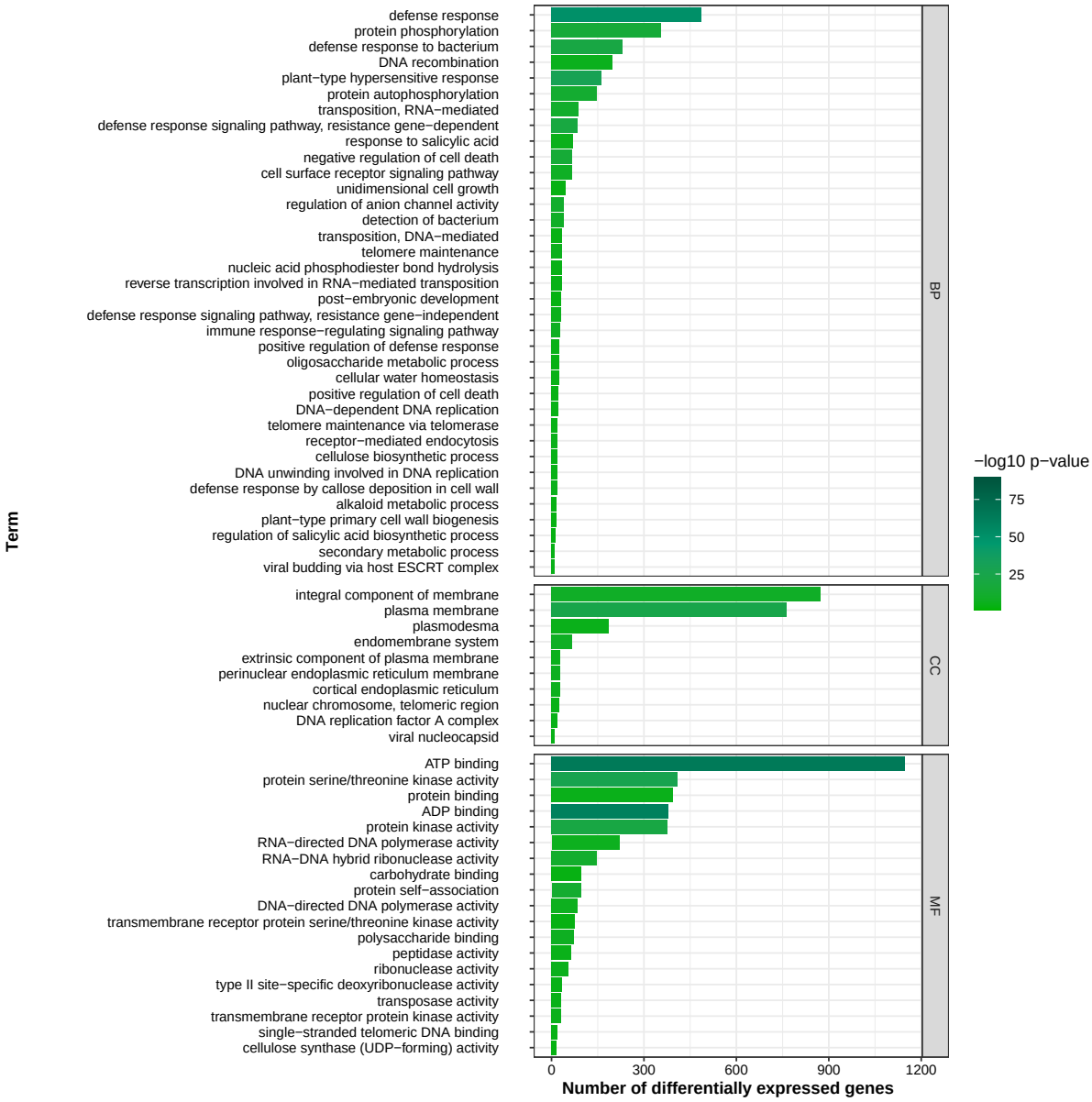

Figure 6: Bar chart of the number of DEGs in each category enriched with common DEGs between the contrasts comparing genotypes within the groups. Gene ontology categories are indicated by BP (Biological Process), CC (Cellular Component) and MF (Molecular Function)

Co-expression enrichment

Our co-expression network was built with the genes passing the expression filter. We obtained 16 modules, of which eleven showed enrichment of 289 Gene Ontology terms. Table 3 presents the Gene Ontology terms enriched in each co-expression module.

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
Module 1	GO:0003735	0.000000000	structural constituent of ribosome
	GO:0006412	0.000000000	translation
	GO:0002181	0.000000000	cytoplasmic translation
	GO:0005840	0.000000000	ribosome
	GO:0022625	0.000000000	cytosolic large ribosomal subunit
	GO:0005829	0.000000000	cytosol
	GO:0005622	0.000000000	intracellular
	GO:0033290	0.000000000	eukaryotic 48S preinitiation complex
	GO:0022627	0.000000000	cytosolic small ribosomal subunit
	GO:0005759	0.000000000	mitochondrial matrix
	GO:0001732	0.000000001	formation of cytoplasmic translation initiation complex
	GO:0016282	0.000000001	eukaryotic 43S preinitiation complex
	GO:0043161	0.000000003	proteasome-mediated ubiquitin-dependent protein catabolic process
	GO:0055114	0.000000004	oxidation-reduction process
	GO:0006099	0.000000018	tricarboxylic acid cycle
	GO:0032153	0.000000021	cell division site
	GO:0005747	0.000000058	mitochondrial respiratory chain complex I
	GO:0005852	0.000000130	eukaryotic translation initiation factor 3 complex
	GO:0030479	0.000000316	actin cortical patch
	GO:0003743	0.000000329	translation initiation factor activity
	GO:0045842	0.000000637	positive regulation of mitotic metaphase/anaphase transition
	GO:0005737	0.000000841	cytoplasm
	GO:0000329	0.000001102	fungal-type vacuole membrane
	GO:0006696	0.000001155	ergosterol biosynthetic process
	GO:0005839	0.000003294	proteasome core complex
	GO:0004298	0.000003823	threonine-type endopeptidase activity
	GO:0016491	0.000004003	oxidoreductase activity
	GO:0043066	0.000004081	negative regulation of apoptotic process
	GO:0010498	0.000004696	proteasomal protein catabolic process
	GO:0006413	0.000005148	translational initiation
	GO:0000502	0.000006289	proteasome complex
	GO:0002183	0.000007142	cytoplasmic translational initiation
	GO:0000272	0.000008485	polysaccharide catabolic process
	GO:0005838	0.000009464	proteasome regulatory particle
	GO:0015986	0.000010074	ATP synthesis coupled proton transport
	GO:0032543	0.000010163	mitochondrial translation
	GO:0010499	0.000012071	proteasomal ubiquitin-independent protein catabolic process

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
	GO:0000276	0.000014544	mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)
	GO:0006457	0.000015568	protein folding
	GO:0006048	0.000015596	UDP-N-acetylglucosamine biosynthetic process
	GO:0051082	0.000018026	unfolded protein binding
	GO:0034622	0.000023142	cellular macromolecular complex assembly
	GO:0008540	0.000026175	proteasome regulatory particle, base subcomplex
	GO:0022624	0.000029125	proteasome accessory complex
	GO:0006120	0.000032853	mitochondrial electron transport, NADH to ubiquinone
	GO:0004175	0.000036743	endopeptidase activity
	GO:0005743	0.000042846	mitochondrial inner membrane
	GO:0036402	0.000050085	proteasome-activating ATPase activity
	GO:0004099	0.000058786	chitin deacetylase activity
	GO:0006119	0.000066176	oxidative phosphorylation
	GO:0000921	0.000068492	septin ring assembly
	GO:0031105	0.000068492	septin complex
	GO:0032160	0.000068492	septin filament array
	GO:1903475	0.000073496	mitotic actomyosin contractile ring assembly
	GO:0006620	0.000075607	posttranslational protein targeting to endoplasmic reticulum membrane
	GO:0009405	0.000076071	pathogenesis
	GO:0015934	0.000076610	large ribosomal subunit
	GO:0000001	0.000077394	mitochondrion inheritance
	GO:0009062	0.000077940	fatty acid catabolic process
	GO:0004129	0.000087082	cytochrome-c oxidase activity
	GO:0030544	0.000101316	Hsp70 protein binding
	GO:0031072	0.000108075	heat shock protein binding
	GO:0005686	0.000113360	U2 snRNP
	GO:0019878	0.000116452	lysine biosynthetic process via aminoadipic acid
	GO:0045899	0.000122924	positive regulation of RNA polymerase II transcriptional preinitiation complex assembly
	GO:0030234	0.000130228	enzyme regulator activity
	GO:0099132	0.000131513	ATP hydrolysis coupled cation transmembrane transport
	GO:0031204	0.000133465	posttranslational protein targeting to membrane, translocation
	GO:0042254	0.000149007	ribosome biogenesis
	GO:0005685	0.000154174	U1 snRNP
	GO:0034515	0.000164803	proteasome storage granule
	GO:0043248	0.000169156	proteasome assembly
	GO:0000027	0.000183653	ribosomal large subunit assembly
	GO:0000050	0.000193815	urea cycle

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
	GO:0004753	0.000203954	saccharopine dehydrogenase activity
	GO:0000028	0.000222564	ribosomal small subunit assembly
	GO:0042788	0.000259576	polysomal ribosome
	GO:0005940	0.000324330	septin ring
	GO:0007264	0.000369241	small GTPase mediated signal transduction
	GO:0008541	0.000389399	proteasome regulatory particle, lid subcomplex
	GO:0006032	0.000410471	chitin catabolic process
Module 2	GO:0043531	0.000004750	ADP binding
Module 3	GO:0000943	0.000000001	retrotransposon nucleocapsid
	GO:0008270	0.000000001	zinc ion binding
	GO:0009507	0.000000003	chloroplast
	GO:0009451	0.000000006	RNA modification
	GO:0015074	0.000000012	DNA integration
	GO:0004519	0.000000056	endonuclease activity
	GO:0003964	0.000000066	RNA-directed DNA polymerase activity
	GO:0004190	0.000000083	aspartic-type endopeptidase activity
	GO:0003676	0.000002277	nucleic acid binding
	GO:0006310	0.000003950	DNA recombination
	GO:0009570	0.000027960	chloroplast stroma
	GO:0007004	0.000049015	telomere maintenance via telomerase
	GO:0006261	0.000050414	DNA-dependent DNA replication
Module 4	GO:0043531	0.000000300	ADP binding
	GO:0047268	0.000001561	galactinol-raffinose galactosyltransferase activity
Module 7	GO:0003735	0.000000000	structural constituent of ribosome
	GO:0005730	0.000000000	nucleolus
	GO:0006412	0.000000000	translation
	GO:0005840	0.000000000	ribosome
	GO:0006364	0.000000000	rRNA processing
	GO:0022625	0.000000000	cytosolic large ribosomal subunit
	GO:0003723	0.000000000	RNA binding
	GO:0022627	0.000000000	cytosolic small ribosomal subunit
	GO:0005622	0.000000000	intracellular
	GO:0005739	0.000000000	mitochondrion
	GO:0032040	0.000000000	small-subunit processome
	GO:0005829	0.000000000	cytosol
	GO:0019843	0.000000000	rRNA binding
	GO:0022626	0.000000000	cytosolic ribosome
	GO:0042254	0.000000000	ribosome biogenesis
	GO:0000462	0.000000000	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
	GO:0000027	0.000000000	ribosomal large subunit assembly
	GO:0042273	0.000000000	ribosomal large subunit biogenesis

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
	GO:0000447	0.000000001	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
	GO:0000480	0.000000001	endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
	GO:0051082	0.000000003	unfolded protein binding
	GO:0006457	0.000000007	protein folding
	GO:0000028	0.000000026	ribosomal small subunit assembly
	GO:0000472	0.000000026	endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
	GO:0008033	0.000000040	tRNA processing
	GO:0030515	0.000000044	snoRNA binding
	GO:0006397	0.000000068	mRNA processing
	GO:0003729	0.000000119	mRNA binding
	GO:0031167	0.000000226	rRNA methylation
	GO:0017056	0.000000229	structural constituent of nuclear pore
	GO:0006414	0.000000286	translational elongation
	GO:0030687	0.000000873	preribosome, large subunit precursor
	GO:0009408	0.000001166	response to heat
	GO:0000049	0.000001168	tRNA binding
	GO:0000463	0.000001366	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
	GO:1990904	0.000001859	ribonucleoprotein complex
	GO:0006606	0.000002675	protein import into nucleus
	GO:0034388	0.000003248	Pwp2p-containing subcomplex of 90S preribosome
	GO:0019919	0.000004851	peptidyl-arginine methylation, to asymmetrical-dimethyl arginine
	GO:0000338	0.000005375	protein deneddylation
	GO:0030686	0.000005630	90S preribosome
	GO:0000055	0.000006631	ribosomal large subunit export from nucleus
	GO:0009507	0.000006957	chloroplast
	GO:0005762	0.000008474	mitochondrial large ribosomal subunit
	GO:0035242	0.000008551	protein-arginine omega-N asymmetric methyltransferase activity
	GO:0005654	0.000009141	nucleoplasm
	GO:0008180	0.000010412	COP9 signalosome
	GO:0031429	0.000010913	box H/ACA snoRNP complex
	GO:0005682	0.000011762	U5 snRNP
	GO:0034513	0.000014963	box H/ACA snoRNA binding
	GO:0003899	0.000015136	DNA-directed 5'-3' RNA polymerase activity
	GO:0016282	0.000016113	eukaryotic 43S preinitiation complex
	GO:0006413	0.000016417	translational initiation

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
	GO:0008168	0.000017298	methyltransferase activity
	GO:0005681	0.000020908	spliceosomal complex
	GO:0004812	0.000020909	aminoacyl-tRNA ligase activity
	GO:0042788	0.000021288	polysomal ribosome
	GO:0034336	0.000021974	misfolded RNA binding
	GO:0060567	0.000021974	negative regulation of DNA-templated transcription, termination
	GO:0008469	0.000022216	histone-arginine N-methyltransferase activity
	GO:0034969	0.000022216	histone arginine methylation
	GO:0003743	0.000026370	translation initiation factor activity
	GO:0006396	0.000026802	RNA processing
	GO:0016554	0.000039355	cytidine to uridine editing
	GO:0000176	0.000042088	nuclear exosome (RNase complex)
	GO:0004386	0.000042349	helicase activity
	GO:0032543	0.000057521	mitochondrial translation
	GO:0001732	0.000061808	formation of cytoplasmic translation initiation complex
	GO:0000398	0.000076151	mRNA splicing, via spliceosome
	GO:0030488	0.000089460	tRNA methylation
	GO:0051117	0.000123216	ATPase binding
	GO:0009536	0.000144513	plastid
	GO:0034511	0.000169454	U3 snoRNA binding
	GO:0032955	0.000200647	regulation of division septum assembly
	GO:0033290	0.000208091	eukaryotic 48S preinitiation complex
	GO:0005736	0.000233738	DNA-directed RNA polymerase I complex
	GO:0042134	0.000243525	rRNA primary transcript binding
	GO:0051028	0.000251357	mRNA transport
	GO:0000469	0.000257832	cleavage involved in rRNA processing
	GO:0005832	0.000258810	chaperonin-containing T-complex
	GO:0006418	0.000300001	tRNA aminoacylation for protein translation
	GO:0003871	0.000302662	5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase activity
	GO:0010157	0.000325097	response to chlorate
	GO:0004328	0.000335979	formamidase activity
	GO:0006383	0.000337829	transcription from RNA polymerase III promoter
	GO:0030295	0.000367275	protein kinase activator activity
	GO:0009295	0.000368216	nucleoid
	GO:0031118	0.000385301	rRNA pseudouridine synthesis
	GO:0016811	0.000392026	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides
	GO:0070181	0.000408561	small ribosomal subunit rRNA binding
	GO:0005732	0.000425031	small nucleolar ribonucleoprotein complex
	GO:0043021	0.000440732	ribonucleoprotein complex binding
	GO:0009631	0.000445880	cold acclimation
	GO:0031428	0.000447993	box C/D snoRNP complex
	GO:0080156	0.000473193	mitochondrial mRNA modification

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
	GO:0005635	0.000489425	nuclear envelope
	GO:0006360	0.000499518	transcription from RNA polymerase I promoter
Module 8	GO:0003964	0.000000000	RNA-directed DNA polymerase activity
	GO:0015074	0.000000000	DNA integration
	GO:0004190	0.000000000	aspartic-type endopeptidase activity
	GO:0006310	0.000000000	DNA recombination
	GO:0003887	0.000000000	DNA-directed DNA polymerase activity
	GO:0004519	0.000000000	endonuclease activity
	GO:0009507	0.000000000	chloroplast
	GO:0000943	0.000000000	retrotransposon nucleocapsid
	GO:0003676	0.000000000	nucleic acid binding
	GO:0004523	0.000000000	RNA-DNA hybrid ribonuclease activity
	GO:0032197	0.000000000	transposition, RNA-mediated
	GO:0004540	0.000000000	ribonuclease activity
	GO:0008233	0.000000000	peptidase activity
	GO:0009570	0.000000000	chloroplast stroma
	GO:0003723	0.000000000	RNA binding
	GO:0008270	0.000000000	zinc ion binding
	GO:0009941	0.000000002	chloroplast envelope
	GO:0046872	0.000001134	metal ion binding
	GO:0006313	0.000007897	transposition, DNA-mediated
	GO:0009535	0.000014685	chloroplast thylakoid membrane
	GO:0004803	0.000016409	transposase activity
Module 9	GO:0006310	0.000000003	DNA recombination
	GO:0003964	0.000000007	RNA-directed DNA polymerase activity
	GO:0004519	0.000000137	endonuclease activity
	GO:0006468	0.000002684	protein phosphorylation
	GO:0046872	0.000005693	metal ion binding
	GO:0004674	0.000021446	protein serine/threonine kinase activity
Module 10	GO:0016021	0.000000000	integral component of membrane
	GO:0006355	0.000000001	regulation of transcription, DNA-templated
	GO:0005515	0.000000002	protein binding
	GO:0003700	0.000000033	DNA binding transcription factor activity
	GO:0006970	0.000000277	response to osmotic stress
	GO:0005794	0.000001711	Golgi apparatus
	GO:0005886	0.000002640	plasma membrane
	GO:0016020	0.000006302	membrane
	GO:0043565	0.000006311	sequence-specific DNA binding
	GO:0007275	0.000007709	multicellular organism development
	GO:0042285	0.000027768	xylosyltransferase activity
	GO:0072583	0.000031484	clathrin-dependent endocytosis
	GO:0006468	0.000040283	protein phosphorylation
	GO:0015031	0.000043071	protein transport
Module 11	GO:0043531	0.000000000	ADP binding
	GO:0009626	0.000000000	plant-type hypersensitive response
	GO:0005524	0.000000010	ATP binding

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
	GO:0009870	0.000000061	defense response signaling pathway, resistance gene-dependent
	GO:0006952	0.000000321	defense response
	GO:0060548	0.000003589	negative regulation of cell death
	GO:0009507	0.000005396	chloroplast
	GO:0033201	0.000012063	alpha-1,4-glucan synthase activity
	GO:0009535	0.000015676	chloroplast thylakoid membrane
	GO:0009011	0.000022960	starch synthase activity
Module 12	GO:0005886	0.000000001	plasma membrane
	GO:0006952	0.000000001	defense response
	GO:0043531	0.000000004	ADP binding
	GO:0016021	0.000000171	integral component of membrane
Module 16	GO:0009535	0.000000000	chloroplast thylakoid membrane
	GO:0015979	0.000000000	photosynthesis
	GO:0009522	0.000000000	photosystem I
	GO:0009507	0.000000000	chloroplast
	GO:0018298	0.000000000	protein-chromophore linkage
	GO:0009523	0.000000000	photosystem II
	GO:0009538	0.000000007	photosystem I reaction center
	GO:0016168	0.000000024	chlorophyll binding
	GO:0009739	0.000000078	response to gibberellin
	GO:0009789	0.000000085	positive regulation of abscisic acid-activated signaling pathway
	GO:0010598	0.000000126	NAD(P)H dehydrogenase complex (plastoquinone)
	GO:0003700	0.000000241	DNA binding transcription factor activity
	GO:0009772	0.000000342	photosynthetic electron transport in photosystem II
	GO:0070413	0.000000400	trehalose metabolism in response to stress
	GO:0009723	0.000000422	response to ethylene
	GO:0009416	0.000000441	response to light stimulus
	GO:0010319	0.000001268	stromule
	GO:0009767	0.000001277	photosynthetic electron transport chain
	GO:0010287	0.000001283	plastoglobule
	GO:0042651	0.000001324	thylakoid membrane
	GO:0007623	0.000001819	circadian rhythm
	GO:0009768	0.000002043	photosynthesis, light harvesting in photosystem I
	GO:0009773	0.000002071	photosynthetic electron transport in photosystem I
	GO:0090229	0.000002188	negative regulation of red or far-red light signaling pathway
	GO:0009409	0.000006625	response to cold
	GO:0006355	0.000012645	regulation of transcription, DNA-templated
	GO:0005992	0.000012767	trehalose biosynthetic process
	GO:0009737	0.000019672	response to abscisic acid
	GO:0009640	0.000019796	photomorphogenesis

Table 3: Gene Ontology terms enriched in each co-expression module.

Module	Category	Overrepresented p-value	Term
	GO:0046524	0.000021765	sucrose-phosphate synthase activity
	GO:0048038	0.000029880	quinone binding
	GO:0031969	0.000039092	chloroplast membrane
	GO:0009882	0.000041604	blue light photoreceptor activity
	GO:0009579	0.000043182	thylakoid
	GO:0080006	0.000129885	internode patterning
	GO:0009941	0.000136129	chloroplast envelope
	GO:0016311	0.000137838	dephosphorylation
	GO:1902448	0.000173188	positive regulation of shade avoidance
	GO:0009735	0.000176429	response to cytokinin
	GO:0019684	0.000211381	photosynthesis, light reaction

We created a Word Cloud representation using a word frequency greater than one in each enriched module to check the most common words (Figure 7).

We used the Gene Set Enrichment Analysis (GSEA) and permuted the genes of the modules 10,000 times in the ranked LFC lists of the following contrasts: i) Low against high biomass; ii) US85-1008 compared to the mean of SES205A and IN84-58; iii) White Transparent compared to the mean of RB72454 and SP80-3280. We found that module 16 was enriched with genes with high absolute LFC values in the three contrasts (Figure 9). Genes in module 16 were positivelly correlated with genes of high LFC in the biomass contrast and in the comparison of US85-1008 with two *S. spontaneum* genotypes. Ranked genes from the contrast comparing the *S. officinarum* White Transparent to the hybrids were negativelly correlated with genes within module 16 .

To visualize the expression profile of each module, we assessed the expression level of the eigengenes (Figure 10). We observed that at least five modules were marked by a expression peak or valley for a single genotype. Module 16 contains genes with higher expression in sucrose-rich genotypes, opposite to module 3. In both cases US85-1008 was in the high expression group. The profile of the eigengene of Module 16 indicates a higher expression in the low biomass group, but without a substantial variability among the samples within the group. According to the GSEA, in this module the low-biomass genotypes did not contain genes with high LFC.

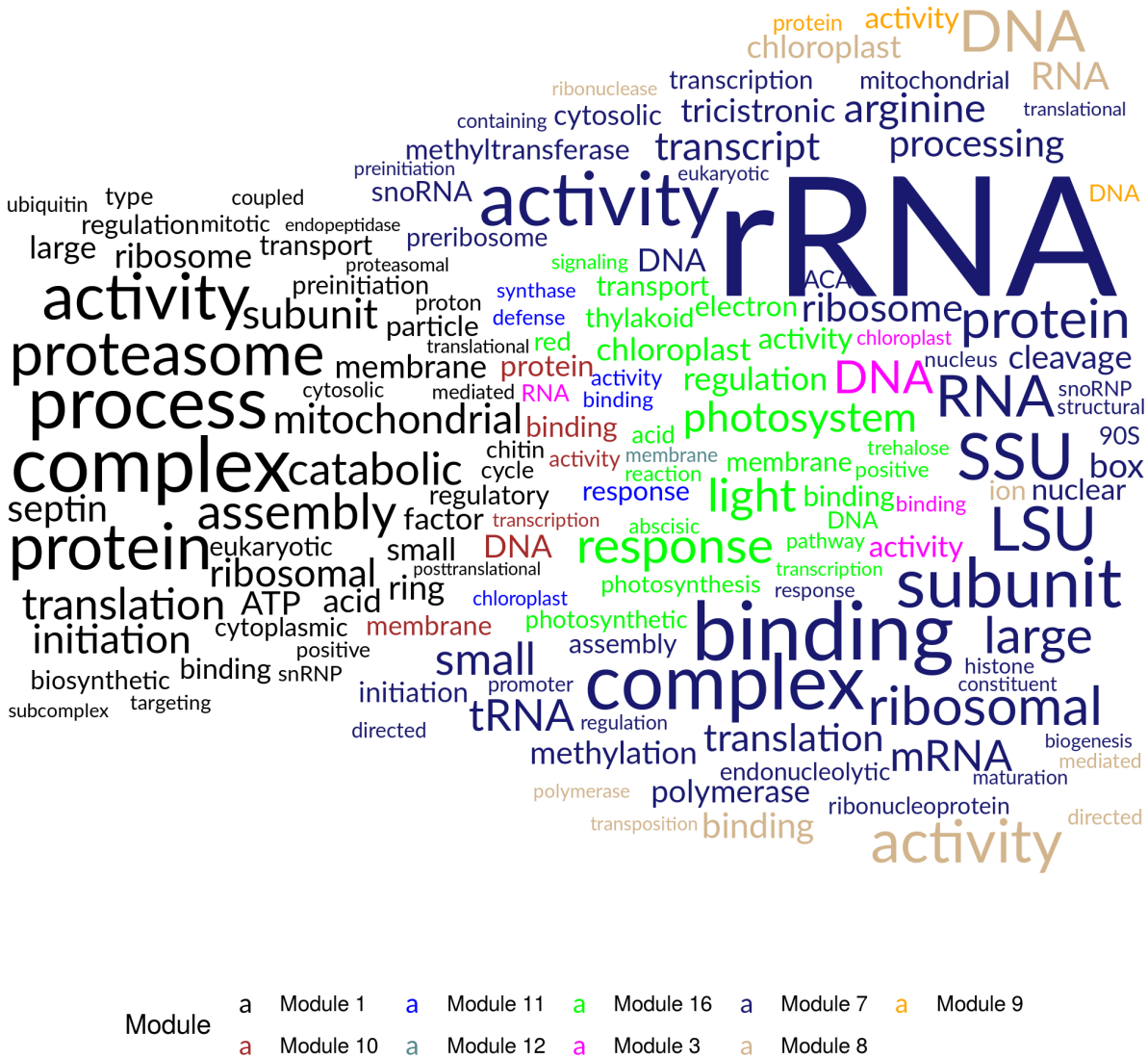

Figure 7: Word cloud using Gene Ontology descriptions for terms enriched with genes in the co-expression modules. This image is restricted to present only nine modules, those that presented enriched GO terms.

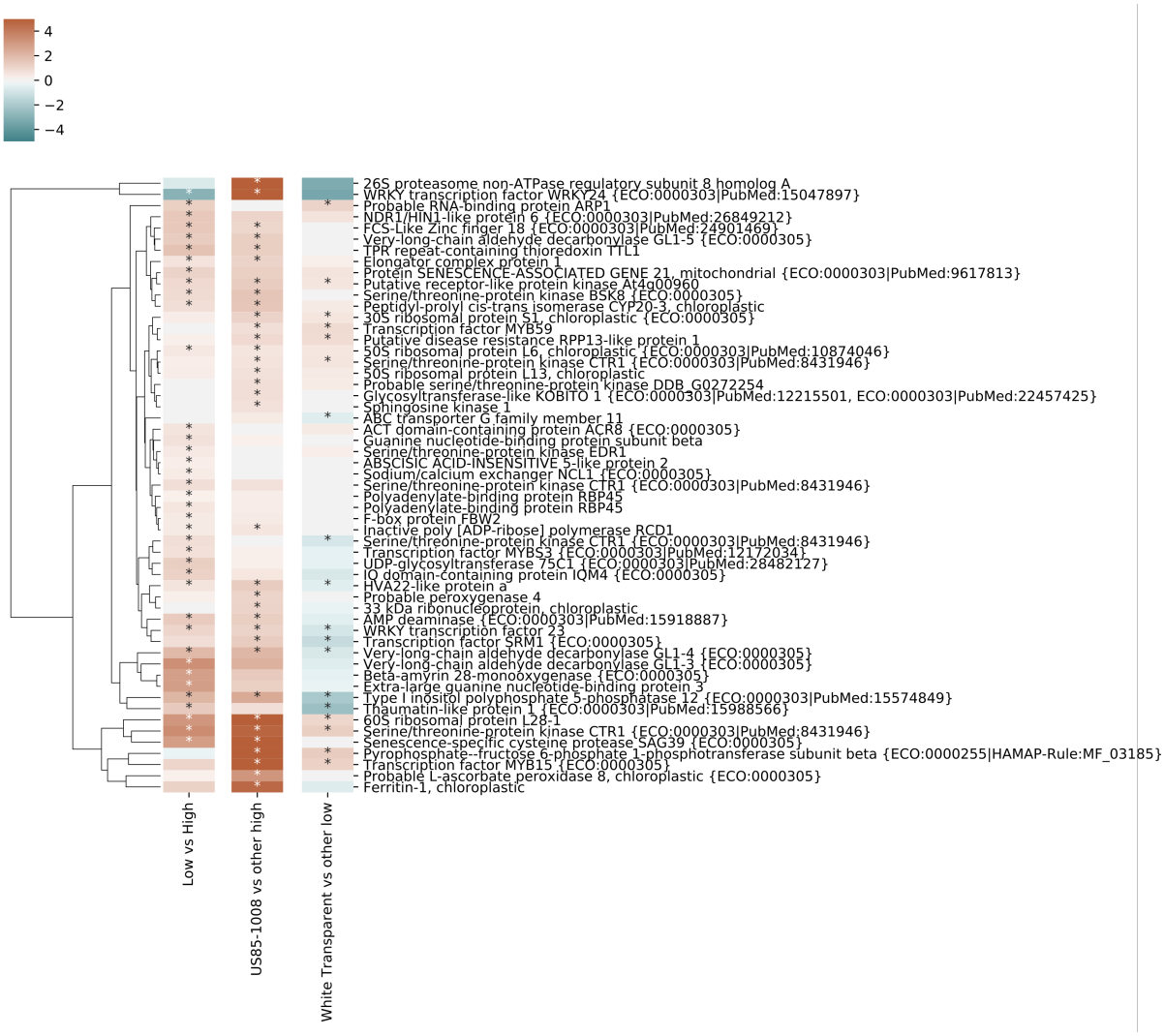

Figure 8: Expression of hormone response DEGs present in the co-expression module 16. These genes were functionally annotated to the biological processes of responses to abscisic acid, cytokinin, ethylene or gibberellin. Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Significantly differentially expressed genes are indicated by asterisks.

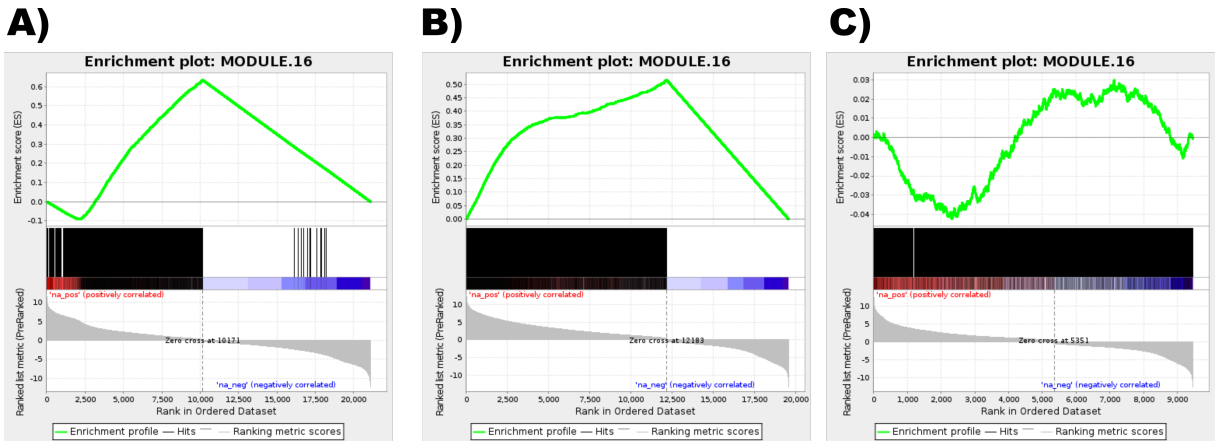

Figure 9: Gene Set Enrichment of the Module 16 co-expressed genes. Gene set enrichment using genes ranked based on absolute LFC. (A) Low-fiber genotypes contrasted to the high-fiber group. (B) US85-1008 contrasted to the mean of SES205A and IN84-58. (C) White Transparent compared to the mean of the hybrids RB72454 and SP80-3280

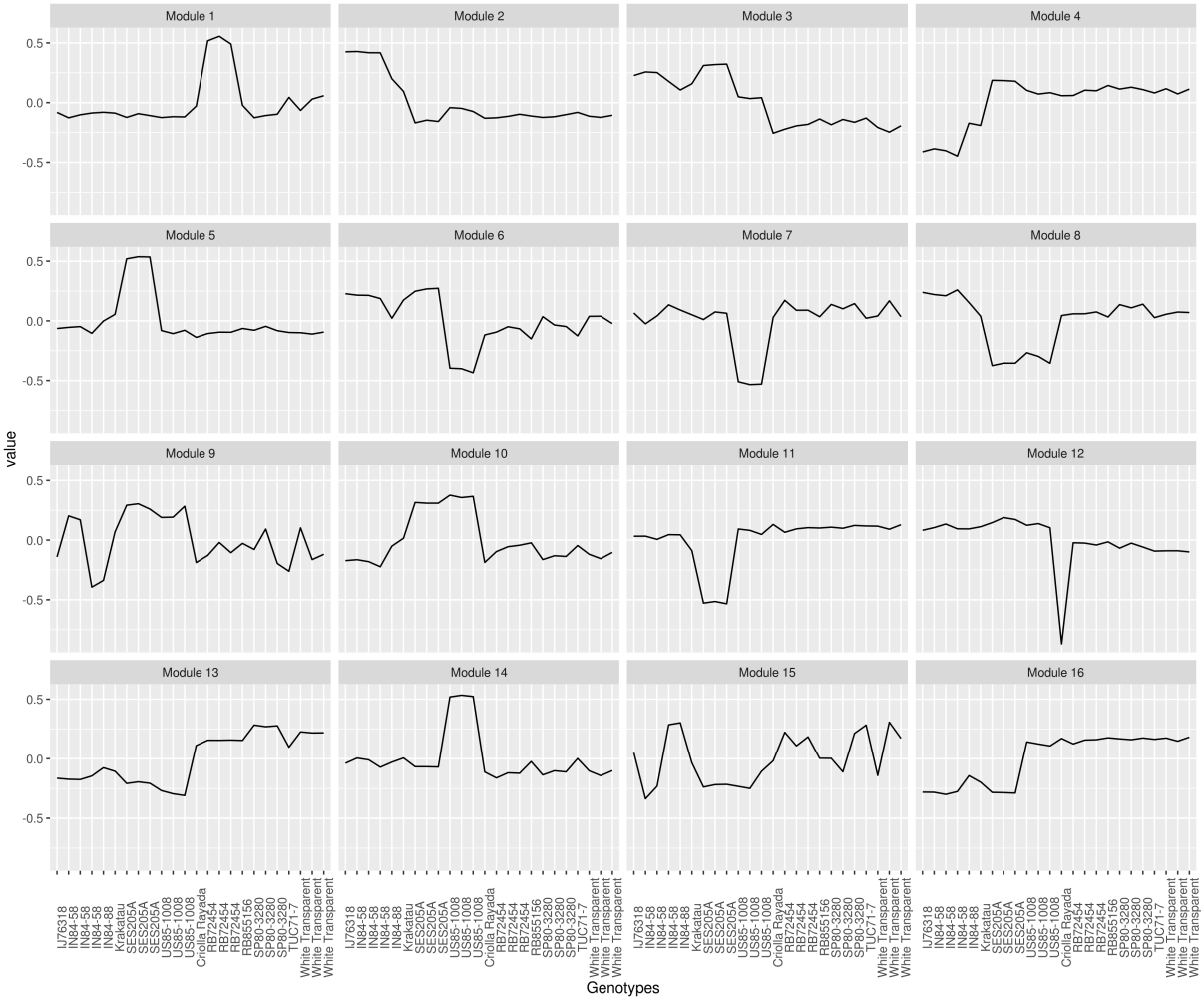

Figure 10: Expression profile of the eigengenes from each module.

Pathway analysis with Gene Ontology terms and MapMan4

We explored the pathways provided by MapMan4 to associate up and downregulated DEGs with metabolic processes. We first used all the isoforms of a gene to map to the functional annotation BINs in Mercator4. Next, in MapMan we used the log fold change of the DEGs identified in each contrast evaluated. Here we present the results of the *metabolism overview* and *lignin* pathways.

Metabolism overview

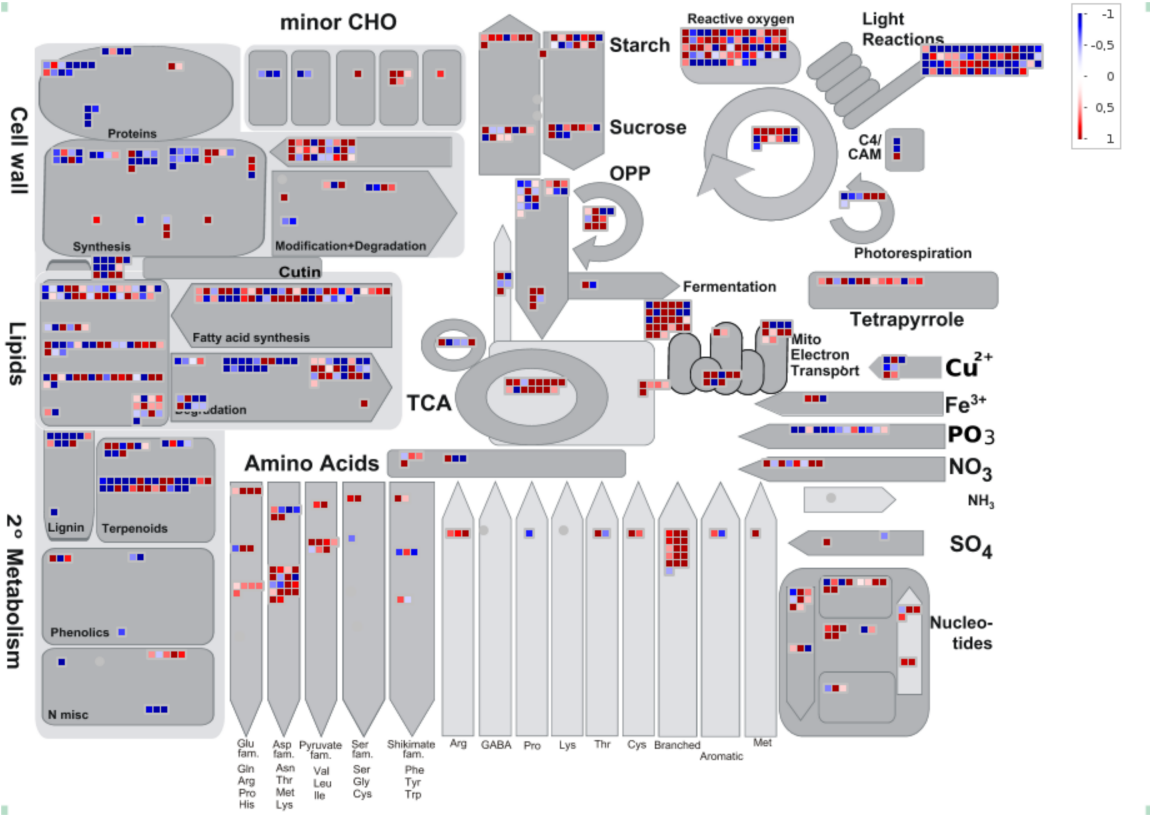

Figure 11: Metabolism overview mapping using the log of fold change of the DEGs from the low biomass genotypes compared to the high biomass group. Genes significantly upregulated were colored in red, while those downregulated were colored in blue.

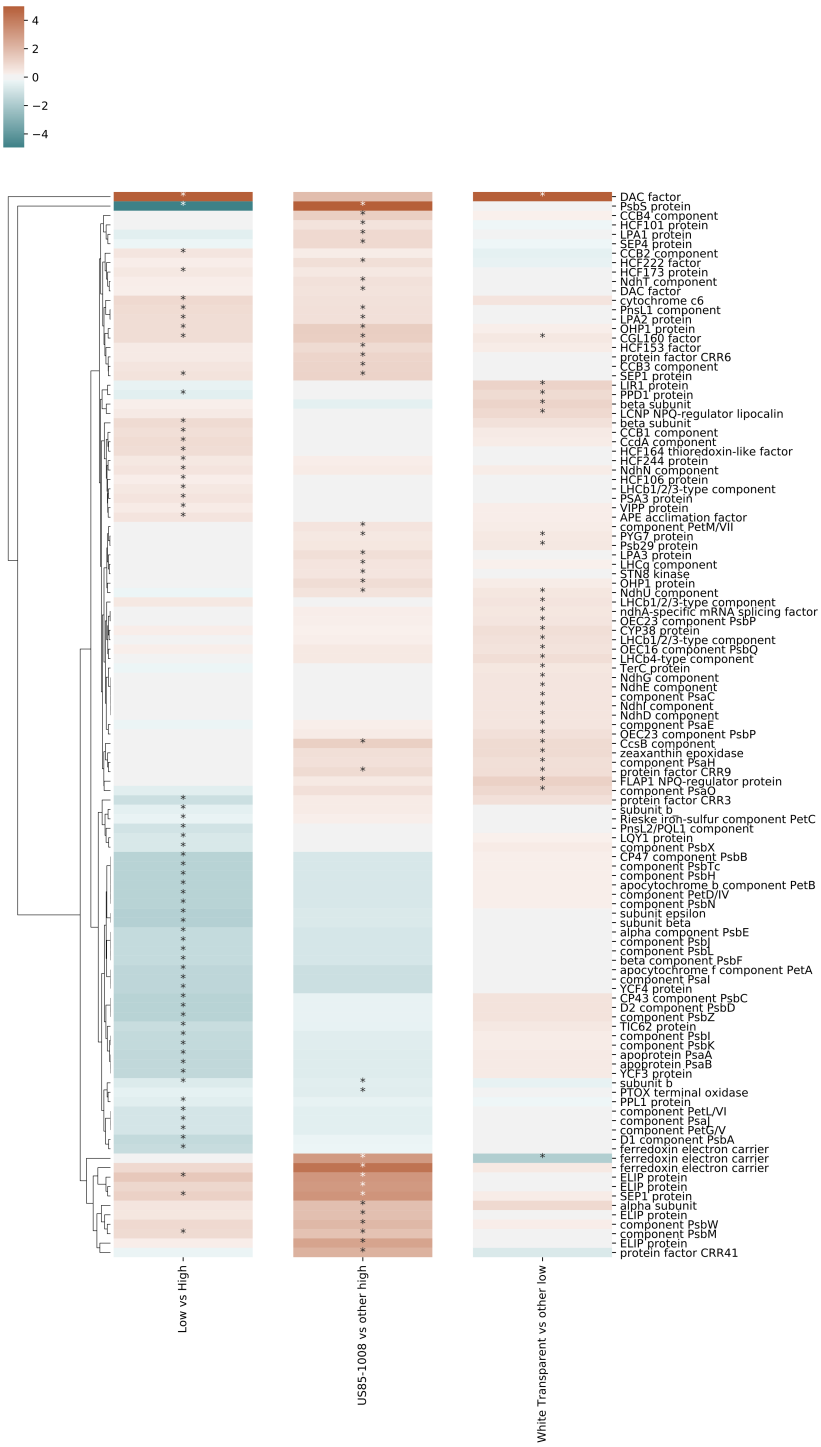

Figure 12: Heatmap for MapMan Photophosphorylatyon annotation. Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Differentially expressed genes of the contrast are indicated by asterisks.

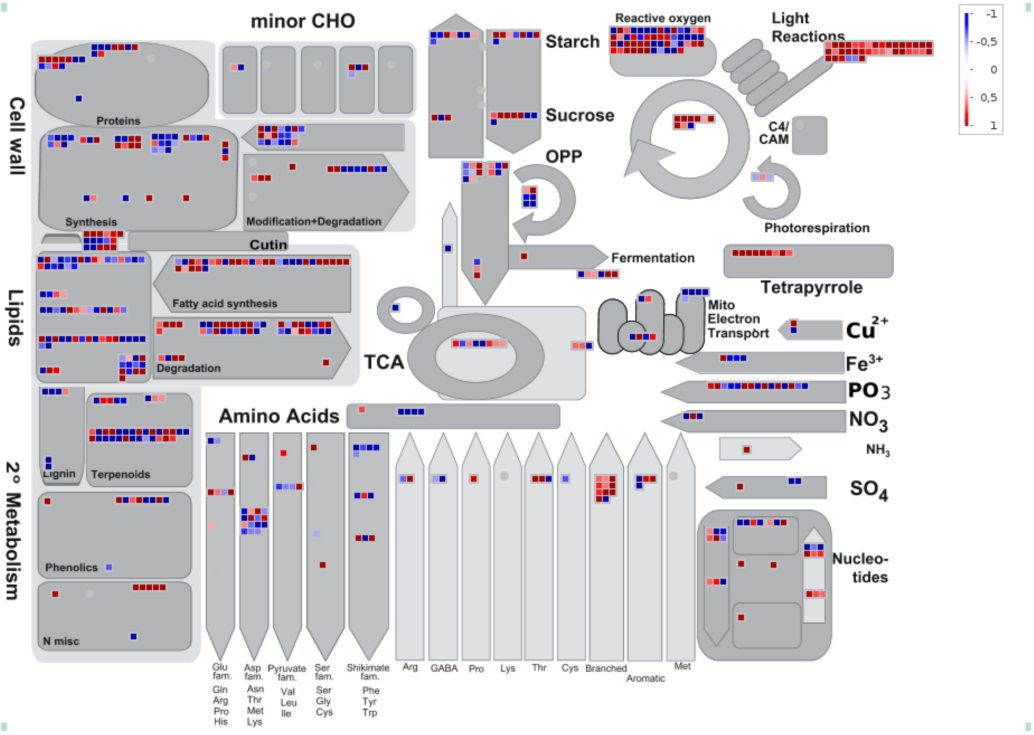

Figure 13: Metabolism overview mapping using the log of fold change of the DEGs from the comparison between US85-1008 and the mean of SES205A and IN84-58. Genes upregulated in US85-1008 were colored in red and those downregulated were colored in blue.

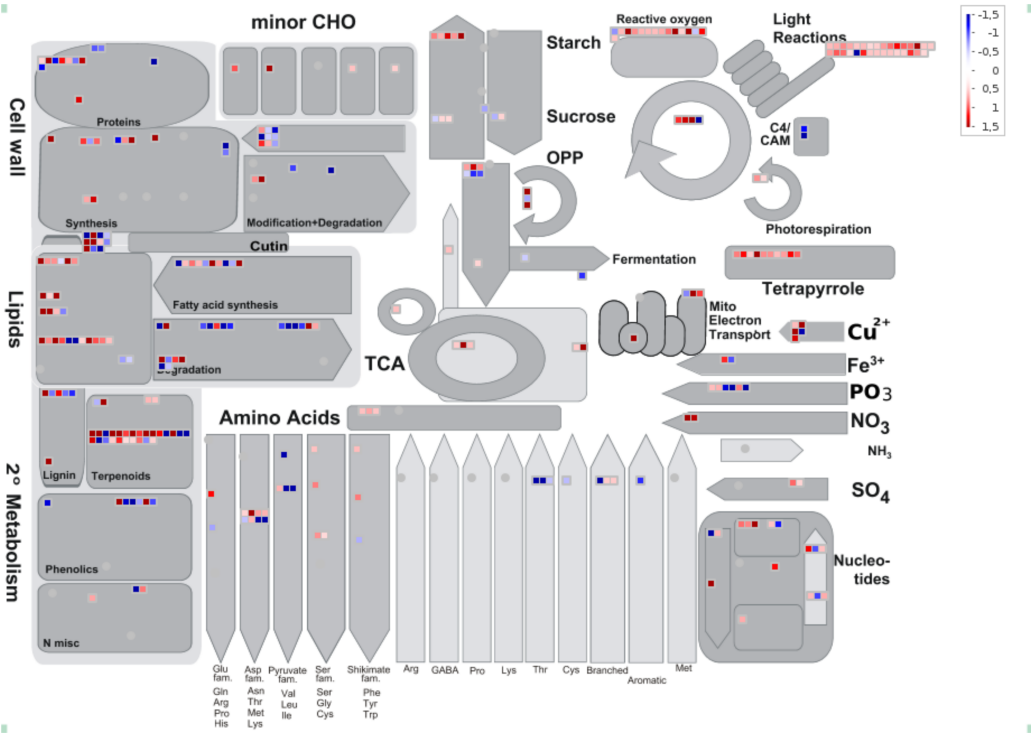

Figure 14: Metabolism overview mapping using the log of fold change of the DEGs from the comparison between White Transparent and the mean of RB72454 and SP80-3280. Genes upregulated in White Transparent were colored in red and those downregulated were colored in blue.

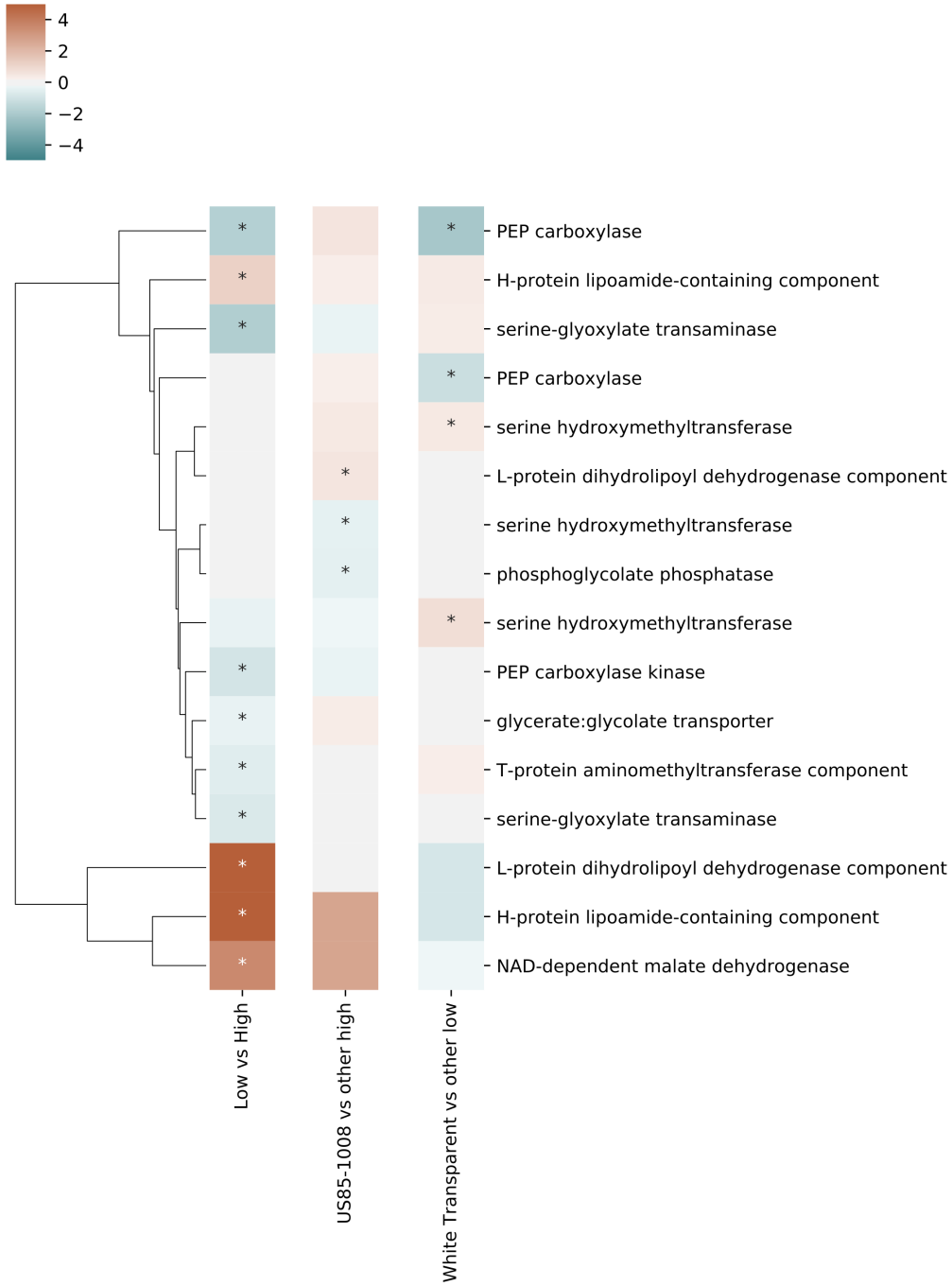

Figure 15: Heatmap for MapMan C4/CAM photosynthesis and photorespiration annotations. Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Differentially expressed genes of the contrast are indicated by asterisks.

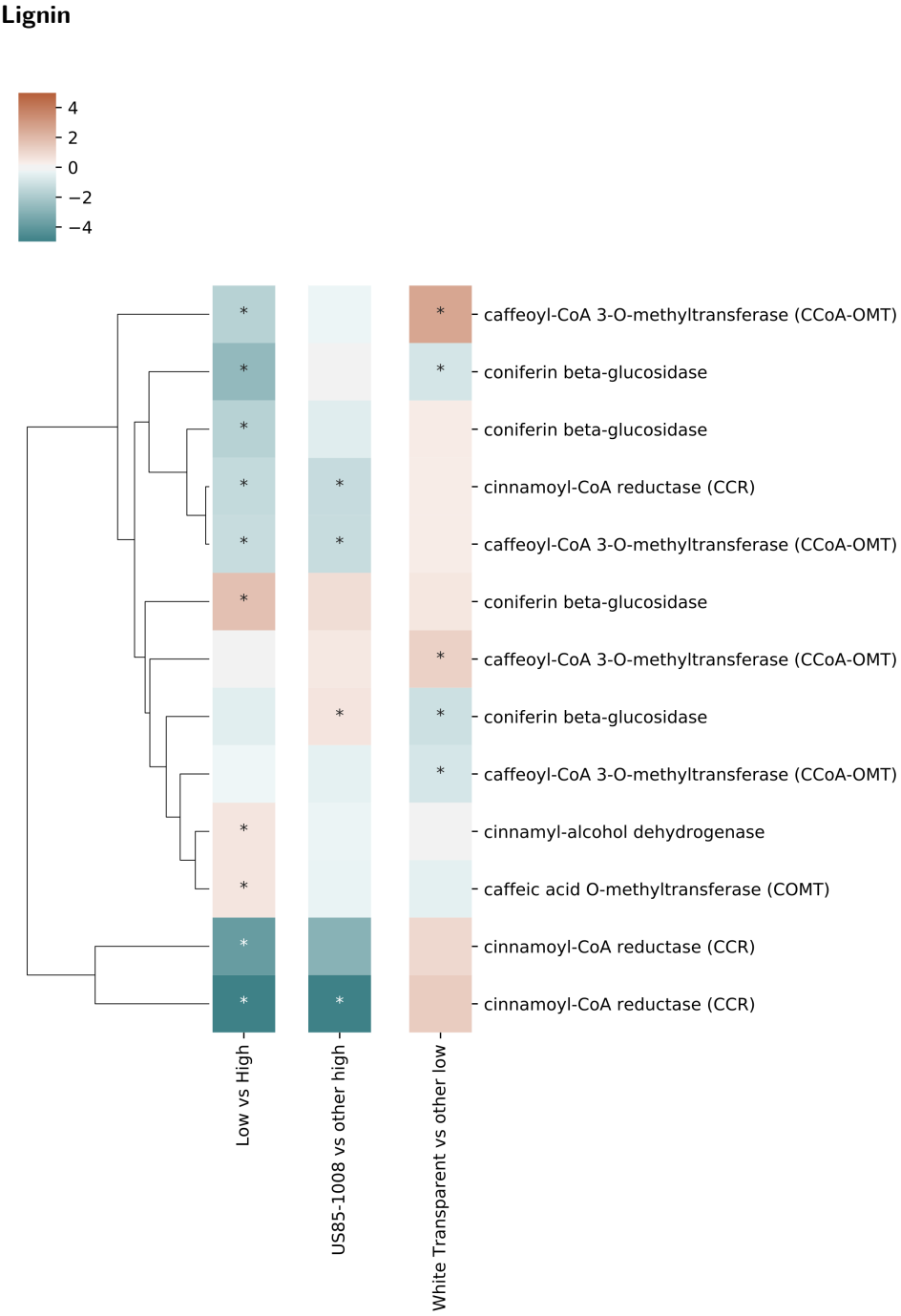

Figure 16: Heatmap for MapMan monolignol synthesis and monolignol glycosylation and deglycosylation. Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Differentially expressed genes of the contrast are indicated by asterisks.

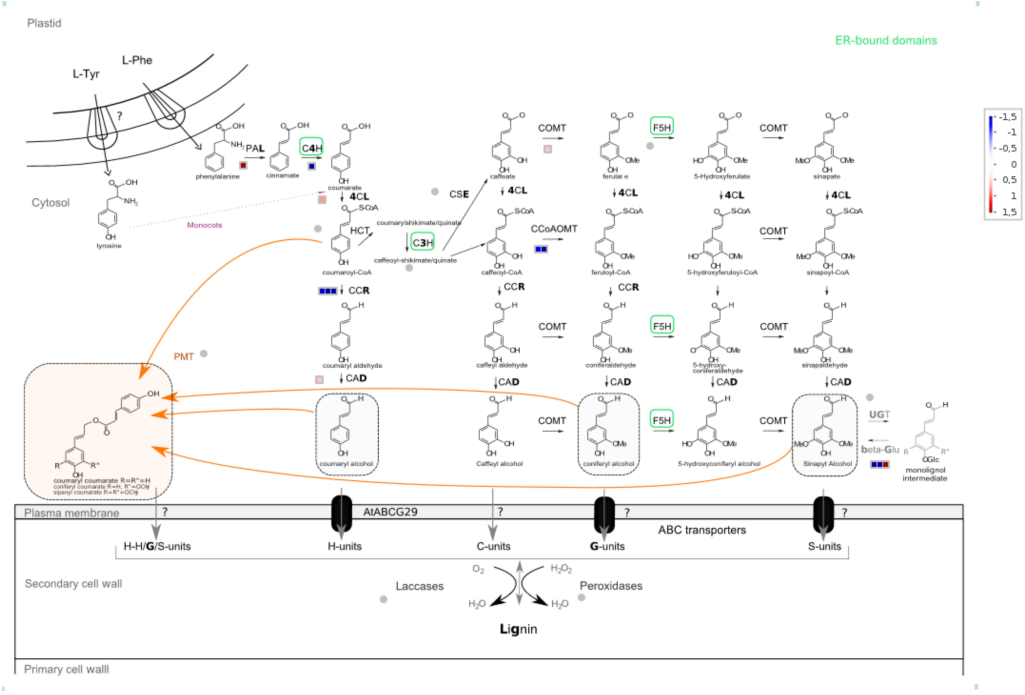

Figure 17: Lignin pathway mapping using the log of fold change of the DEGs from the low biomass genotypes compared to the high biomass group. Genes significantly upregulated were colored in red, while those downregulated were colored in blue.

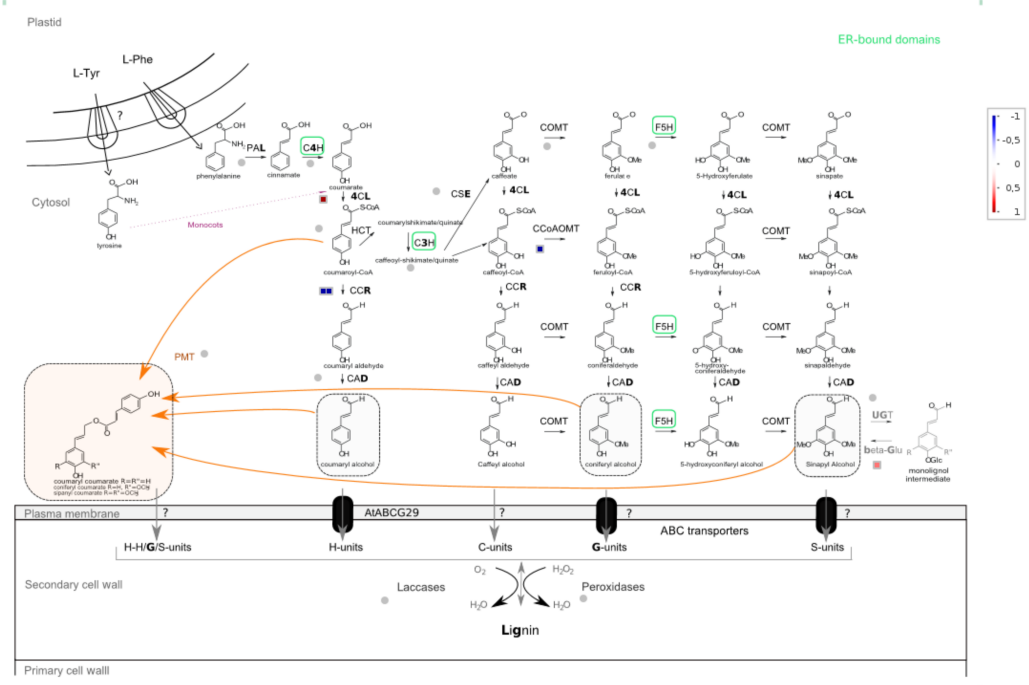

Figure 18: Lignin pathway mapping using the log of fold change of the DEGs from the comparison between US85-1008 and the mean of SES205A and IN84-58. Genes upregulated in US85-1008 were colored in red and those downregulated were colored in blue.

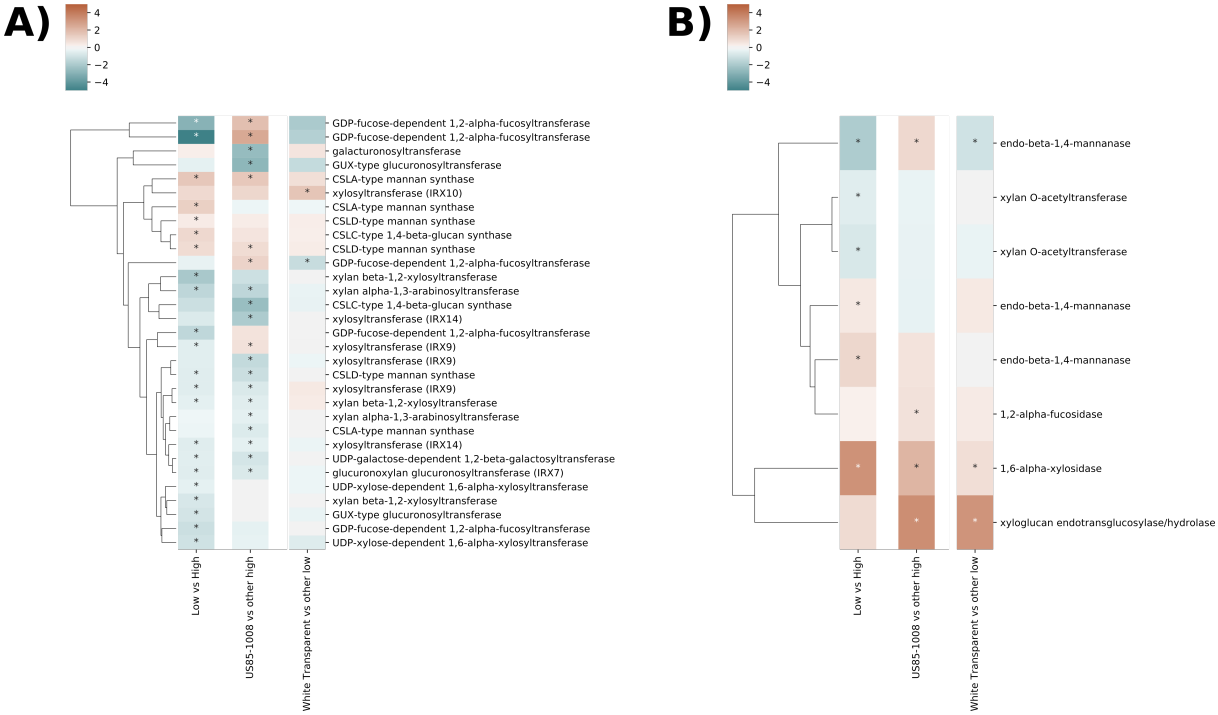

Figure 19: Heatmap for MapMan synthesis (A) and modification and degradation (B) of the cell wall compounds. Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Differentially expressed genes of the contrast are indicated by asterisks.

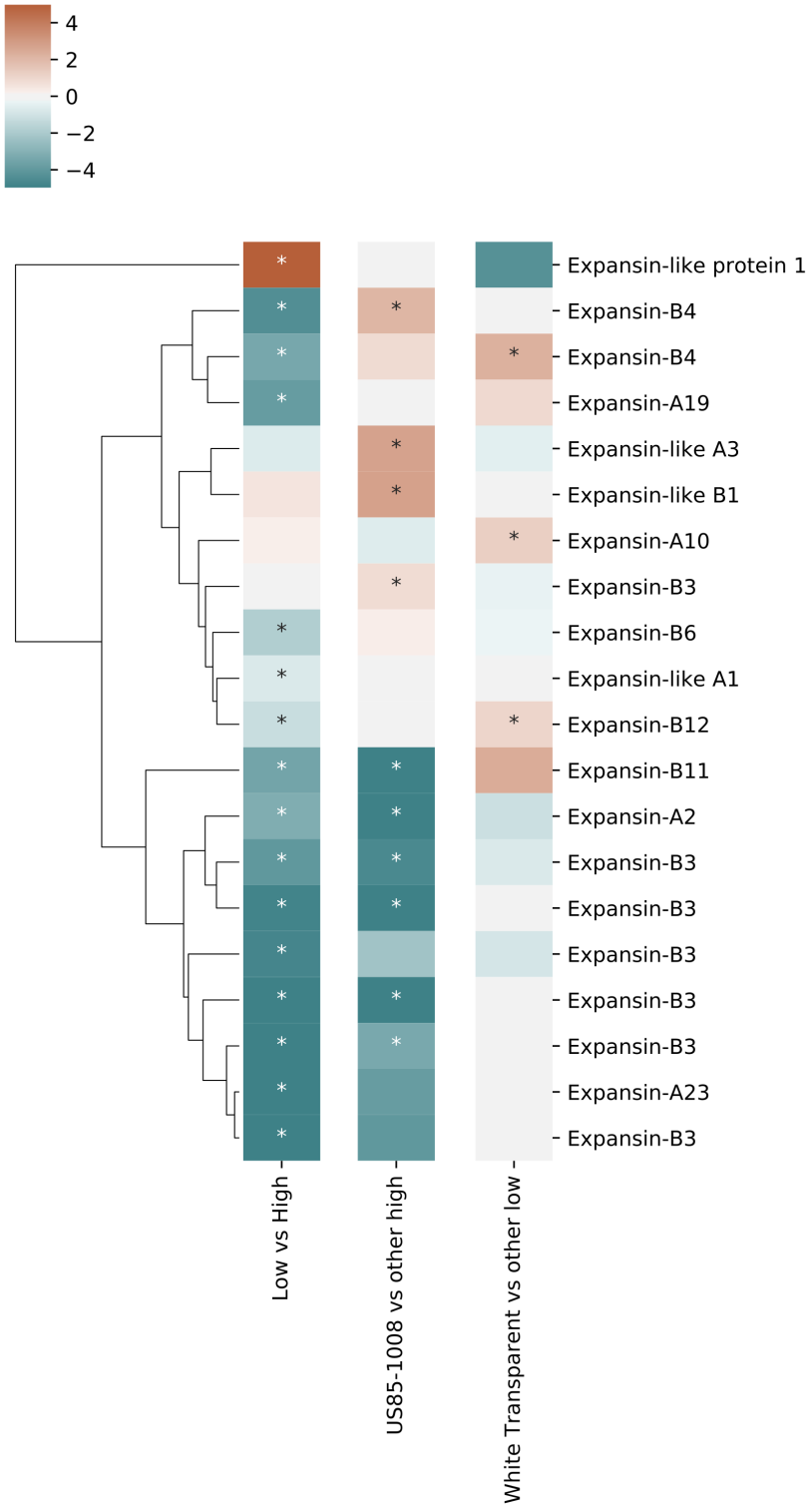

Figure 20: Expression of DEGs coding for expansins. Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Differentially expressed genes of the contrast are indicated by asterisks.

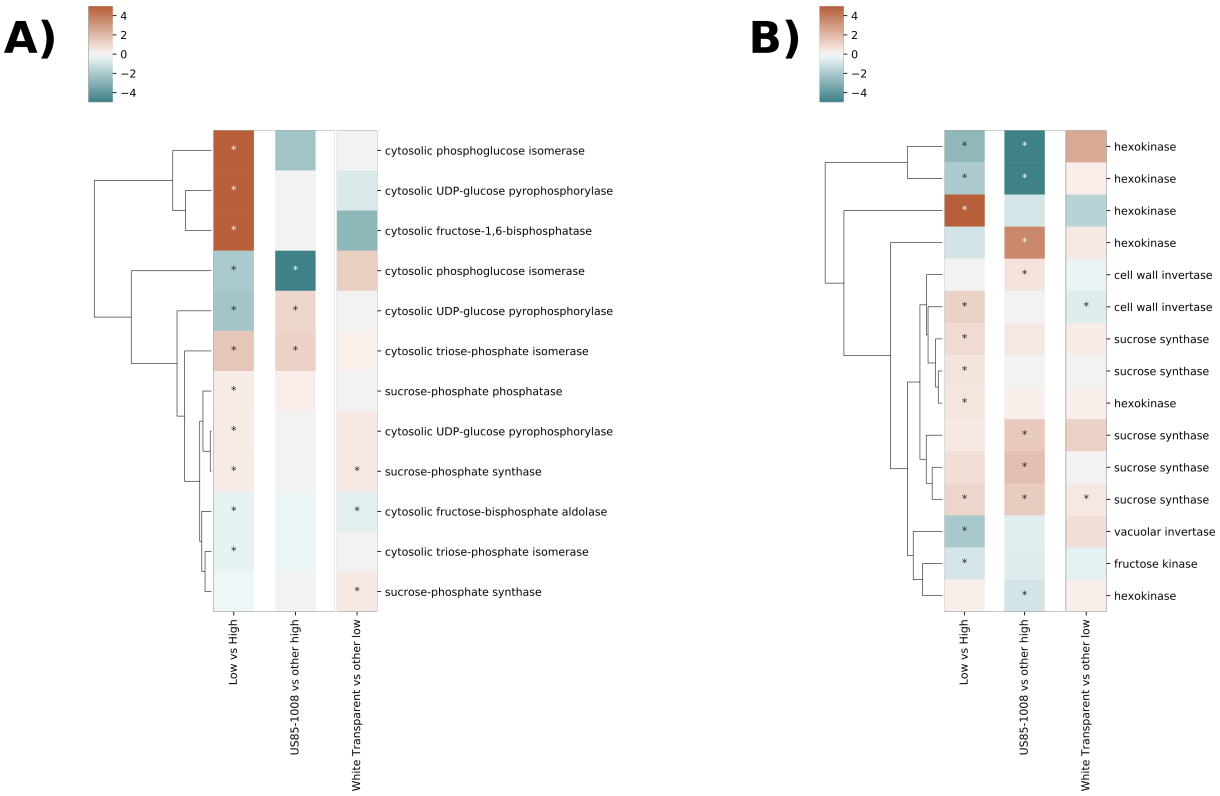

Figure 21: Heatmap for MapMan sucrose metabolism of synthesis (A) and degradation (B). Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Differentially expressed genes of the contrast are indicated by asterisks.

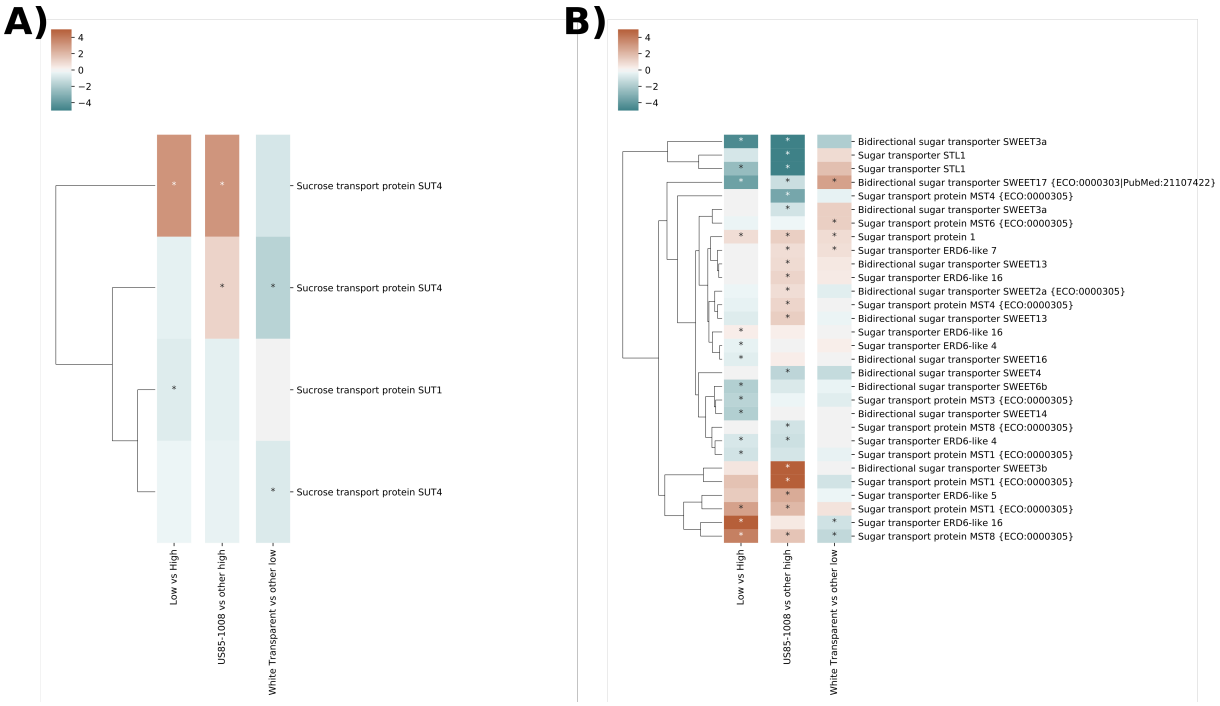

Figure 22: Expression of DEGs coding for sucrose transport proteins (A) and sugar transporters (B). Each column in the heatmap represent the genes fold changes according to a contrast: (i) *Low vs High* is the comparison between the low biomass to the high biomass group; (ii) *US85-1008 vs other high* represents the comparison of US85-1008 to wild high-fiber canes; (iii) *White Transparent vs other low* is the contrast of White Transparent against the low-fiber hybrids. Differentially expressed genes of the contrast are indicated by asterisks.

Expression of investigated genes

In this subsection we present the log of counts per million (logCPM) in the three contrasts for genes investigated by their biological relevance. Expression of four genes encoding sucrose synthase (SuSy) are in Table 4.

Table 4: Expression of sucrose synthase genes. The expression levels, in logCPM, were obtained from the contrast comparing the two phenotypically distinct groups and from the contrast within each group. Asterisk indicates if the gene is differentially expressed in the contrast.

	Low fiber vs High fiber	**US85-1008 vs other high fiber**	***S. officinarum* vs other low fiber**
trinity_dn11006_c0_g1	3.659	3.047*	3.829
trinity_dn11963_c0_g1	6.04*	5.692	6.317
trinity_dn141746_c0_g1	-0.338	-0.355*	-0.473
trinity_dn14183_c0_g1	1.672*	1.262	2.048
trinity_dn931_c0_g1	6.833*	6.412*	7.186*

References

[1] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nature Methods. 2017;14(4):417–419.

Additional file 4 - Supporting information for differentially expressed transcripts

Comparing the expression of genes and transcripts

We checked the similarity of results when the quantification was done for the whole transcriptome, considering each individual isoform, or when we grouped expression levels at gene-level. The number of expressed genes and transcripts can be found on Table 1. In both cases, the number represents roughly 30% of the complete reference used.

	Number passing the expression filter	Percentage of total reference
Genes	47676	0.27
Transcripts	133232	0.30

Table 1: Number of genes and transcripts kept after minimum expression filter for each of the quantification methods.

We also compared if the differentially expressed transcripts (DETs) corresponded to the differentially expressed genes (DEGs) when the analysis was performed grouping counts at the gene level. First, we used the contrast comparing biomass groups. The number of DEGs and DETs can be seen on Table 2. For 15,188 DEGs, at least one of its transcripts was differentially expressed (Figure 1).

	Genes	Transcripts
Down	10903	21996
Not DE	26602	90850
Up	10171	20386

Table 2: Number of downregulated (Down), not differentially expressed (Not DE) and upregulated (Up) genes or transcripts .

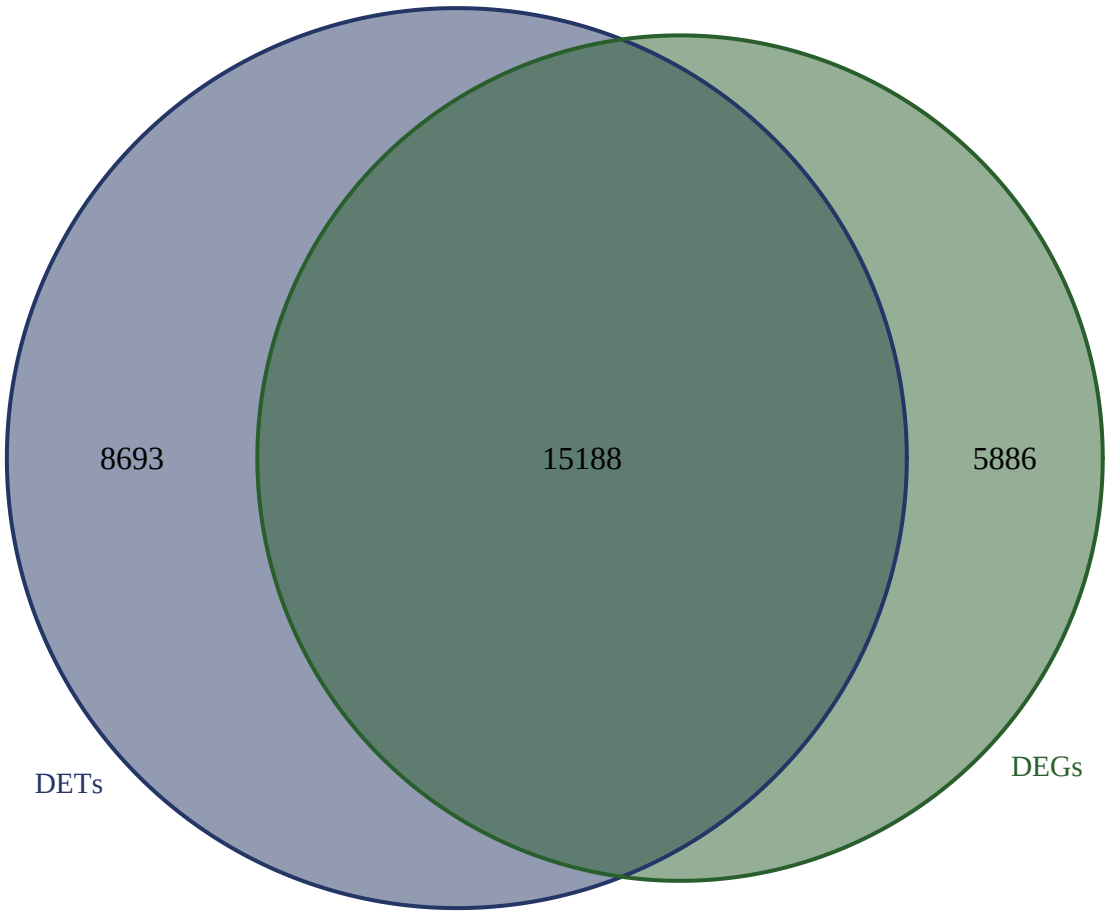

Figure 1: Venn Diagram showing the overlap between the list of differentially expressed genes and differentially expressed transcripts.

We observed that the same six processes enriched among DEGs were also enriched in the DETs (Table 3). Analysis of DETs showed enrichment of terms related to transposition, defense response and enzymatic activities. Enriched terms related to *photosynthesis* (*glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating) activity* and *geranylgeranyl-diphosphate geranylgeranyltransferase activity*) that carrying out differential expression analysis of isoforms revealed more specific processes. In the case of *glyceraldehyde-3-phosphate dehydrogenase* (GADPH), we noted that only two non-DEGs presented DETs (Figure 4).

We focused the search of DEGs and DETs to a group of annotated genes with a common gene ontology term. We chose the *photosyntesis* biological process as an example. Only 18 DEGs had at least one isoform as differentially expressed, 12 genes found as differentially expressed did not have DETs and 47 genes had differentially expressed isoforms but were not differentially expressed when counts were grouped into the gene level (Figure 2). Five photosynthesis DEGs did not have at least one DET (Figure 3).

Category	Description	Genes	DEGs	Transcripts	DETs	Enriched DEGs	Enriched DETs
GO:0002181	cytoplasmic translation	81	58	129	59	Yes	Yes
GO:0003964	RNA-directed DNA polymerase activity	1212	592	1841	712	Yes	Yes
GO:0004190	aspartic-type endopeptidase activity	1045	500	1743	649	Yes	Yes
GO:0004519	endonuclease activity	945	457	1526	557	Yes	Yes
GO:0004523	RNA-DNA hybrid ribonuclease activity	653	326	878	357	Yes	Yes
GO:0015074	DNA integration	986	474	1562	592	Yes	Yes
GO:0000943	retrotransposon nucleocapsid	768	344	1217	423	No	Yes
GO:0003676	nucleic acid binding	1857	834	4312	1334	No	Yes
GO:0003887	DNA-directed DNA polymerase activity	384	179	626	233	No	Yes
GO:0004310	farnesyl-diphosphate farnesyltransferase activity	7	3	48	25	No	Yes
GO:0004365	glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating) activity	16	11	55	30	No	Yes
GO:0004514	nicotinate-nucleotide diphosphorylase (carboxylating) activity	4	1	13	10	No	Yes
GO:0004540	ribonuclease activity	227	101	311	120	No	Yes
GO:0004674	protein serine/threonine kinase activity	1778	753	5026	1535	No	Yes
GO:0004803	transposase activity	108	50	244	110	No	Yes
GO:0005524	ATP binding	5497	2298	16431	4844	No	Yes
GO:0006310	DNA recombination	1090	504	1890	663	No	Yes
GO:0006313	transposition, DNA-mediated	139	73	283	125	No	Yes
GO:0006696	ergosterol biosynthetic process	19	12	67	36	No	Yes
GO:0006952	defense response	1840	765	4157	1352	No	Yes
GO:0008171	O-methyltransferase activity	57	29	164	77	No	Yes
GO:0008270	zinc ion binding	1686	699	4323	1308	No	Yes
GO:0008615	pyridoxine biosynthetic process	7	4	19	14	No	Yes
GO:0008825	cyclopropane-fatty-acyl-phospholipid synthase activity	2	1	10	9	No	Yes
GO:0009443	pyridoxal 5'-phosphate salvage	4	2	20	14	No	Yes
GO:0009870	defense response signaling pathway, resistance gene-dependent	225	94	452	161	No	Yes
GO:0010942	positive regulation of cell death	75	34	152	65	No	Yes

Category	Description	Genes	DEGs	Transcripts	DETs	Enriched DEGs	Enriched DETs
GO:0016767	geranylgeranyl-diphosphate geranylgeranyltransferase activity	6	3	37	22	No	Yes
GO:0016866	intramolecular transferase activity	33	18	88	44	No	Yes
GO:0017148	negative regulation of translation	74	39	200	81	No	Yes
GO:0019438	aromatic compound biosynthetic process	28	14	73	36	No	Yes
GO:0030598	rRNA N-glycosylase activity	41	22	83	52	No	Yes
GO:0032197	transposition, RNA-mediated	342	155	465	185	No	Yes
GO:0032199	reverse transcription involved in RNA-mediated transposition	132	62	197	78	No	Yes
GO:0032201	telomere maintenance via semi-conservative replication	3	2	7	7	No	Yes
GO:0042301	phosphate ion binding	10	8	24	17	No	Yes
GO:0043531	ADP binding	1199	517	2620	925	No	Yes
GO:0043657	host cell	9	6	18	14	No	Yes
GO:0046718	viral entry into host cell	29	17	42	26	No	Yes
GO:0046905	phytoene synthase activity	4	1	35	20	No	Yes
GO:0046983	protein dimerization activity	438	199	1245	444	No	Yes
GO:0051286	cell tip	10	9	12	10	No	Yes
GO:0051996	squalene synthase activity	7	3	48	25	No	Yes
GO:0060548	negative regulation of cell death	179	73	355	130	No	Yes
GO:0070987	error-free translesion synthesis	2	2	7	7	No	Yes
GO:0071768	mycolic acid biosynthetic process	1	0	10	9	No	Yes
GO:0075732	viral penetration into host nucleus	19	12	23	17	No	Yes
GO:0090305	nucleic acid phosphodiester bond hydrolysis	136	64	214	83	No	Yes
GO:0090729	toxin activity	30	13	65	35	No	Yes

Table 3: Gene Ontology terms enriched among DEGs or DETs using the contrast between biomass groups. The number of genes, differentially expressed genes, transcripts and differentially expressed transcripts for each GO term are shown. The last two columns indicate if the term was enriched among DEGs and DETs, respectively.

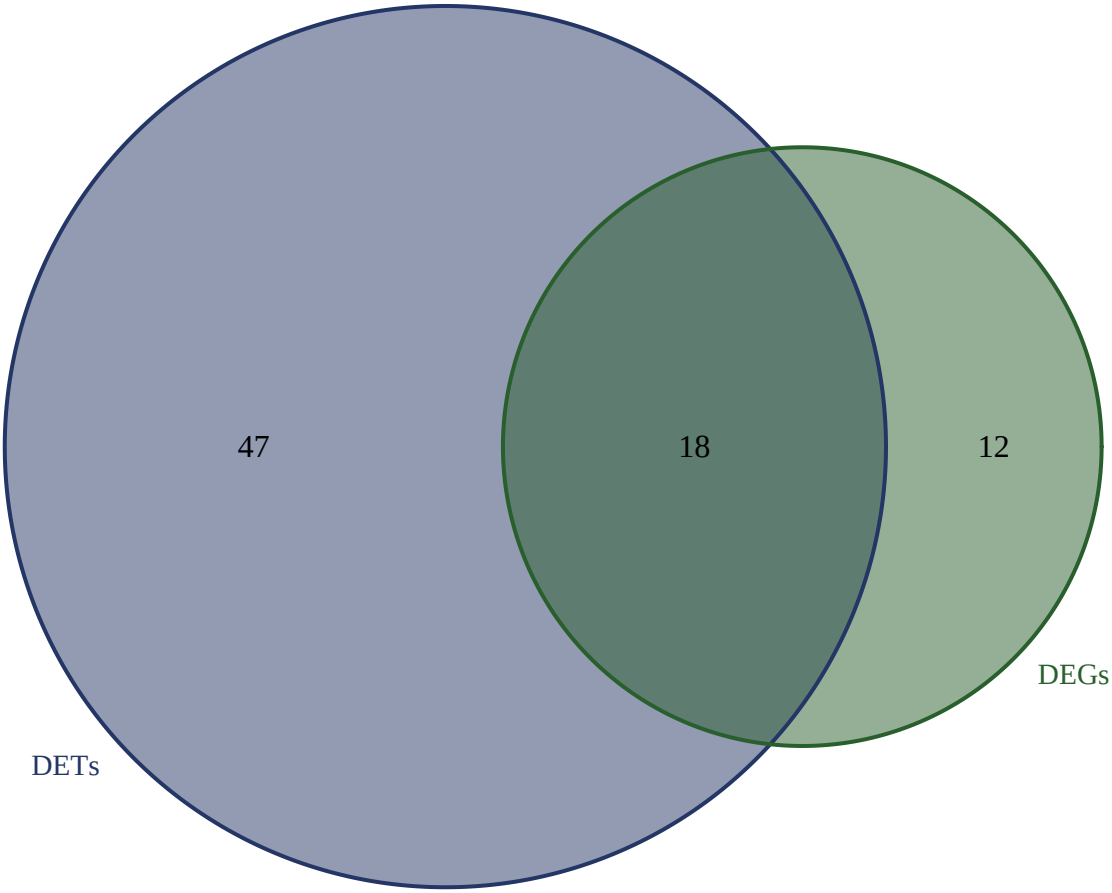

Figure 2: Venn Diagram showing overlap between the DEGs and DETs of the *photosynthesis* biological process.

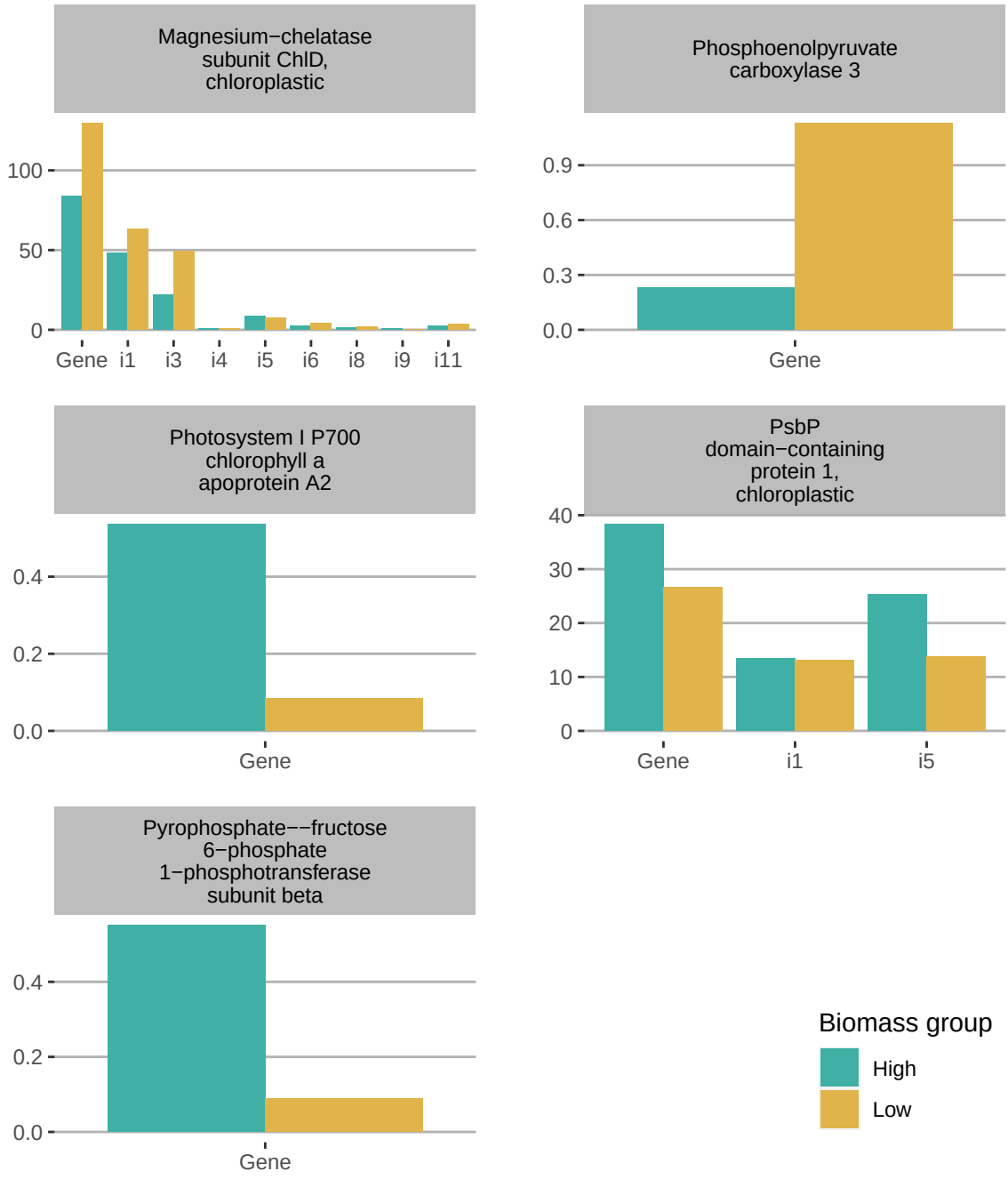

Figure 3: Genes associated to photosynthesis without differentially expressed transcripts. Expression, in counts per million, was measured for each biomass group. In the $x - axis$, transcripts passing the expression filter are shown beside the corresponding gene, indicated with the prefix **i**.

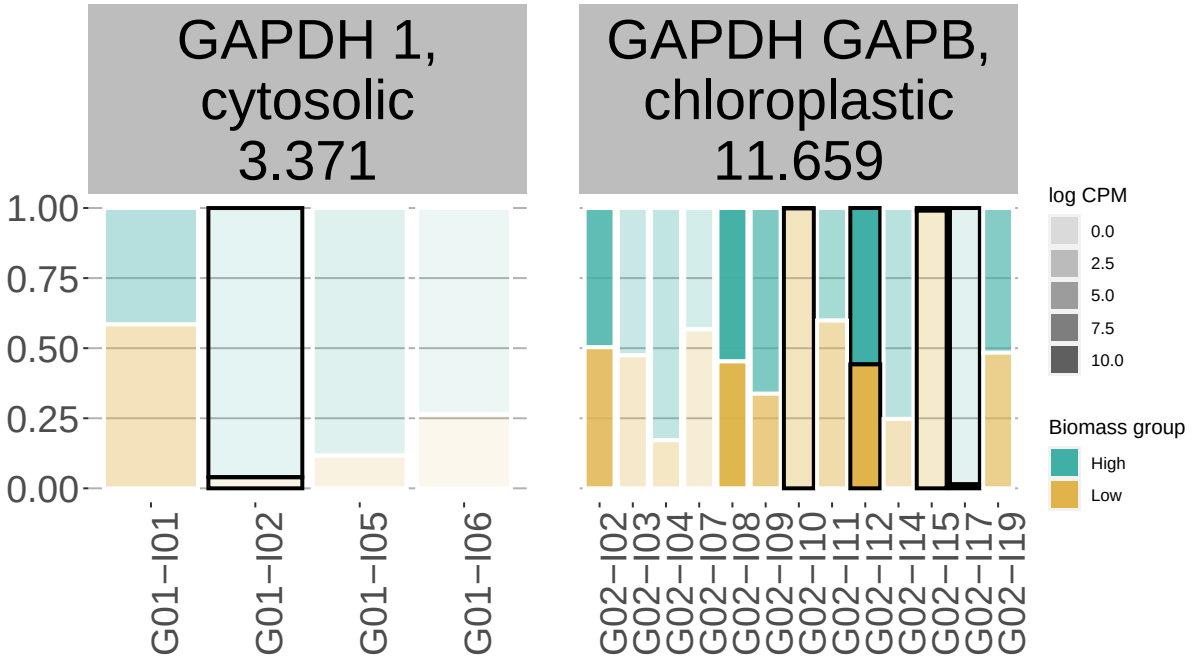

Figure 4: Expression differentially expressed transcripts corresponding to non-differentially expressed *glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating) activity* related genes. The identificator of each gene provides the log of the counts per million of it. For the isoforms, the measure of expression is in counts per million shown as a proportion between the biomass groups, in different colors. The intensity of the color represents the logarithm of the counts per million (log CPM) of the transcript. Differentially expressed transcripts have their edges in black.

3 A HIERARCHICAL BAYESIAN MODEL TO ASSESS ALLELE-SPECIFIC EXPRESSION IN MIXED-PLOIDY SPECIES REVEALS EXPRESSION BIASES IN SUGARCANE

Abstract

Allele-specific expression (ASE) represents differences in the magnitude of expression between alleles of the same gene. Allelic imbalance in diploids occurs if the ratio of expression between both alleles shows deviations from the expected equivalent expression. However, this is not straightforward for polyploids, especially autopolyploids, as knowledge about the dosage of each allele is required for accurate estimation of ASE. This is the case for the genomically complex *Saccharum* species, characterized by high levels of ploidy and aneuploidy. We propose a model to test for allelic imbalance in *Saccharum* that can be easily expanded to other polyploids. As a test case we used genotyping data and RNA-Sequencing libraries from leaves of six sugarcane accessions. A hierarchical Beta-Binomial model was used to test if allele expression followed the expectation based on genomic allele dosage. The doses of the alleles were used in a prior Beta distribution for modeling the proportion of the reference allele from RNA counts. This proportion was then used in a Binomial distribution to model the number of RNA-seq reads showing this allele. We used the Bayesian Markov chain Monte Carlo procedure to draw samples from the a *posteriori* distribution. We defined a polymorphism as showing ASE when the relative genomic dose was outside the highest density interval of the posterior distribution in a certain genotype. Some of the genes evaluated in each accession showed ASE and were related to a broad range of processes, mostly associated with general metabolism, organelles, responses to stress and responses to stimuli. In addition, the frequency of genes with ASE in high-level functional terms was similar among the genotypes. The highest frequencies of ASE occurred in sugarcane hybrids, suggesting some influence of the interspecific hybridization in these genotypes. Although the number of polymorphisms we evaluated was still somewhat limited, this study was the first to assess genome-wide ASE in a high- and mixed-ploidy system using estimated doses of the alleles.

Keywords: Allelic imbalance; Polyploid; Allele dosage; Bayes; *Saccharum*

3.1 Introduction

Sugarcane is one of the most important polyploid crops, and is cultivated in 26.8 million hectares worldwide [14]. Profitable and sustainable production relies on high-yielding cultivars developed by breeding programs [36]. Sugarcane breeders can use molecular markers and genomic sequences to explore the variability among *Saccharum* accessions, and enhance knowledge about the molecular basis of desired traits [23]. However, modern cultivars are complex polyploids, which poses challenges for analyzing their genomes. Although such cultivars have a basic chromosome set of x = 10, they are highly polyploid and aneuploid interspecific hybrids, resulting in a genome of approximately 10 Gbp [11, 32, 33, 36].Association between genotypic and phenotypic data is thus not trivial in sugarcane. Instead of relying only on genomic information, approaches using transcriptomes have proven useful in investigation of likely cellular functions of putative genes, aiming to obtain molecular markers from functional genomic regions. Thus, analyses of transcriptomic data have made it possible to assess gene expression to compare different organs and developmental stages [5, 24] and to contrast specific genotypes [44] or groups of accessions [19, 6].

Differential expression analysis identifies significant changes in the intensity of gene expression, revealing possible changes in metabolic pathways according to contrasting factors used in the experi-

mental design [46, 40]. However, there is also variation inherent to the allelic origin of each transcript, because a heterozygous locus can have more than one haplotype being transcribed. The magnitude of variation between the expression of the haplotypes can differ, resulting in preferentially expressed alleles [4]. Significant differences in the expression of the alleles are due to effects in proximal regulation, changes in the reading frame and epigenetic modifications [4]. Therefore, to measure allele-specific expression (ASE), polymorphisms must be detected and the expression level of each allele be obtained via RNA sequencing [18, 35]. The objective is to detect deviations from equivalent expression between the alleles (*i.e.*, allelic imbalance), as well as to compare the relative allelic proportion in samples from different environments [13, 47, 43].

In diploids, tests for allele-specific expression often use a binomial model with an expected probability of equivalent expression between the alleles (Figure 1 - A). Then, ASE stems from significant deviations from similar expression levels of both alleles. In polyploids, the allelic frequency in the homology group can influence the relative expression levels. Therefore, the doses of the alleles in each heterozygous site must be estimated for accurate assessment of ASE. Pham and colleagues [31] onsidered the possible dosage values in autotetraploid potato - simplex, duplex and triplex [10, 25] - to determine the expected probability of allele counts. This is a case of studying allele-specific expression for an organism with fixed ploidy (Figure 1 - B). They found from 2,180 to 5,270 genes showing preferentially expressed alleles in their experimental conditions - combination of six genotypes and two organs. Furthermore, all potato genotypes had more genes with ASE in the tuber than in the leaves, the former showing enrichment of genes coding for proteins responsible for the localization of macromolecules and transport processes. These authors emphasized that ASE reflected the breeding history of this crop, as it was more frequent in the target of selection - the tuber. On the other hand, they also reported the occurrence of ASE in genes related to traits introgressed from wild genotypes.

Polyploidy arises by whole genome duplications (WGD), originating as autopolyploids; or by hybridization between related species, resulting in allopolyploids [22, 41]. While the former event creates multiple sets of homologous chromosomes, the latter results in parental subgenomes that can be grouped in sets of homoeologous chromosomes [41]. The six *Saccharum* species are polyploids with a large number of chromosomes [32, 50]. Most of the sugarcane cultivars are hybrids between *Saccharum officinarum* and *S. spontaneum*, with variable and genotype-specific numbers of chromosomes [32, 33]. Because both species are considered autopolyploids [50], commercial sugarcane cultivars are interspecific hybrids that can be genomically classified as auto-allopolyploids [51].Recently, sugarcane breeding has focused on the variability from wild accessions to explore traits for bioenergy production [7]. There is an interest in genes associated with important traits in sugarcane breeding - higher biomass production, resistance to diseases and tolerance to adverse environmental conditions. To that end, knowledge about gene regulation can provide useful targets for marker-association studies.

Recent research has addressed the detection of allele-specific expression in sugarcane. After determining haplotypes of particular genomic regions, a variable number of polymorphisms were found within the genes where allele expression was correlated to the dosage [9, 39]. Sforça and colleagues [39] also reported difficulties in observing all the haplotypes of a region, inferring missing haplotypes based on expression data when possible. Another approach used the tetraploid *S. spontaneum* genome [52]to investigate alleles of specific gene families [2].These results show that expression of alleles from genes coding for the Dof transcription family differed depending on the tissues examined, the developmental stages or hormone treatments. They also found that the *cis*-elements of the alleles of the same gene were associated with different functions. These studies pioneered research on allelic expression in sugarcane, but they focused on specific genic regions for a small group of genes. It would be informative to have a global view of the frequency of allele-specific expression in sugarcane, considering the transcriptomes of different *Saccharum* accessions.

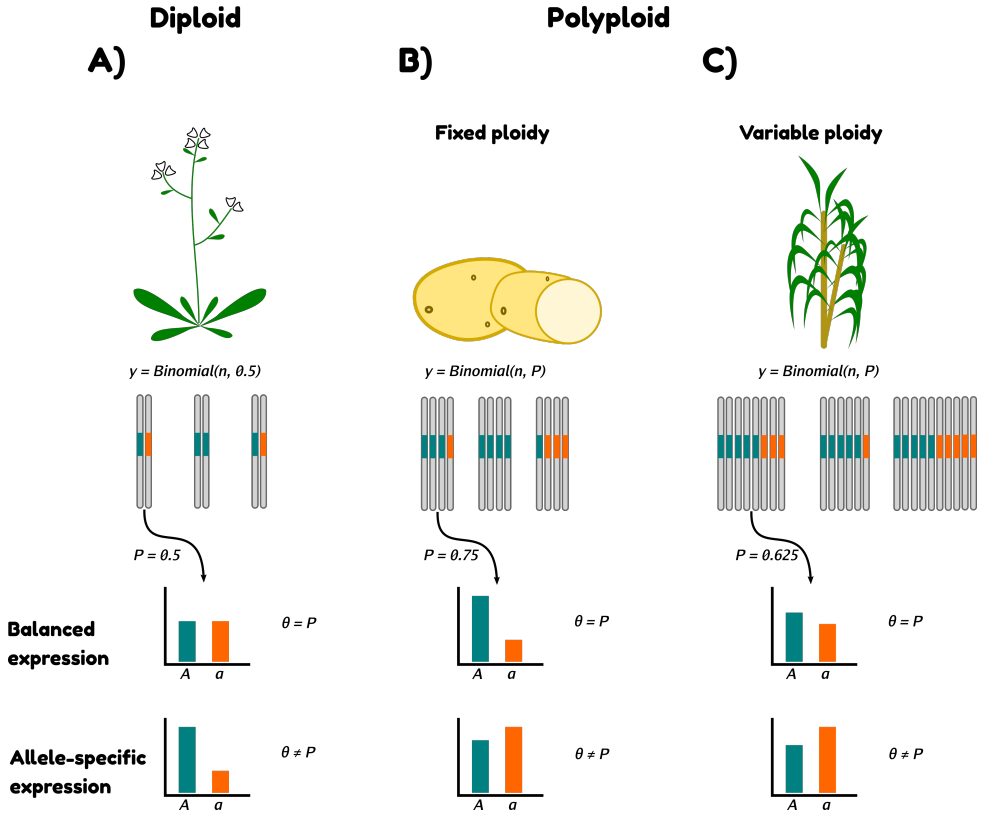

Figure 1: Allele-specific expression studies in different ploidy scenarios. Plants must be properly genotyped to identify homozygous and heterozygous loci. In diploid plant species, we test for allele-specific expression using a known probability of 0.5 of sampling reads with the reference allele (A). For polyploids, we rely on knowledge of the doses of each allele to calculate the proportion P. If the ploidy is fixed - the same in all homology groups -, P changes according to the doses (B). In polyploids with a variable number of homologous chromosomes per group, we need to properly estimate the ploidy of each group and use the allele doses to calculate P (C). If the proportion of the reference allele from the RNA-Sequencing (θ) is significantly different from P, the gene is said to show allele-specific expression.

Pham and colleagues [31] used a fixed ploidy of four homologs per group to detect SNPs with ASE in tetraploid potato (Figure 1 - B). However, in a crop such as sugarcane, the ASE models must deal with variable ploidy levels [15], respecting cytological results that demonstrate homoeology and aneuploidy (Figure 1 - C) [32, 45]. Nowadays, it is feasible to assess allele-specific expression in sugarcane by combining the expression data from RNA-Sequencing studies with the allelic dosages estimated through an appropriate pipeline for an organism with non-fixed ploidy [30, 38]. Our main objective was to test for allele-specific expression using a model leveraging the doses of the alleles as prior information. Here we show the use of a Beta-Binomial distribution to model ASE in *Saccharum*. Finally, we suggest that this model can be easily applied to unravel ASE in other complex polyploid species.

3.2 Material and Methods

3.2.1 Biological material, SNP calling pipeline and quantification of allele expression

Genotypic and transcriptomic information of *Saccharum* genotypes was used to investigate the expression of different alleles. To characterize expression profiles, a set of six genotypes was selected from a previous gene expression study [6] - IN84-58, RB72454, SES205A, SP80-3280, US85-1008 and White Transparent. These genotypes represent two groups of accessions contrasting in key biomass traits - fiber content and tillering capacity. Genotypes of the high biomass group include the hybrid US85-1008 and

the *S. spontaneum* genotypes IN84-58 and SES205. The low biomass group included the *S. officinarum* White Transparent and the hybrids RB72454 and SP80-3280. Briefly, we collected portions of the first visible dewlap leaves (+1) from six-month-old sugarcane plants and extracted the total RNA from the middle section of each leaf. Pooled libraries were sequenced in two lanes of an Illumina HiSeq 2500 platform, in paired-end mode (2 × 100 bp). Information regarding those genotypes can be found in the supplementary material (Table 1 in Additional File 1). Herein we used as a reference the longest isoforms of a transcriptome assembled *de novo* using the RNA-Seq reads of the full set of genotypes [6].

A panel of 245 *Saccharum* accessions forms the Brazilian Panel of Sugarcane Genotypes (BPSG), which is composed of wild accessions and hybrids from Brazilian and foreign breeding programs [26]. These accessions were genotyped using the genotype-by-sequencing (GBS) protocol [12] with the PstI restriction enzyme. Library preparation was planned to provide a higher sequencing depth for some genotypes, by including duplicate samples in multiple library plates, including White Transparent, IN8458, RB72454 and, in particular, SP80-3280. A pipeline for SNP discovery in polyploids was performed with TASSEL4-POLY [30], using BOWTIE2 [21] to align the GBS reads. First, for SNP discovery we used the standard Tassel4-Poly pipeline with the following main modifications: a minimum minor allele frequency of 0.01 (*mnMAF*) and a minimum minor allele count (*mnMAC*) of 40. Next, the ploidy and allelic dosages for each site were estimated with SUPERMASSA [38] and VCF2SM [30]. We used the Hardy-Weinberg inference model with a minimum call rate of 50%, a naïve posterior threshold of 0.5 and a minimum posterior probability to keep a variant of 0.5. Ploidy levels ranging from four to 16 were tested, then filtered for polymorphic sites with the most likely ploidy being between six and 14. The SNP calling process took into account all genotypes from the BPSG, but only those present in our RNA-Seq data were kept for downstream analysis. The VCF file was filtered to remove sites where the genotypes were homozygous or had missing calls, as well as those identified as insertions or deletions.

HISAT2 [20] was used to align the RNA-Seq reads to the *de novo* transcriptome. Quantification of read counts of each allele was performed with the GATK ASEREADCOUNTER tool [4, 10, 25] for each aligned library. Counts of the reference allele and the total counts for each SNP for each genotype were scored. Reads from both lanes of the same sample were grouped, as no batch effect was identified. Sites with at least ten RNA-Seq reads were retained and positions showing low expression were removed.

3.2.2 Model to test for allele-specific expression in *Saccharum*

To assess the occurrence of allelic imbalance in a given SNP, we tested if the expression of the reference allele was equal to its relative dosage in the genome, given the estimated ploidy. For the i-th SNP of genotype k, α_{ik} and β_{ik} were the dosage of the reference and the alternative alleles, respectively (Figure 2 - Genotyping). First, the genomic ratio was calculated as the dosage of the reference allele divided by the corresponding ploidy level ($P_{ik} = \frac{\alpha_{ik}}{\alpha_{ik}+\beta_{ik}}$). Next, the proportion of the reference allele was estimated from the RNA-Seq count, denoted by θ_{ik}. Then, we tested the null hypothesis of no significant difference between these two ratios:

$$H_0 : \theta_{ik} = P_{ik}$$

A model following the Beta-Binomial distribution was proposed to test this hypothesis (Figure 2 - I, II and III). First, we modeled the number of RNA-Seq reads of the reference allele of the i-th SNP, on the r-th replicate of the k-th genotype - y_{irk}, following a Binomial distribution (Figure 2 - II):

$$y_{irk} \sim Binomial(n_{irk}, \theta_{ik}),$$

where n_{irk} represents the total number of reads of the SNP for the corresponding sample. The prior distribution of the parameter θ_{ik} was modeled by a Beta distribution (Figure 2 - I), using as parameters the dosages of the alleles:

$$\theta_{ik} \sim Beta(\alpha_{ik}, \beta_{ik})$$

The posterior distribution of θ_{ik} was sampled via a Markov chain Monte Carlo (MCMC) procedure, using the Bayesian framework of STAN [3] in the RSTAN package [42]. Four chains with 10,000 iterations each and a burn-in of 1,000 iterations were used. Convergence of the model was assessed by investigation of the effective sample size, autocorrelation between the chain draws, the scale reduction factor and visual inspections of traceplots. We deemed a SNP as showing a preferentially expressed allele if the ratio P_{ik} was outside of the highest density interval (HDI) of θ_{ik} (Figure 2 - III). A gene with at least one SNP with allelic-specific expression was called as having ASE (ASEG).

3.2.3 Enrichment analysis

Gene Ontology (GO) terms was evaluated for enrichment with ASEGs. To that end we used the ASEGs as the set of selected genes, compared against the background of all the genes with at least one heterozygous SNP. Tests were performed with the GOSEQ R package [49]. Terms with an FDR-adjusted p-value less than 5% [1] were considered overrepresented.

3.3 Results

3.3.1 Number of polymorphisms obtained with the polyploid genotyping pipeline

A total of 63,712 polymorphic sites were identified in the *de novo* transcriptome reference for the 245 genotypes of the panel. We kept 37,902 SNPs after removing monomorphic or missing sites in the six genotypes of the RNA-Seq dataset. By doing so, we only kept SNPs that were heterozygous in at least one of the genotypes. We also removed polymorphisms identified as indels, keeping 27,041 sites. Most of the SNPs sequenced at higher depth were dodecaploid, for all genotypes, with lower frequencies for lower ploidy levels (Figure 1 in Additional file 1).This finding is in agreement with cytological observations, as twelve is the most frequent ploidy among the homologs of *Saccharum* hybrids [33]. Less stringent filters resulted in different distributions, with higher frequencies of hexaploid and octaploid loci. This may reflect lower accuracy for polymorphisms detected at lower depth of sequencing.

Another important observation was that the total number of SNPs was almost constant among the genotypes when no depth filter was applied (Figure 1 in Additional file 1). However, the number of heterozygous SNPs was higher in hybrids and *S. officinarum* (Table 2 in Additional file 1).When increasing the minimum depth filter, the genotypes SES205A and US85-1008 had fewer SNPs than the others. During the GBS protocol, these were the only genotypes without replication in the sequencing libraries. Furthermore, 75% of the transcripts had up to 2,665 bp, with an average of roughly four SNPs (Figure 2 in Additional file 1). It also observed that longer transcripts did not necessarily have more SNPs. This is likely explained by the inherent limitation of GBS to only detect SNPs in positions adjacent to the restriction enzyme recognition site. Overall, these figures show that the markers identified with the GBS pipeline are appropriate for genotyping and comparing different accessions.

After removing indels, missing and monomorphic sites, quantification of allele expression was performed with ASEREADCOUNTER for 26,995 SNPs identified in 6,722 transcripts. We used the heterozygous sites in each genotype (numbers in Table 2 in Additional file 1) to test if the RNA-Seq proportion between both alleles deviated from the ratio observed in the GBS reads, indicating a likely imbalance between the alleles.

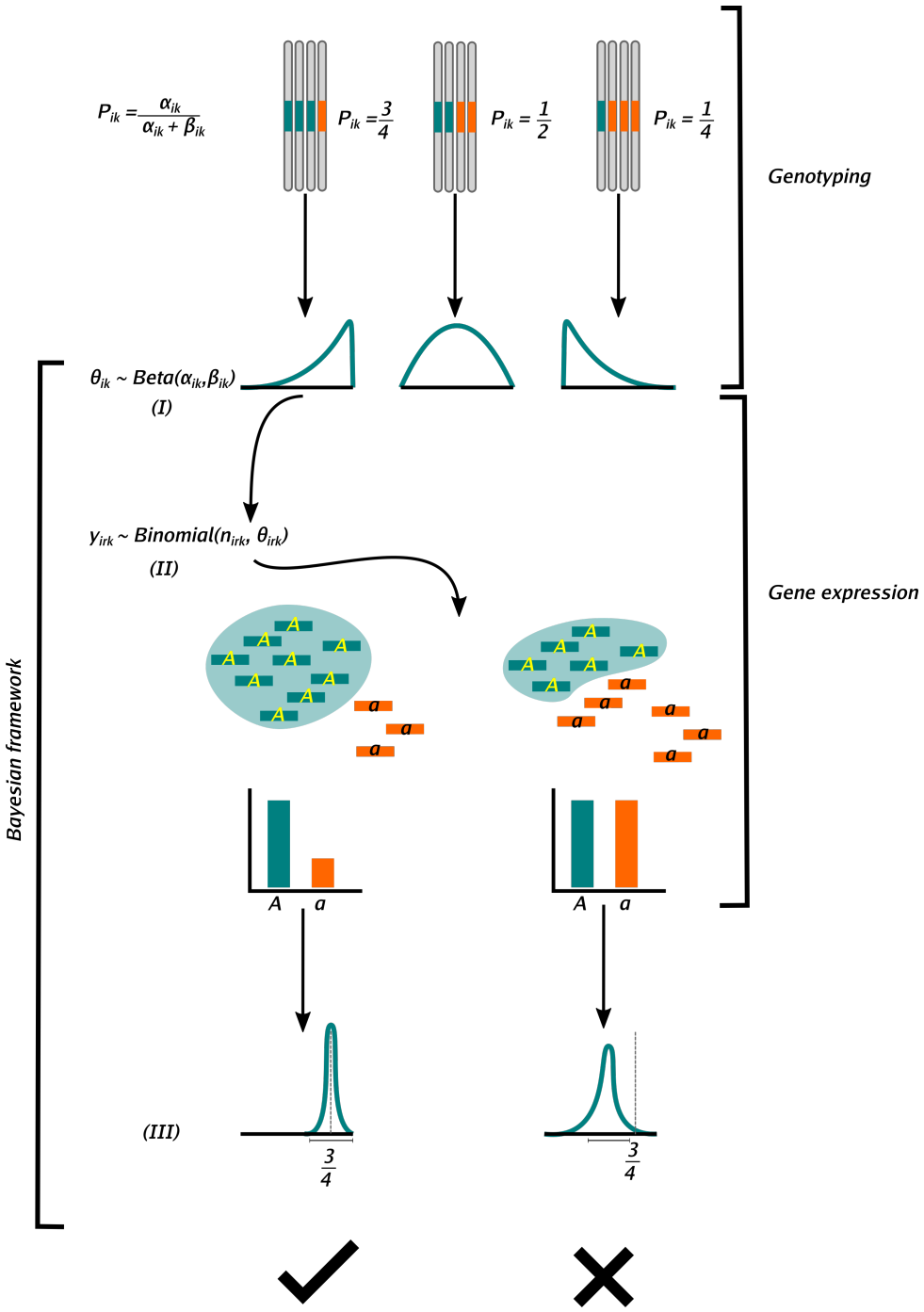

Figure 2: Schematic view of the allele-specific expression modeling. The genotyping section of the figure shows examples of different relative doses of the reference allele, P_{ik}. The reference allele is colored in blue, while the alternative allele is shown in orange. In the Bayesian framework the allele doses were used as shape parameters of a beta distribution (I), which was used as the prior distribution of θ_{ik}. From the first example of prior, we show two possible scenarios of posterior distributions. The prior was conjugated to the Binomial likelihood, which was used to model the counts associated to the reference allele (blue bars) from the total counts generated by RNA-Sequencing (II). Last, we show the posterior distribution (III), from which we tested for allele-specific expression. The check mark indicates an example of a gene with balanced expression, while the X represents a case of a gene with imbalanced expression.

3.3.2 Preferentially expressed alleles

The SNPs used to test for preferential expression were the heterozygous loci with a minimum of ten genomic reads. For all genotypes, genes showing ASE were the majority (Figure 3 and Table 3 in Additional file 1). No evidence of positional bias of the SNPs showing ASE was found (Figure 3 in Additional file 1) and we also found no evidence that SNPs in highly expressed genes were more likely to show ASE (Figure 4 in Additional file 1). Dissimilarity among genotypes calculated with ASE-SNPs was similar to that obtained with all loci. First, using either the relative dosage estimated with the genotypic data or the relative expression calculated from the RNA-Seq, hybrid genotypes were clustered with *S. officinarum* (Figure 5 - A and 5 - C in Additional file 1). A second cluster was formed by the *S. spontaneum* accessions. These groups were also consistent when using only SNPs classified as showing ASE (Figure 5 - B and 5 - D in Additional file 1), revealing that the occurrence of ASE may be used to estimate distances between accessions.

The three hybrid genotypes - RB72454, SP80-3280 and US85-1008 - had the highest number of ASE-SNPs and also the highest number of ASEGs (Figure 3). We noticed that most genes with ASE occurred exclusively in a single genotype after evaluating all possible intersections of ASEGs. However, results regarding the functional annotation were similar among genotypes (Table 4, 5, 6 and 7 in Additional file 2). We found ASEGs coding for stress-related proteins, especially disease resistance proteins. Among the disease resistance gene analogs (RGAs) and RPP genes, we found an ASEG coding for the protein *ENHANCED DISEASE RESISTANCE 2*. In this gene, four SNPs revealed allele-specific expression, and from the three common ASE-SNPs between SP80-3280 and RB72454, two showed higher expression of the alternative allele (Figure 6 in Additional file 1). The protein coded by this gene is potentially involved with hypersensitive response and in preventing senescence induced by ethylene. Curiously, a gene coding for a *probable ethylene response sensor 2* was among the ASEGs for the same intersection (Table 5 in Additional file 2).

Sucrose content has traditionally been the focus of sugarcane breeding programs and, more recently, there has been an increasing interest in developing high fiber cultivars. Hence, ASEGs related to carbohydrate metabolism were investigated. However, a clear pattern for genes involved with carbohydrate partitioning was not identified, even in genotypes in the same phenotypic group - high or low biomass. On the other hand, genes related to this biological process were classified as ASEGs in individual genotypes. For example, we detected a gene coding for *UTP–glucose-1-phosphate uridylyltransferase*, an enzyme involved in the synthesis of UDP-glucose, was detected. This gene had preferentially expressed alleles in all low fiber genotypes and in two high fiber accessions - IN84-58 and US85-1008 (Table 4 and 5 in Additional file 2 and Figure 7 in Additional file 1). Interestingly, the genes coding for *Sucrose-phosphate synthase* and *Sucrose transport protein SUT4*, proteins respectively involved with sucrose synthesis and transport, showed significant ASE in IN84-58. Similarly to carbohydrate metabolism, we could not find evidence of any association between photosynthesis-related ASEGs and the two phenotypic groups. Moreover, we identified genes in these processes for which all the accessions had biased expression towards the same allele. In the case of the gene coding for *RuBisCO large subunit-binding protein*, SNPs of the six genotypes showed preferential expression of the reference allele (Figure 8 in Additional file 1). The same was true for almost all ASE-SNPs found in the *phosphoenolpyruvate carboxylase 3* coding gene (Figure 9 in Additional file 1).

Functional enrichment tests were used to evaluate if ASEGs were acting on similar processes. No GO term was significantly enriched with ASEGs. This result is possibly explained by the limited number of genes with detected polymorphisms that passed the filtering steps - roughly one thousand per genotype. We then checked the frequency of ASEGs in each GO Term (Table 6 and 7 in Additional file 2) and found that GO terms with the highest frequencies of ASEGs were often found in common

for all genotypes. As expected, high-level GO terms had the most ASEGs, followed by many metabolic processes and terms associated to the biosynthesis of cellular compounds. This shows that many ASEGs are possibly directly involved with maintaining the metabolism. For the GO terms, the frequency of ASEGs shared among genotypes (Table 6 in Additional file 2) is higher than the frequencies of ASEGs found in individual genotypes (Table 7 in Additional file 2). This indicates that high-level GO terms have many genotype-specific ASEGs. Hence, genes with allele-specific expression were found seemingly at random in the same pathway when considering different genotypes.

3.4 Discussion

Silva [7] has shown that dosage-effects and gene duplication are key factors contributing to variations in gene expression levels in sugarcane. Allele-specific expression adds another layer to the complexity of interpreting gene expression in both autopolyploids and allopolyploids. For sugarcane, this phenomenon was investigated in genes with known functions [2, 9]. As we evaluated a larger set of expressed genes, the ASEGs found in our study were associated with a wide range of functional roles, mostly with high level metabolic processes. We found no differences among hybrids and wild genotypes regarding ASEGs related to the biosynthesis, modification or degradation of particular compounds. Indeed, we observed that most ASEGs were exclusive to an individual accession (Figure 3) rather than to groups of genotypes, and that the number of ASEGs in high level GO terms was similar among the accessions. The lack of co-occurrence of ASEGs in specific pathways can be explained by two concurrent hypotheses. First, that allele-specific expression in sugarcane is genotype-specific, occurring for different genes in high level pathways. Second, there are ASEGs shared among a few genotypes that can be associated with particular functional roles. The second hypothesis could explain the few ASEGs in more specific terms (Table 8 in Additional file 2), such as the defense gene with higher expression of the alternative alleles in hybrids (Figure 6 in Additional file 1).

Previous efforts unraveled allele-specific expression in groups of sugarcane genes. Vilela and colleagues [9]found most of the SNPs in the TOR coding gene with the expression of different alleles matching the corresponding doses of the haplotypes. For the *Phytochrome C* coding gene, however, they identified allele-specific expression towards the main haplotype. In another endeavor, Sforça and colleagues compared the expression proportion to the genomic proportion of SNPs found in haplotypes of the genes HP600 and CENP-C [39]. Both genes had SNPs showing significant differences between the genomic and the transcriptomic proportions of the haplotypes. Allele expression has also been studied in combining the *S. spontaneum* genome [52] and transcriptomic datasets. Recently, Cai and colleagues [2] used the upstream region of Dof transcription factors and found *cis*-elements associated with different functions in plants. Furthermore, these authors identified differences in upstream regions of the alleles of the same gene coding for a transcription factor. They also found alleles showing specific expression depending on tissue, developmental stage and different hormone treatments.

These studies focused on specific haplotypes of a few genes [9, 39] or evaluated specific gene families [2]. To achieve a global view of allele-specific expression in sugarcane, we took a *de novo* transcriptome as a reference and estimated the allele dosage based on SNPs identified from GBS data. Estimating the doses is common for genotyping polyploids [8, 15, 16, 39]. To test for allele-specific expression in sugarcane leaves, we hypothesized that the expression of the alleles followed the allelic dosages. Our results showed that more than half of the evaluated genes had at least one SNP showing ASE (Figure 3 and Table 3 in Additional file 1). We did not verify any bias associated with ASE-SNPs (Figures 3 and 4 in Additional file 1),which also correctly clustered genotypes (Figure 5 in Additional file 1). Thus, neither a restricted coverage of polymorphisms nor differences in multiplexing apparently hampered the detection of SNPs with ASE. Moreover, our results indicate that interspecific hybridization may have

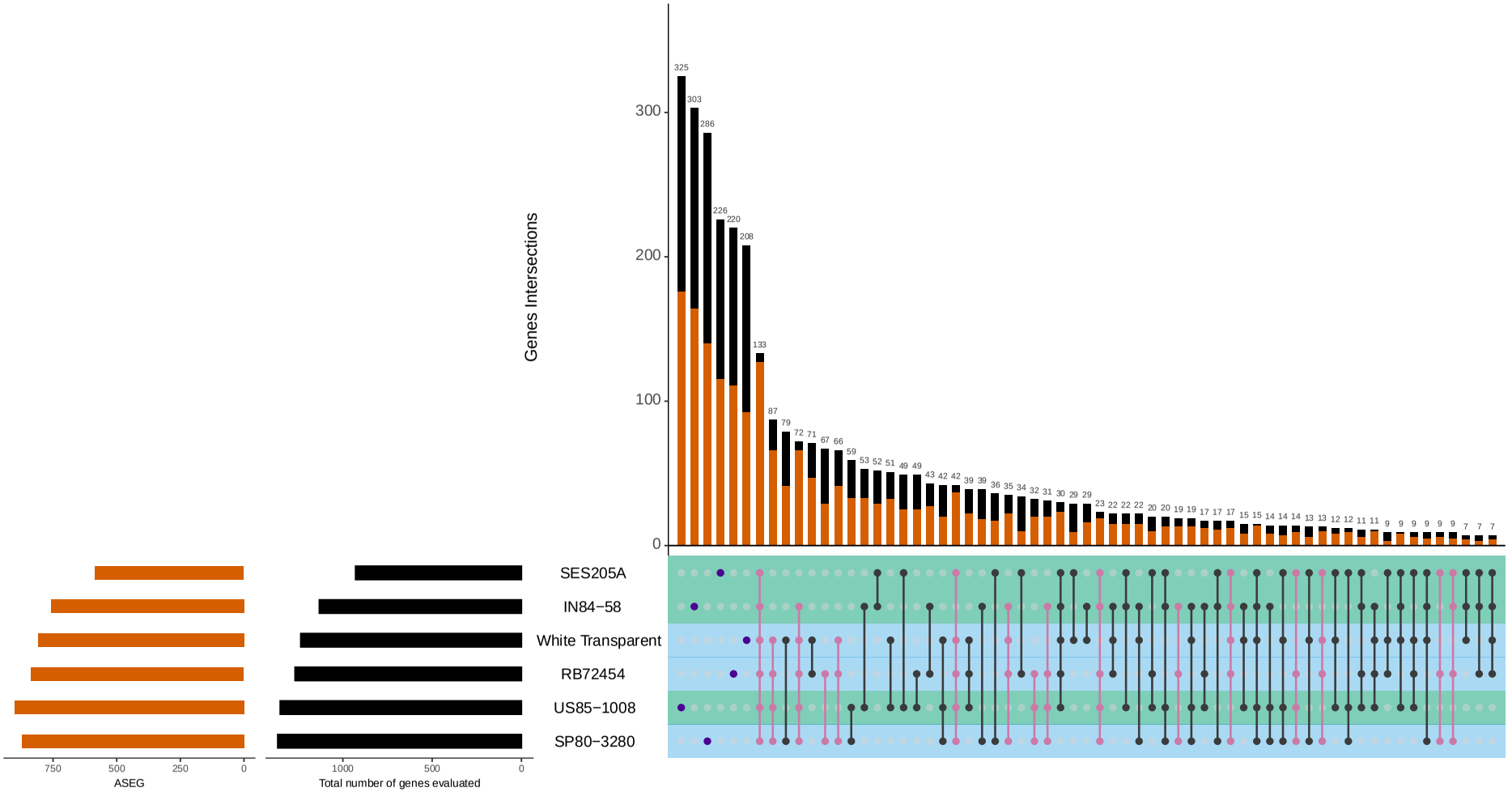

Figure 3: Intersections among the genes showing allele-specific expression (ASEGs) detected for each genotype. The number of ASEGs in each genotype is shown in orange and the total of number genes (ASEGs and non-ASEGs) is in black, on the left part of the plot. The right plot indicates all possible intersections among the genotypes, with ASEGs and non-ASEGs colored with the same scheme in the barplot. High fiber genotypes are shown with a green bar, and low fiber genotypes are in blue. Purple dots indicate the exclusive genes of each genotype, pink dots represent the intersections where SP80-3280 and RB72454 are present. The remaining intersections are colored in black.

caused changes in allele expression, as the highest numbers of ASEGs were found in RB72454, SP80-3280 and US85-1008. Nevertheless, we note that this observation should be interpreted with caution, as the sampling of polymorphisms with GBS is limited and subject to biases [28].

Variation in the expression levels due to allele-dosage effects can be expected in polyploids, and this could lead to variable phenotypic effects [29]. Appropriate knowledge of allele dosage in polyploid organisms is required to test for allele-specific expression. ASE tests used in diploids are often based on a Binomial distribution using the null hypothesis that both alleles are expressed equally ($\theta = 0.5$) [4, 13]. Genotyping of organisms with a fixed ploidy is feasible [10, 25, 31], with markers possibly having different allelic dosages. Unfortunately, this test is not suitable for organisms with variable ploidy levels, as loci can show multiple categories of heterozytes for each ploidy level. In this scenario, cytological observations on sugarcane reveal different ploidies in the homoeologous groups [33], expanding the categories of allelic dosages [15].

Knowledge about the complete haplotypes of the homologs/homoeologs from genomic data can improve ploidy estimation [39]. Using SNPs, we restricted our analysis to two alleles, although in many loci the number of alleles is probably higher. For identifying multiple alleles, we should use a haplotype-based approach, which requires a large marker density or longer sequencing reads [27, 37]. However, determining the ploidy of the genomic regions for a large number of loci is still nontrivial for complex polyploids. In this scenario, the best alternative still relies on estimating the doses of alleles using molecular markers [8, 15, 16]. WWith this approach, knowledge of the doses has been used for constructing genetic maps and improving the performance of predictive models [8, 16]. In addition, this information can be used to test for allele-specific expression as done for species with fixed ploidy [31]. For those with variable ploidy levels, models should account for the dosages of alleles in each marker. This is the scenario for our *Saccharum* dataset, in which we aimed to estimate the posterior distribution of the proportion of the reference allele. We used a Bayesian hierarchical model considering the estimated doses - obtained through genotyping - as parameters of a prior Beta distribution. Because the counts of the allele - from the expression data - follow a Binomial distribution, we modeled allele-specific expression for polyploids with a Beta-Binomial distribution (Figure 2).

Sugarcane cultivars, which are interspecific hybrids, can also show different regulation of alleles coming from different homoeologs. Unfortunately, we currently do not have enough information to identify the homoeologs but only the polymorphisms. A limitation would arise if non-identifiable duplicated genes are treated as single-copy, potentially biasing read mapping [39]. Lastly, as stated by Vilela and colleagues [9], we can only speculate which mechanisms are responsible for biased expression, including the regulation of promoter regions or epigenetic changes [4]. For a deeper investigation of the causes of ASE, multiple omics approaches should be integrated. Through genomics, assessment of the upstream and downstream regions can reveal polymorphisms in *cis*-elements. These regions can be also investigated for epigenetic modifications affecting gene regulation. In any case, by combining genomic and transcriptomic data we can identify ASEGs independently from the underlying causes.

Testing for allele-specific expression is relevant to understand differences in tissues, conditions or genotypes. Previous studies in plants emphasize how this phenomenon is common among the expressed genes. Allelic-specific expression was found in more than 50% of the genes in the maize ear of a hybrid cultivar, with a similar number of ASEGs found independently of the developmental stage [18]. They also found a higher contribution of the alleles from one parent, but this was less pronounced during floret differentiation. Ereful and colleagues [13] studied allelic imbalance combining rice genotypes (parents and F1 hybrids) and drought conditions (plants under normal water regime or following a dry-down protocol). They suggested that the occurrence of ASE was more associated to the genotype than due to water stress. However, depending on the crop, more ASEGs can be found in specific tissues. Pham and colleagues [31] found evidence that allele-specific expression is more frequent in potato tubers than in leaves, probably

due to the selection for carbohydrate accumulation in the tubers.

Allopolyploids - or even a diploid interspecific hybrid - can be compared to their diploid parents to verify the occurrence of expression level dominance and also homoeolog-specific expression [48]. The study of alleles in allopolyploids relies on assessing the expression of the homoeologs to test for both expression level dominance and homoeolog-specific expression. This depends on previous knowledge of the allopolyploid parents, as the aim is to verify possible biases in gene expression towards a subgenome [17]. The analysis is similar to that performed in the maize hybrid by Hu and colleagues [18] to determine the parental alleles with biased expression. In cotton, transgressive expression and expression level dominance of the A or D genome, together, were more frequent than additive expression [48]. However, homoeolog-specific expression was balanced between the subgenomes, which was partially explained by differential regulation of one parental homoeolog despite the expression level dominance of the other parental genome. In addition, allele-specific expression of polyploids can be more associated with the genotype than to other factors. Similarly, Powell and colleagues [34] stated that homoeolog expression bias was inherent to the wheat genotype, while the infection by necrotrophic *Fusarium pseudograminearum* mostly altered the magnitude of expression of the subgenomes. Knowledge of gene expression in the parental subgenomes is still lacking in the literature for studying expression level dominance and homoeolog-specific expression in sugarcane.

Genes showing preferential allele expression can be used for targeted genotyping to discover QTL regions associated with a trait. ASEGs found in rice subjected to different drought treatments were closely located to eight markers surrounding QTLs with effects on grain yield under drought [13].When implemented in breeding of polyploid crops, estimation of doses can improve phenotypic predictions compared to the diploid approximation for heterozygous loci. According to De Lara and colleagues [8], the predictive ability of genomic selection models was higher when considering allele dosages in the autotetraploid *Panicum maximum*. In addition to using genomic doses, knowledge of expression biases could improve the accuracy of predictive models for plant breeding, especially in the genomic selection context. Depending on the trait evaluated, ASEGs are potential targets for associating genomic regions and phenotypes. Nowadays, sugarcane breeding focuses on bioenergy-associated traits [7]. Assessing the regulation of allele expression in *Saccharum* can provide targets to help in this process.

Allele-specific expression of a large set of expressed genes has been evaluated in plants [13, 18], but these studies are still limited in polyploids [34, 31, 48]. Polyploidy has a significant role in the evolution of plants and many crops are recognizable polyploids, while others experienced ancient polyploidization [29].Among the most important polyploid crops, sugarcane presents a complex genome, with a large set of chromosomes. Indeed, the wild species used to develop modern cultivars, *S. officinarum* and *S. spontaneum*, are polyploids showing up to ten groups of homologs, at least six chromosomes per group and aneuploidy is frequent in some accessions [33]. For this reason, we aimed to shed light on the occurrence of allele-specific expression in the genus by assessing a set of wild and hybrid genotypes. This is the first report of allele-specific expression in sugarcane using a large set of genes and multiple genotypes. To achieve this objective, we developed a model appropriate for assessing allele-specific expression in mixed-ploid organisms. This model can be easily applied to other polyploids, both with fixed and variable ploidy levels.

References

[1] Benjamini, Y. e Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

[2] Cai, M., Lin, J., Li, Z., Lin, Z., Ma, Y., Wang, Y., e Ming, R. (2020). Allele specific expression of Dof genes responding to hormones and abiotic stresses in sugarcane. *PLOS ONE*, 15(1):e0227716.

[3] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., e Riddell, A. (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1).

[4] Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., e Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, 16(1):1–12.

[5] Casu, R. E., Jarmey, J. M., Bonnett, G. D., e 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.

[6] Correr, F. H., Hosaka, G. K., Barreto, F. Z., Valadão, I. B., Balsalobre, T. W. A., Furtado, A., Henry, R. J., Carneiro, M. S., e Margarido, G. R. A. (2020). Differential expression in leaves of Saccharum genotypes contrasting in biomass production provides evidence of genes involved in carbon partitioning. *BMC Genomics*, 21(1):673.

[7] da Silva, J. A. (2017). The Importance of the Wild Cane Saccharum spontaneum for Bioenergy Genetic Breeding. *Sugar Tech*, 19(3):229–240.

[8] de C. Lara, L. A., Santos, M. F., Jank, L., Chiari, L., Vilela, M. d. M., Amadeu, R. R., dos Santos, J. P. R., Pereira, G. d. S., Zeng, Z.-B., e Garcia, A. A. F. (2019). Genomic Selection with Allele Dosage in Panicum maximum Jacq. *G3: Genes|Genomes|Genetics*, 9(8):2463–2475.

[9] De Mendonça Vilela, M., Del Bem, L. E., Van Sluys, M. A., De Setta, N., Kitajima, J. P., Cruz, G. M. Q., Sforça, D. A., De Souza, A. P., Ferreira, P. C. G., Grativol, C., Cardoso-Silva, C. B., Vicentini, R., e Vincentz, M. (2017). Analysis of Three Sugarcane Homo/Homeologous Regions Suggests Independent Polyploidization Events of Saccharum officinarum and Saccharum spontaneum. *Genome Biology and Evolution*, 9(2):266–278.

[10] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., e Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

[11] Diniz, A. L., Ferreira, S. S., Ten-Caten, F., Margarido, G. R., dos Santos, J. M., Barbosa, G. V. S., Carneiro, M. S., e Souza, G. M. (2019). Genomic resources for energy cane breeding in the post genomics era. *Computational and Structural Biotechnology Journal*, 17:1404–1414.

[12] Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., e Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6(5):e19379.

[13] Ereful, N. C., Liu, L. Y., Tsai, E., Kao, S. M., Dixit, S., Mauleon, R., Malabanan, K., Thomson, M., Laurena, A., Lee, D., Mackay, I., Greenland, A., Powell, W., e Leung, H. (2016). Analysis of Allelic Imbalance in Rice Hybrids Under Water Stress and Association of Asymmetrically Expressed Genes with Drought-Response QTLs. *Rice*, 9(1).

[14] FAOSTAT (2021). Food and agriculture organization of the United Nations. *Statistical database (http://www.fao.org/faostat/en/#data/QC/), accessed in 20 January 2021.*

[15] Garcia, A. A. F., Mollinari, M., Marconi, T. G., Serang, O. R., Silva, R. R., Vieira, M. L. C., Vicentini, R., Costa, E. A., Mancini, M. C., Garcia, M. O. S., Pastina, M. M., Gazaffi, R., Martins, E. R. F., Dahmer, N., Sforça, D. A., Silva, C. B. C., Bundock, P., Henry, R. J., Souza, G. M., van Sluys, M.-A., Landell, M. G. A., Carneiro, M. S., Vincentz, M. A. G., Pinto, L. R., Vencovsky, R., e Souza, A. P. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Scientific Reports*, 3(1):3399.

[16] Gemenet, D. C., da Silva Pereira, G., De Boeck, B., Wood, J. C., Mollinari, M., Olukolu, B. A., Diaz, F., Mosquera, V., Ssali, R. T., David, M., Kitavi, M. N., Burgos, G., Felde, T. Z., Ghislain, M., Carey, E., Swanckaert, J., Coin, L. J. M., Fei, Z., Hamilton, J. P., Yada, B., Yencho, G. C., Zeng, Z.-B., Mwanga, R. O. M., Khan, A., Gruneberg, W. J., e Buell, C. R. (2020). Quantitative trait loci and differential gene expression analyses reveal the genetic basis for negatively associated β-carotene and starch content in hexaploid sweetpotato [Ipomoea batatas (L.) Lam.]. *Theoretical and Applied Genetics*, 133(1):23–36.

[17] Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E., e Wendel, J. F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist*, 196(4):966–971.

[18] Hu, X., Wang, H., Diao, X., Liu, Z., Li, K., Wu, Y., Liang, Q., Wang, H., e Huang, C. (2016). Transcriptome profiling and comparison of maize ear heterosis during the spikelet and floret differentiation stages. *BMC Genomics*, 17(1):1–18.

[19] Kasirajan, L., Hoang, N. V., Furtado, A., Botha, F. C., e Henry, R. J. (2018). Transcriptome analysis highlights key differentially expressed genes involved in cellulose and lignin biosynthesis of sugarcane genotypes varying in fiber content. *Scientific Reports*, 8(1):1–16.

[20] Kim, D., Langmead, B., e Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360.

[21] Langmead, B. e Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

[22] Lavania, U. C. (2013). Polyploidy, body size, and opportunities for genetic enhancement and fixation of heterozygosity in plants. *The Nucleus*, 56(1):1–6.

[23] Mancini, M. C., Cardoso-Silva, C. B., Costa, E. A., Marconi, T. G., Garcia, A. A. F., e De Souza, A. P. (2017). New Developments in Sugarcane Genetics and Genomics. In Buckeridge, M. S. e De Souza, A. P., editors, *Advances of Basic Science for Second Generation Bioethanol from Sugarcane*, pages 159–174. Springer International Publishing, Cham.

[24] Mattiello, L., Riaño-Pachón, D. M., Martins, M. C. M., da Cruz, L. P., Bassi, D., Marchiori, P. E. R., Ribeiro, R. V., Labate, M. T. V., Labate, C. A., e Menossi, M. (2015). Physiological and transcriptional analyses of developmental stages along sugarcane leaf. *BMC Plant Biology*, 15(1):300.

[25] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., e DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

[26] Medeiros, C., Balsalobre, T. W. A., e Carneiro, M. S. (2020). Molecular diversity and genetic structure of Saccharum complex accessions. *PLOS ONE*, 15(5):e0233211.

[27] N'Diaye, A., Haile, J. K., Cory, A. T., Clarke, F. R., Clarke, J. M., Knox, R. E., e Pozniak, C. J. (2017). Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. *PLoS ONE*, 12(1):1–24.

[28] Nguyen, T. K. e Lim, J.-H. (2019). Tools for Chrysanthemum genetic research and breeding: Is genotyping-by-sequencing (GBS) the best approach? *Horticulture, Environment, and Biotechnology*, 60(5):625–635.

[29] Osborn, T. C., Chris Pires, J., Birchler, J. A., Auger, D. L., Chen, Z. J., Lee, H. S., Comai, L., Madlung, A., Doerge, R. W., Colot, V., e Martienssen, R. A. (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics*, 19(3):141–147.

[30] Pereira, G. S., Garcia, A. A. F., e Margarido, G. R. (2018). A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. *BMC Bioinformatics*, 19(1):1–10.

[31] Pham, G. M., Newton, L., Wiegert-Rininger, K., Vaillancourt, B., Douches, D. S., e Buell, C. R. (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *Plant Journal*, 92(4):624–637.

[32] Piperidis, G., Piperidis, N., e D'Hont, A. (2010). Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. *Molecular Genetics and Genomics*, 284(1):65–73.

[33] Piperidis, N. e D'Hont, A. (2020). Sugarcane genome architecture decrypted with chromosome-specific oligo probes. *The Plant Journal*, page tpj.14881.

[34] Powell, J. J., Fitzgerald, T. L., Stiller, J., Berkman, P. J., Gardiner, D. M., Manners, J. M., Henry, R. J., e Kazan, K. (2017). The defence-associated transcriptome of hexaploid wheat displays homoeolog expression and induction bias. *Plant Biotechnology Journal*, 15(4):533–543.

[35] Romanel, A., Lago, S., Prandi, D., Sboner, A., e Demichelis, F. (2015). ASEQ: Fast allele-specific studies from next-generation sequencing data. *BMC Medical Genomics*, 8(1):1–12.

[36] Scortecci, K. C., Creste, S., Jr., T. C., Xavier, M. A., Landell, M. G. A., Figueira, A., e Benedito, V. A. (2012). Challenges, Opportunities and Recent Advances in Sugarcane Breeding. In *Plant Breeding*, page 352. InTech.

[37] Sehgal, D. e Dreisigacker, S. (2019). Haplotypes-based genetic analysis: Benefits and challenges. *Vavilovskii Zhurnal Genetiki i Selektsii*, 23(7):803–808.

[38] Serang, O., Mollinari, M., e Garcia, A. A. F. (2012). Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS ONE*, 7(2):1–13.

[39] Sforça, D. A., Vautrin, S., Cardoso-Silva, C. B., Mancini, M. C., Romero-da Cruz, M. V., Pereira, G. d. S., Conte, M., Bellec, A., Dahmer, N., Fourment, J., Rodde, N., Van Sluys, M.-A., Vicentini, R., Garcia, A. A. F., Forni-Martins, E. R., Carneiro, M. S., Hoffmann, H. P., Pinto, L. R., Landell, M. G. d. A., Vincentz, M., Berges, H., e de Souza, A. P. (2019). Gene Duplication in the Sugarcane Genome: A Case Study of Allele Interactions and Evolutionary Patterns in Two Genic Regions. *Frontiers in Plant Science*, 10(May).

[40] Soneson, C. e Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91.

[41] Spoelhof, J. P., Soltis, P. S., e Soltis, D. E. (2017). Pure polyploidy: Closing the gaps in autopolyploid research. *Journal of Systematics and Evolution*, 55(4):340–352.

[42] Stan Development Team (2018). RStan: the R interface to Stan.

[43] Tuch, B. B., Laborde, R. R., Xu, X., Gu, J., Chung, C. B., Monighetti, C. K., Stanley, S. J., Olsen, K. D., Kasperbauer, J. L., Moore, E. J., Broomer, A. J., Tan, R., Brzoska, P. M., Muller, M. W., Siddiqui, A. S., Asmann, Y. W., Sun, Y., Kuersten, S., Barker, M. A., De La Vega, F. M., e Smith, D. I. (2010). Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS ONE*, 5(2).

[44] Vicentini, R., Bottcher, A., Dos Santos Brito, M., Dos Santos, A. B., Creste, S., De Andrade Landell, M. G., Cesarino, I., e Mazzafera, P. (2015). Large-scale transcriptome analysis of two sugarcane genotypes contrasting for lignin content. *PLoS ONE*, 10(8):e0134909.

[45] Vieira, M. L. C., Almeida, C. B., Oliveira, C. A., Tacuatiá, L. O., Munhoz, C. F., Cauz-Santos, L. A., Pinto, L. R., Monteiro-Vitorello, C. B., Xavier, M. A., e Forni-Martins, E. R. (2018). Revisiting meiosis in sugarcane: Chromosomal irregularities and the prevalence of bivalent configurations. *Frontiers in Genetics*, 9(JUN):1–12.

[46] Wang, Z., Gerstein, M., e Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

[47] Wood, D. L., Nones, K., Steptoe, A., Christ, A., Harliwong, I., Newell, F., Bruxner, T. J., Miller, D., Cloonan, N., e Grimmond, S. M. (2015). Recommendations for accurate resolution of Gene and isoform allele-specific expression in RNA-seq data. *PLoS ONE*, 10(5):1–27.

[48] Yoo, M. J., Szadkowski, E., e Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, 110(2):171–180.

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

[50] Zhang, J., Nagai, C., Yu, Q., Pan, Y. B., Ayala-Silva, T., Schnell, R. J., Comstock, J. C., Arumuganathan, A. K., e Ming, R. (2012). Genome size variation in three Saccharum species. *Euphytica*, 185(3):511–519.

[51] Zhang, J., Sharma, A., Yu, Q., Wang, J., Li, L., Zhu, L., Zhang, X., Chen, Y., e Ming, R. (2016a). Comparative structural analysis of Bru1 region homeologs in Saccharum spontaneum and S. officinarum. *BMC Genomics*, 17(1).

[52] Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., Zhu, F., Jones, T., Zhu, X., Bowers, J., Wai, C. M., Zheng, C., Shi, Y., Chen, S., Xu, X., Yue, J., Nelson, D. R., Huang, L., Li, Z., Xu, H., Zhou, D., Wang, Y., Hu, W., Lin, J., Deng, Y., Pandey, N., Mancini, M., Zerpa, D., Nguyen, J. K., Wang, L., Yu, L., Xin, Y., Ge, L., Arro, J., Han, J. O., Chakrabarty, S., Pushko, M., Zhang, W., Ma, Y., Ma, P., Lv, M., Chen, F., Zheng, G., Xu, J., Yang, Z., Deng, F., Chen, X., Liao, Z., Zhang, X., Lin, Z., Lin, H., Yan, H., Kuang, Z., Zhong, W., Liang, P., Wang, G., Yuan, Y., Shi, J., Hou, J., Lin, J., Jin, J., Cao, P., Shen, Q., Jiang, Q., Zhou, P., Ma, Y., Zhang, X., Xu, R., Liu, J., Zhou, Y., Jia, H., Ma, Q., Qi, R., Zhang, Z., Fang, J., Fang, H., Song, J., Wang, M., Dong, G., Wang, G., Chen, Z., Ma, T., Liu, H., Dhungana, S. R., Huss, S. E., Yang, X., Sharma, A., Trujillo, J. H., Martinez, M. C., Hudson, M., Riascos, J. J., Schuler, M., Chen, L. Q., Braun, D. M., Li, L., Yu, Q., Wang, J., Wang, K., Schatz, M. C., Heckerman, D., Van Sluys, M. A., Souza, G. M., Moore, P. H., Sankoff, D., VanBuren, R., Paterson, A. H., Nagai, C., e Ming, R. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nature Genetics*, 50(11):1565–1573.

Additional file 1

Supplementary figures

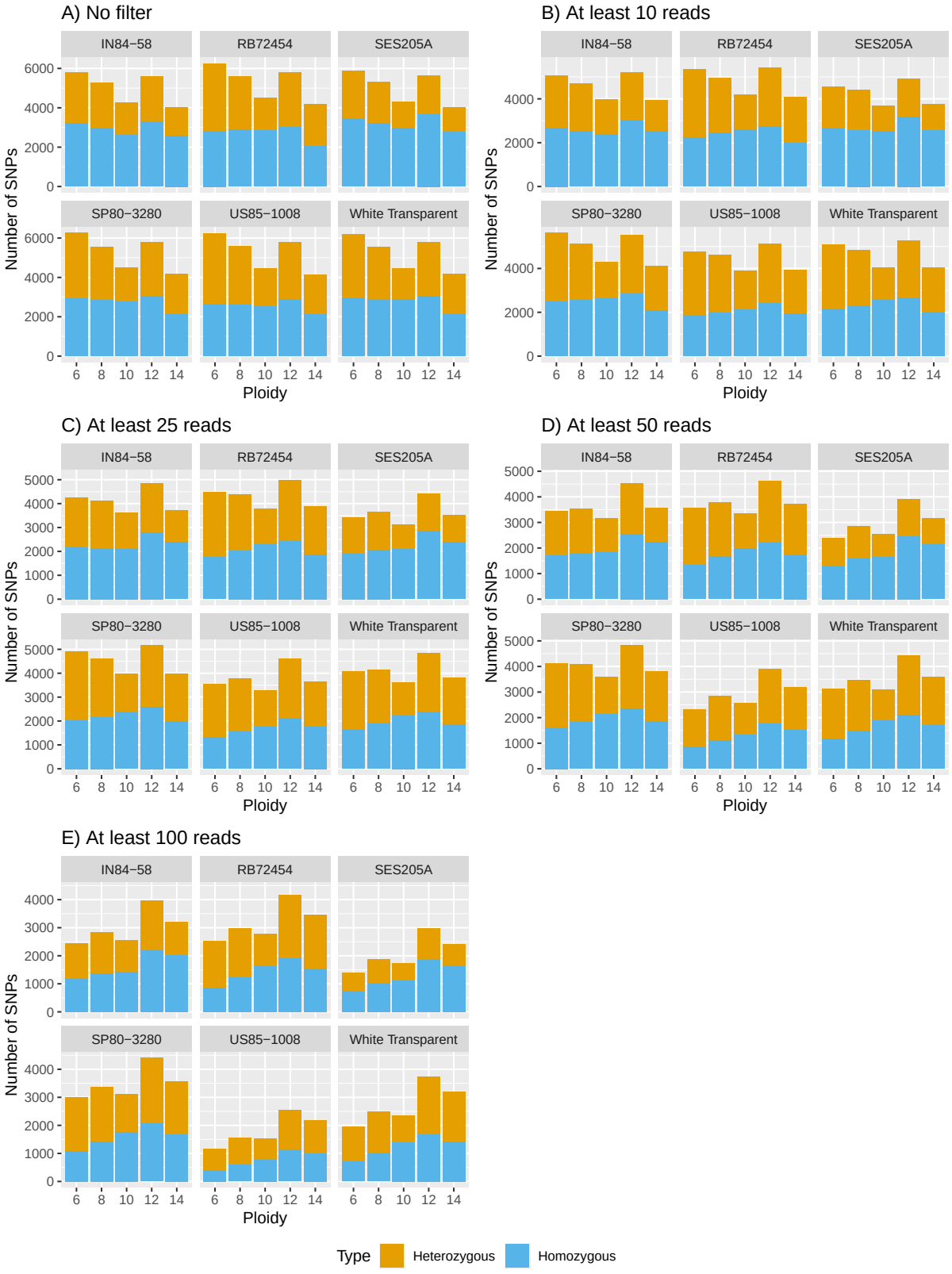

Figure 1: Number of SNPs according to the ploidy levels for each genotype. The number of homozygous and heterozygous SNPs for each ploidy level are presented in different scenarios: without any filter (A); with a minimum count of 10 (B), 25 (C), 50 (D) and 100 (E) GBS reads. Heterozygous SNPs are shown in orange, while the homozygous ones are in blue. Each subplot identifies a different genotype.

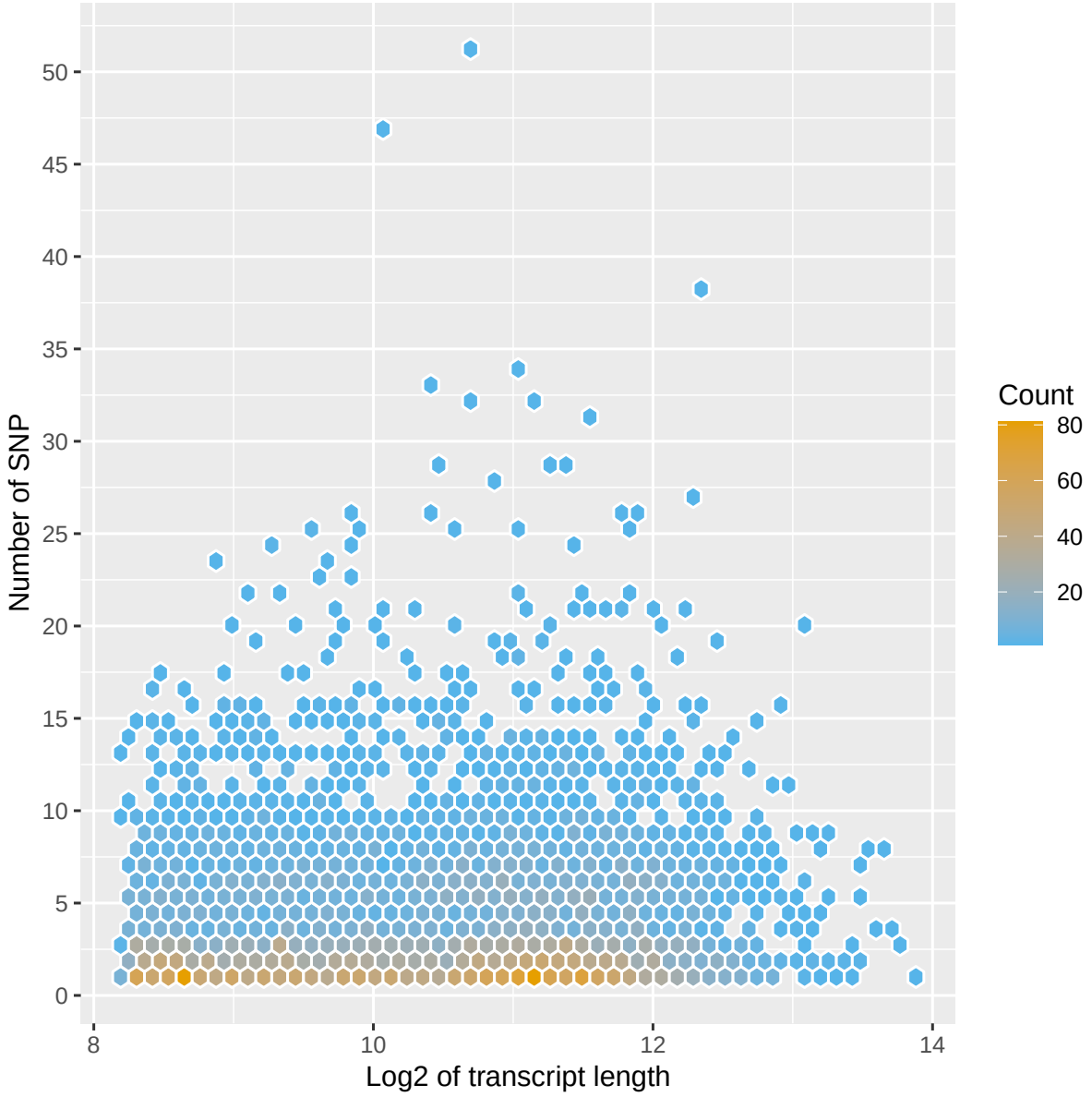

Figure 2: Number of SNPs as a function of transcript length. Counts are represented by a gradient from low (blue) to high numbers (orange).

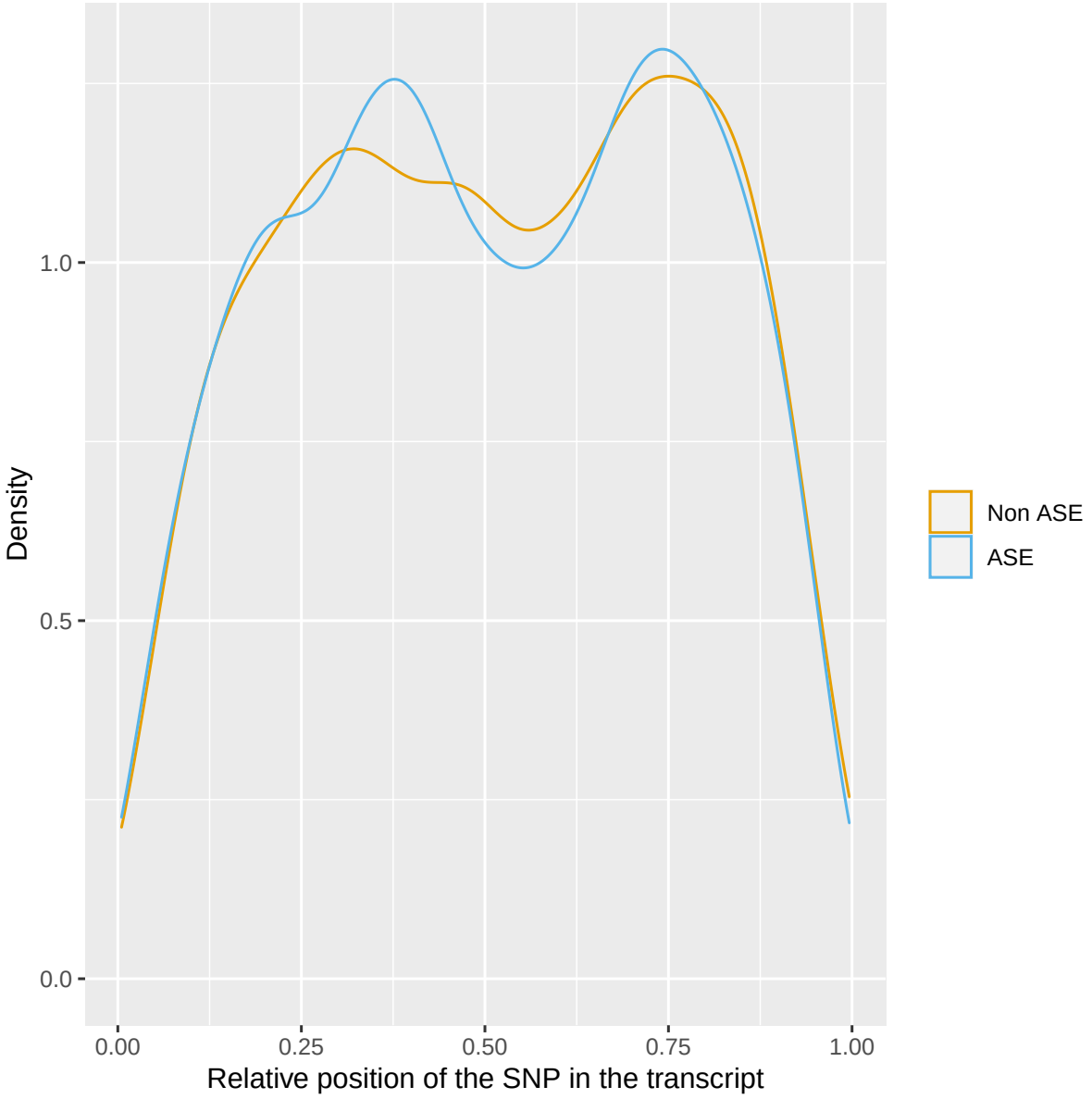

Figure 3: Distribution of SNPs along the normalized length of transcripts. SNPs with no allele-specific expression (ASE) are shown in orange and those with significant ASE are in blue.

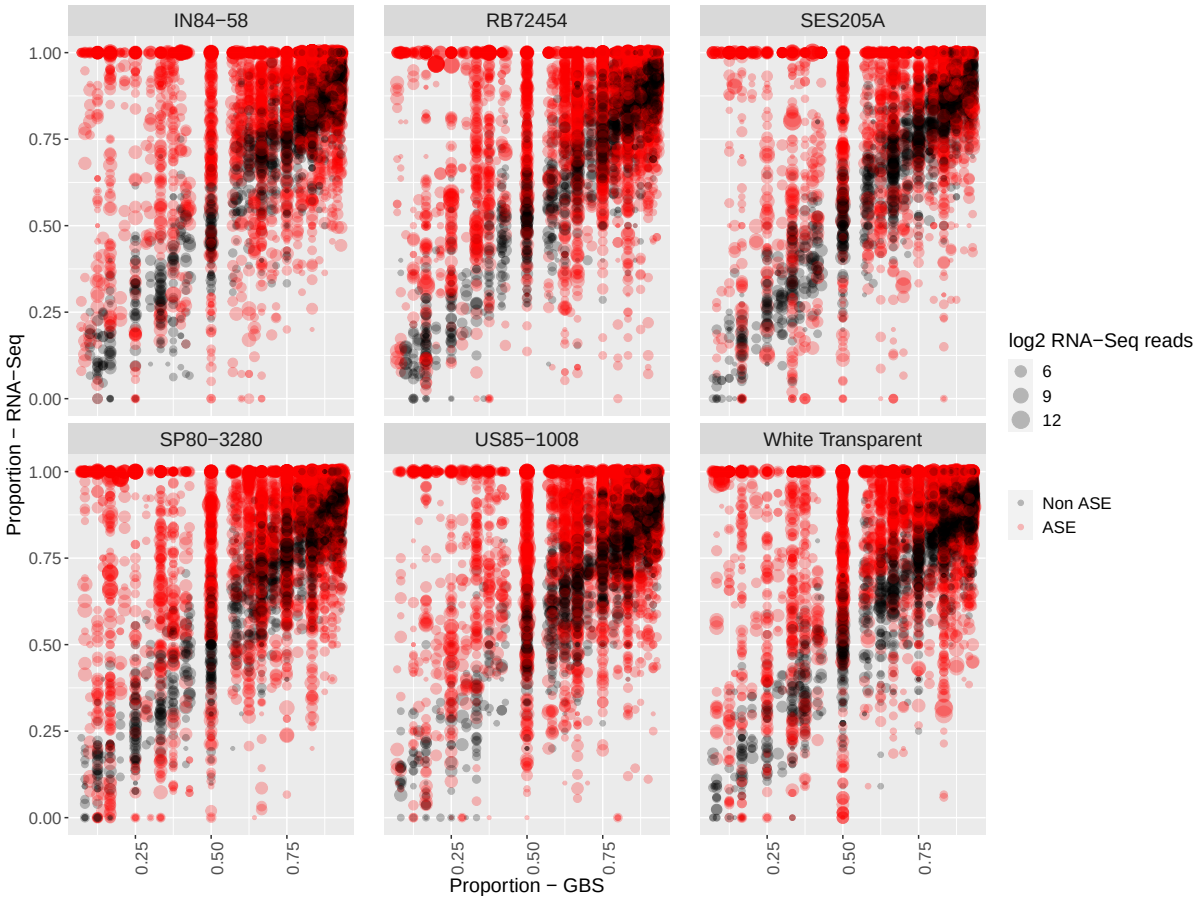

Figure 4: Relationship between the proportion of reads with the reference allele in genomic and expression datasets. SNPs with significant ASE are colored in red and non-ASE SNPs are colored in black. The size of each point is proportional to the overall expression level of both alleles of each SNP. Data for each genotype is shown in a different subplot.

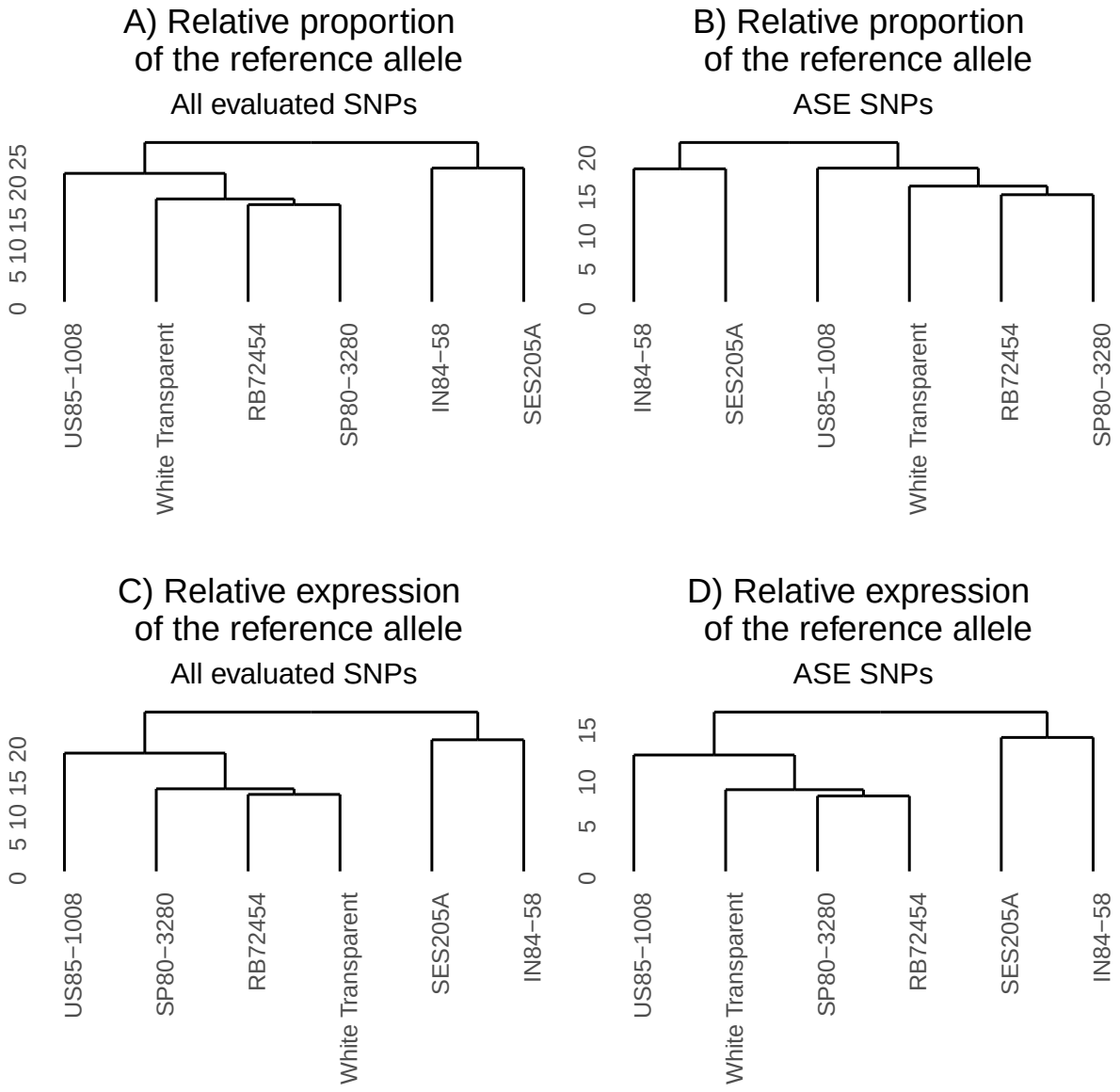

Figure 5: Dendrograms of the hierarchical clustering of genotypes based on heterozygous SNPs. Genotypes were clustered based on the genomic proportion of the reference allele using all heterozygous SNPs (A) and SNPs with allele-specific expression (ASE) only (B). The relative expression of the reference allele from all heterozygous SNPs (C) and ASE SNPs (D) was also used to cluster the genotypes.

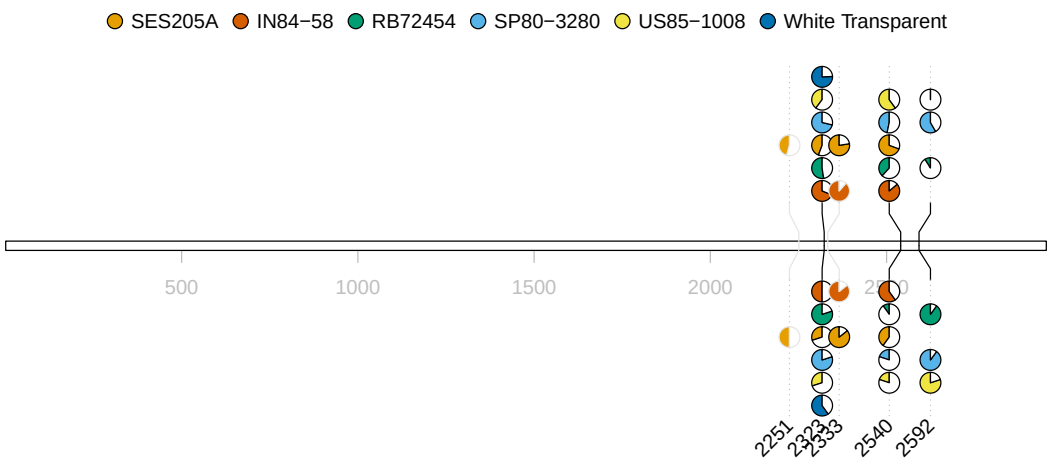

Figure 6: SNPs identified in the gene coding for *ENHANCED DISEASE RESISTANCE 2*. SNPs showing ASE have black borders, while those not showing ASE have grey borders. Relative genomic dosage of the alleles is represented by pie charts in the bottom part of the plot. The expressed proportion of each allele is represented by pie charts at the top. Different colors represent SNPs identified in each genotype (see legend). Colored shaded areas in the pie charts represent the relative dosage or the expressed proportion of the reference allele, while the alternative allele is in white.

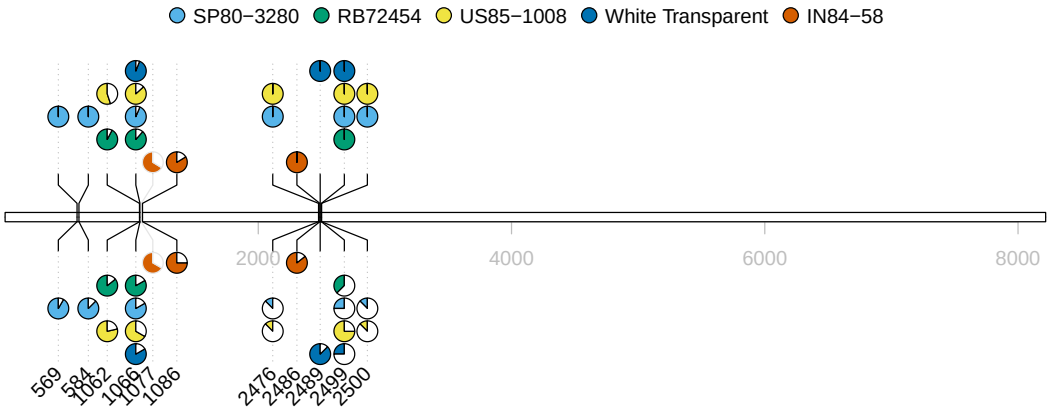

Figure 7: SNPs identified in the gene coding for *UTP–glucose-1-phosphate uridylyltransferase*. SNPs showing ASE have black borders, while those not showing ASE have grey borders. Relative genomic dosage of the alleles is represented by pie charts in the bottom part of the plot. The expressed proportion of each allele is represented by pie charts at the top. Different colors represent SNPs identified in each genotype (see legend). Colored shaded areas in the pie charts represent the relative dosage or the expressed proportion of the reference allele, while the alternative allele is in white.

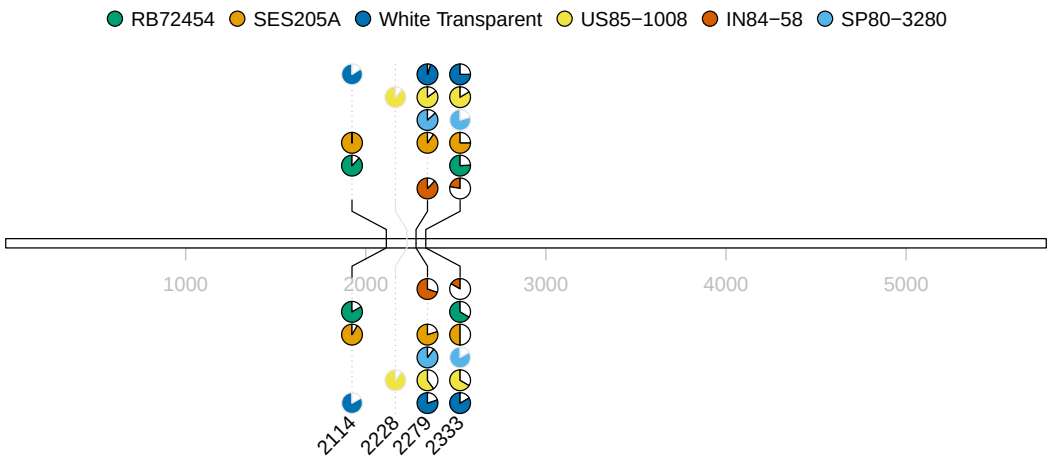

Figure 8: SNPs identified in the gene coding for *RuBisCO large subunit-binding protein subunit alpha, chloroplastic*. SNPs showing ASE have black borders, while those not showing ASE have grey borders. Relative genomic dosage of the alleles is represented by pie charts in the bottom part of the plot. The expressed proportion of each allele is represented by pie charts at the top. Different colors represent SNPs identified in each genotype (see legend). Colored shaded areas in the pie charts represent the relative dosage or the expressed proportion of the reference allele, while the alternative allele is in white.

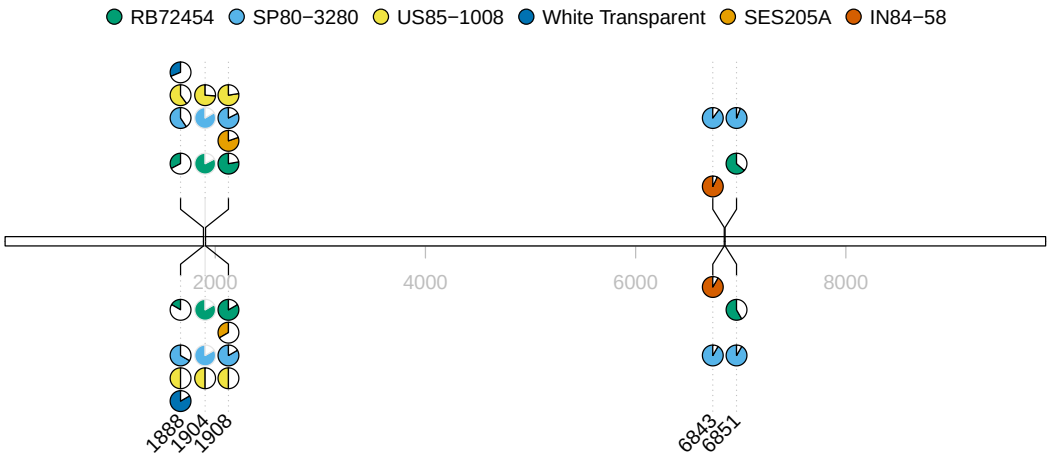

Figure 9: SNPs identified in the gene coding for *Phosphoenolpyruvate carboxylase 3*. SNPs showing ASE have black borders, while those not showing ASE have grey borders. Relative genomic dosage of the alleles is represented by pie charts in the bottom part of the plot. The expressed proportion of each allele is represented by pie charts at the top. Different colors represent SNPs identified in each genotype (see legend). Colored shaded areas in the pie charts represent the relative dosage or the expressed proportion of the reference allele, while the alternative allele is in white.

Supplementary tables

Table 1: Accessions of the Brazilian Panel of Sugarcane Genotypes used in the study. For each accession we show its species description, where hybrids are considered as *Saccharum* spp. The accessions were classified according to the phenotype - high or low biomass.

Accessions	Species	Biomass group
IN84-58	*Saccharum spontaneum*	High
SES205A	*Saccharum spontaneum*	High
US85-1008	*Saccharum* spp. (hybrid)	High
White Transparent	*Saccharum officinarum*	Low
RB72454	*Saccharum* spp. (hybrid)	Low
SP80-3280	*Saccharum* spp. (hybrid)	Low

Table 2: Number of heterozygous SNPs in each genotype.

Genotypes	Number of heterozygous SNPs
IN84-58	6209
RB72454	8459
SES205A	4745
SP80-3280	8573
US85-1008	8901
White Transparent	8053

Table 3: Number of SNPs with significant allele-specific expression (ASE), SNPs without ASE (non-ASE), genes with ASE (ASEG) and genes without ASE (non-ASEG) according to the genotypes.

Genotype	Non-ASE	ASE	Non-ASEG	ASEG
IN84-58	872	1421	374	758
RB72454	1238	1751	432	836
SES205A	738	1051	344	585
SP80-3280	1403	1763	493	872
US85-1008	1308	1808	451	900
White Transparent	1238	1582	427	809

Additional file 2

Supplementary tables

Table 4: Functional annotation of genes showing allele-specific expression identified in each genotype.

Description	IN84.58	White.Transparent	RB72454	SP80.3280	US85.1008	SES205A
Transcription factor TGAL7 {ECO:0000305}	yes	yes	yes	yes	yes	no
Diacylglycerol O-acyltransferase 3 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Protein bfr2	yes	yes	yes	no	yes	no
Zinc finger A20 and AN1 domain-containing stress-associated protein 1	yes	yes	yes	no	yes	yes
WAT1-related protein At3g18200	yes	no	no	no	no	no
Transcription factor TGAL10 {ECO:0000305}	yes	no	no	yes	no	yes
Neutral ceramidase	yes	yes	no	no	no	no
Histone deacetylase 14	yes	yes	yes	yes	yes	yes
Probable RNA-binding protein ARP1	yes	no	no	no	no	yes
LON peptidase N-terminal domain and RING finger protein 1	yes	no	no	no	yes	no
Zinc finger protein WIP4	yes	yes	no	yes	no	yes
Ankyrin repeat-containing protein At5g02620	yes	no	no	yes	no	no
AUGMIN subunit 5 {ECO:0000303|PubMed:22505726}	yes	no	no	no	no	yes
Probable disease resistance protein RPP1 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Calmodulin-binding protein 60 B {ECO:0000303|PubMed:11782485}	yes	yes	yes	yes	yes	no
Calcium-transporting ATPase 5, plasma membrane-type {ECO:0000305}	yes	yes	yes	yes	yes	yes
bZIP transcription factor TRAB1 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Probable inactive purple acid phosphatase 27	yes	no	no	yes	yes	yes
Aldehyde dehydrogenase family 7 member A1	yes	yes	yes	yes	yes	no
Dynamin-related protein 12A	yes	yes	no	no	yes	no
Protein FLX-like 2	yes	yes	no	yes	no	yes
Probable galactinol–sucrose galactosyltransferase 2	yes	yes	no	no	yes	yes
Stress-related protein	yes	yes	yes	yes	yes	yes
Stomatal closure-related actin-binding protein 1 {ECO:0000303|PubMed:21719691}	yes	yes	yes	yes	yes	yes
Linoleate 9S-lipoxygenase 6	yes	yes	yes	no	no	no
Dymeclin	yes	yes	no	yes	no	no
Protein ANTHESIS POMOTING FACTOR 1 {ECO:0000303|PubMed:27968983}	yes	yes	yes	yes	yes	no
Monooxygenase 2 {ECO:0000303|PubMed:10216258}	yes	no	no	no	no	no
Proline-rich receptor-like protein kinase PERK13	yes	yes	no	yes	no	no
Metal transporter Nramp3	yes	no	yes	no	yes	yes
G-type lectin S-receptor-like serine/threonine-protein kinase At2g19130	yes	yes	yes	yes	yes	yes
L-type lectin-domain containing receptor kinase IX.1 {ECO:0000303|PubMed:19773388}	yes	yes	no	yes	yes	yes
Uncharacterized protein At1g03900	yes	no	yes	yes	no	no
Double-stranded RNA-binding protein 5	yes	no	yes	no	no	no
Coatomer subunit gamma-2	yes	no	no	yes	no	no
bZIP transcription factor 27 {ECO:0000303|PubMed:11906833}	yes	no	no	yes	yes	yes

Protein NARROW LEAF 1 {ECO:0000303|PubMed:18562767}	yes	yes	yes	yes	no	yes
Serotonin N-acetyltransferase 1, chloroplastic {ECO:0000305}	yes	no	no	no	no	no
Cysteine-rich receptor-like protein kinase 28	yes	yes	yes	no	yes	yes
Glutathione S-transferase U18	yes	yes	no	no	yes	yes
Plasmodesmata-located protein 8 {ECO:0000303|PubMed:28786767}	yes	no	no	no	no	yes
Transcription factor PHYTOCHROME INTERACTING FACTOR-LIKE 15 {ECO:0000303|PubMed:17485859}	yes	yes	yes	yes	yes	yes
ADP-ribosylation factor GTPase-activating protein AGD12	yes	yes	yes	yes	yes	yes
RNA polymerase sigma factor sigC	yes	yes	no	no	no	no
Protein ESSENTIAL FOR POTEXVIRUS ACCUMULATION 1 {ECO:0000303|PubMed:27402258}	yes	yes	yes	no	yes	yes
Suppressor of mec-8 and unc-52 protein homolog 1	yes	yes	yes	yes	no	no
Protein YIPF1 homolog	yes	yes	no	no	yes	no
LINE-1 retrotransposable element ORF2 protein	yes	yes	no	yes	yes	yes
Damage-control phosphatase At2g17340 {ECO:0000303|PubMed:27322068}	yes	no	yes	no	no	no
Pheophytinase, chloroplastic	yes	no	yes	yes	yes	yes
Auxin response factor 25	yes	no	no	no	yes	no
Actin-histidine N-methyltransferase {ECO:0000250|UniProtKB:Q86TU7}	yes	yes	no	yes	no	no
eIF-2-alpha kinase GCN2 {ECO:0000250|UniProtKB:Q9HGN1}	yes	no	no	no	no	yes
YTH domain-containing protein ECT4 {ECO:0000305}	yes	no	no	no	no	no
Protein DETOXIFICATION 33 {ECO:0000303|PubMed:11739388}	yes	yes	yes	yes	yes	yes
ATP-dependent RNA helicase SUV3, mitochondrial {ECO:0000303|PubMed:23808500}	yes	no	no	no	no	no
Carbonic anhydrase, chloroplastic	yes	no	yes	yes	no	no
Dicarboxylate transporter 2.1, chloroplastic	yes	no	no	yes	yes	no
3-ketoacyl-CoA synthase 4 {ECO:0000303|PubMed:18465198}	yes	yes	yes	yes	yes	yes
Probable protein phosphatase 2C 39	yes	yes	yes	yes	yes	no
UTP–glucose-1-phosphate uridylyltransferase	yes	yes	yes	yes	yes	no
Glucose-1-phosphate adenylyltransferase small subunit 2, chloroplastic/amyloplastic/cytosolic {ECO:0000305}	yes	yes	yes	yes	yes	yes
Receptor-like protein EIX2 {ECO:0000305}	yes	no	no	yes	yes	no
Soluble starch synthase 2-2, chloroplastic/amyloplastic	yes	yes	no	no	no	no
Auxilin-related protein 2	yes	yes	yes	yes	yes	no
Serine decarboxylase 1	yes	no	no	no	no	no
U-box domain-containing protein 63	yes	yes	no	yes	yes	yes
Carotene epsilon-monooxygenase, chloroplastic	yes	yes	yes	no	yes	yes
Cyclin-dependent kinases regulatory subunit 1	yes	no	no	no	no	no
DEAD-box ATP-dependent RNA helicase 27	yes	no	yes	yes	no	yes
Cell division cycle protein 27 homolog B	yes	yes	no	no	no	no
DEAD-box ATP-dependent RNA helicase 37	yes	no	no	no	no	no
Probable calcium-transporting ATPase 9, plasma membrane-type {ECO:0000305}	yes	yes	yes	no	yes	no
Potassium transporter 7	yes	no	yes	no	no	no
Transposon Tf2-6 polyprotein	yes	yes	yes	yes	no	yes
WRKY transcription factor SUSIBA2 {ECO:0000305|PubMed:12953112}	yes	no	no	no	yes	no
Plasma membrane ATPase 1	yes	yes	yes	no	yes	yes
Dr1-associated corepressor	yes	no	no	no	yes	no
Polyadenylate-binding protein RBP45	yes	yes	yes	yes	yes	no
Probable inactive dual specificity protein phosphatase-like At4g18593	yes	no	no	no	no	no
Protein TIFY 3 {ECO:0000305}	yes	no	yes	yes	yes	no
Probable acylpyruvase FAHD1, mitochondrial {ECO:0000305}	yes	yes	no	yes	no	no
Craniofacial development protein 2	yes	yes	yes	yes	yes	no

BTB/POZ and MATH domain-containing protein 4	yes	no	yes	no	no	no
1-aminocyclopropane-1-carboxylate oxidase homolog 3	yes	no	yes	no	no	yes
Protein TIFY 6a {ECO:0000305}	yes	no	no	no	yes	yes
Peroxisomal membrane protein PMP22	yes	no	no	no	no	no
Protein CROWDED NUCLEI 1 {ECO:0000303|PubMed:24308514}	yes	yes	no	no	yes	no
RNA-binding protein 42	yes	yes	yes	yes	yes	no
Epoxide hydrolase A {ECO:0000303|PubMed:16511284}	yes	no	no	yes	no	no
Peptide methionine sulfoxide reductase B1, chloroplastic	yes	yes	no	yes	yes	yes
Siroheme synthase {ECO:0000255|HAMAP-Rule:MF_01646}	yes	no	no	no	no	no
UDP-N-acetylglucosamine–N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase {ECO:0000255|HAMAP-Rule:MF_00033}	yes	no	no	no	no	no
GDSL esterase/lipase At4g01130	yes	no	no	no	no	no
Uncharacterized protein YKR070W	yes	yes	yes	yes	yes	yes
Metallothionein-like protein 1B	yes	yes	yes	yes	yes	yes
Protein NRT1/ PTR FAMILY 8.3	yes	yes	yes	yes	yes	yes
Calcium-dependent protein kinase 19 {ECO:0000305}	yes	yes	no	yes	no	no
Pentatricopeptide repeat-containing protein At4g30825, chloroplastic	yes	no	no	yes	no	no
Protein CHLOROPLAST ENHANCING STRESS TOLERANCE, chloroplastic {ECO:0000305}	yes	yes	yes	yes	no	no
Protein NRT1/ PTR FAMILY 5.10	yes	yes	yes	yes	yes	yes
Protein MICRORCHIDIA 6 {ECO:0000303|PubMed:22555433}	yes	no	no	yes	yes	no
Probable calcium-binding protein CML27	yes	yes	yes	yes	yes	no
Uncharacterized protein MJ1408	yes	yes	yes	yes	yes	no
Polyadenylate-binding protein 2	yes	yes	yes	yes	no	yes
Probable 2-oxoglutarate-dependent dioxygenase At5g05600 {ECO:0000305}	yes	no	no	no	yes	no
Auxin response factor 1	yes	yes	yes	yes	yes	yes
Ferredoxin–nitrite reductase, chloroplastic	yes	yes	yes	yes	yes	yes
Callose synthase 9	yes	no	yes	yes	yes	yes
Ubiquitin domain-containing protein DSK2b	yes	yes	no	yes	yes	yes
Filament-like plant protein 3 {ECO:0000303|PubMed:11972898}	yes	no	no	yes	no	no
Reactive Intermediate Deaminase A, chloroplastic {ECO:0000303|PubMed:25070638}	yes	yes	yes	yes	yes	yes
BTB/POZ and MATH domain-containing protein 1	yes	yes	yes	yes	yes	yes
Receptor-like serine/threonine-protein kinase SD1-8	yes	yes	yes	yes	no	yes
Probable glutathione S-transferase GSTU6	yes	yes	no	no	no	no
Putative adagio-like protein 2	yes	yes	yes	yes	yes	no
Transmembrane emp24 domain-containing protein p24delta3	yes	yes	yes	no	yes	yes
RNA-binding protein 25	yes	yes	yes	yes	yes	yes
APETALA2-like protein 5 {ECO:0000305}	yes	no	no	no	no	yes
Universal stress protein A-like protein	yes	no	no	no	no	no
Disease resistance protein RGA5 {ECO:0000305}	yes	no	yes	yes	yes	yes
Protein NAR1 {ECO:0000303|PubMed:23734982}	yes	no	no	yes	no	no
Putative vacuolar protein sorting-associated protein 13A	yes	yes	yes	yes	yes	no
Transcription factor VIP1	yes	no	yes	no	no	yes
Inorganic phosphate transporter 2-1, chloroplastic	yes	no	yes	no	yes	no
Heat shock cognate 70 kDa protein	yes	no	no	no	no	no
Lipase	yes	yes	yes	yes	yes	yes
RING-H2 finger protein ATL74	yes	no	no	no	no	no
Calmodulin-binding protein 60 C {ECO:0000303|PubMed:11782485}	yes	yes	yes	yes	yes	yes
Non-lysosomal glucosylceramidase {ECO:0000305}	yes	yes	yes	yes	yes	yes

RNA pseudouridine synthase 4, mitochondrial	yes	no	yes	no	yes	no
Transposon Ty3-I Gag-Pol polyprotein	yes	yes	yes	no	yes	no
Katanin p60 ATPase-containing subunit A-like 2 {ECO:0000255|HAMAP-Rule:MF_03025}	yes	no	no	no	no	no
Palmitoyl-acyl carrier protein thioesterase, chloroplastic	yes	no	no	yes	yes	no
HBS1-like protein	yes	no	no	no	no	no
Protein NRT1/ PTR FAMILY 2.7	yes	no	no	no	yes	yes
Non-functional NADPH-dependent codeinone reductase 2	yes	yes	yes	yes	yes	yes
5'-nucleotidase domain-containing protein DDB_G0275467	yes	no	yes	yes	no	yes
Delta-1-pyrroline-5-carboxylate dehydrogenase 12A1, mitochondrial	yes	no	no	no	yes	no
Putative disease resistance protein RGA1	yes	yes	yes	yes	yes	yes
Probable protein phosphatase 2C 55	yes	yes	yes	yes	yes	no
Secoisolariciresinol dehydrogenase	yes	no	no	no	no	no
Protein NDH-DEPENDENT CYCLIC ELECTRON FLOW 5 {ECO:0000305}	yes	no	no	no	no	no
Pentatricopeptide repeat-containing protein At3g12770	yes	yes	no	yes	yes	yes
Glucose-6-phosphate isomerase 1, chloroplastic	yes	yes	yes	yes	yes	no
Wall-associated receptor kinase 2	yes	no	no	no	no	no
Heat stress transcription factor A-4d	yes	no	yes	no	yes	no
Serine carboxypeptidase-like 50	yes	yes	yes	yes	yes	yes
YTH domain-containing protein ECT3 {ECO:0000305}	yes	no	yes	yes	yes	yes
Ras-related protein Rab-2-A	yes	no	yes	no	no	no
Ras-related protein Rab-2-B	yes	no	no	yes	no	yes
BTB/POZ domain-containing protein At1g30440	yes	yes	yes	yes	yes	yes
WAT1-related protein At5g64700	yes	no	yes	yes	yes	no
Triacylglycerol lipase SDP1	yes	yes	yes	yes	yes	yes
Low affinity sulfate transporter 3	yes	yes	no	yes	yes	yes
Protein WRKY1	yes	yes	yes	yes	no	no
Polyadenylate-binding protein 8	yes	no	no	no	yes	no
Alcohol dehydrogenase-like 7	yes	no	no	no	no	no
O-glucosyltransferase rumi	yes	no	yes	yes	no	yes
Ribosomal RNA small subunit methyltransferase, chloroplastic {ECO:0000305}	yes	yes	yes	no	yes	no
Protein NETWORKED 2D {ECO:0000303|PubMed:22840520}	yes	yes	yes	yes	no	no
Putative 3,4-dihydroxy-2-butanone kinase	yes	no	yes	yes	yes	no
BTB/POZ domain-containing protein POB1	yes	yes	no	yes	no	no
Flowering time control protein FCA {ECO:0000305}	yes	yes	no	no	no	no
Small GTPase LIP1 {ECO:0000303|PubMed:17683937}	yes	no	no	yes	no	no
Importin-5	yes	no	yes	yes	no	no
Arginine–tRNA ligase, chloroplastic/mitochondrial {ECO:0000305}	yes	yes	yes	yes	yes	no
Transcription-associated protein 1	yes	yes	yes	yes	yes	yes
Protein DETOXIFICATION 23 {ECO:0000303|PubMed:11739388}	yes	no	no	yes	no	yes
NADP-dependent malic enzyme, chloroplastic {ECO:0000305}	yes	yes	yes	yes	yes	yes
U-box domain-containing protein 75 {ECO:0000305}	yes	no	no	no	no	no
GEM-like protein 1	yes	yes	yes	yes	yes	yes
GPI transamidase component PIG-S	yes	no	no	no	no	no
Putative RNA-binding protein YlmH	yes	yes	yes	no	yes	yes
Brassinosteroid LRR receptor kinase BRI1 {ECO:0000305}	yes	no	no	no	yes	no
Phototropin-2	yes	yes	yes	yes	yes	yes
Long chain acyl-CoA synthetase 7, peroxisomal	yes	no	no	no	no	no
Trans-cinnamate 4-monooxygenase	yes	yes	yes	yes	yes	no
BTB/POZ and TAZ domain-containing protein 3	yes	yes	yes	yes	no	no
RNA polymerase sigma factor sigE, chloroplastic/mitochondrial	yes	yes	yes	yes	yes	no

Glutamate receptor 3.1	yes	yes	yes	yes	yes	no
Nudix hydrolase 16, mitochondrial	yes	yes	yes	yes	no	yes
Guanylate-binding protein 2	yes	no	no	yes	no	no
Probable pterin-4-alpha-carbinolamine dehydratase, chloroplastic {ECO:0000305}	yes	yes	yes	yes	yes	yes
Probable inactive ATP-dependent zinc metalloprotease FTSHI 3, chloroplastic	yes	yes	yes	yes	yes	yes
Protein ALWAYS EARLY 3	yes	no	no	no	no	no
5'-adenylylsulfate reductase-like 4	yes	yes	yes	yes	yes	yes
Lycopene epsilon cyclase, chloroplastic {ECO:0000303|PubMed:8837512}	yes	yes	no	yes	yes	no
Two-component response regulator-like PRR73	yes	no	no	no	no	no
Phosphate transporter PHO1-3	yes	yes	no	yes	yes	yes
Protein DA1-related 1 {ECO:0000303|PubMed:18483219}	yes	yes	no	no	no	yes
Peroxisomal membrane protein 11-5	yes	no	yes	yes	yes	yes
Pentatricopeptide repeat-containing protein At5g67570, chloroplastic	yes	yes	yes	yes	yes	no
Vacuolar protein sorting-associated protein 2 homolog 1	yes	no	no	yes	no	yes
DEAD-box ATP-dependent RNA helicase 32	yes	yes	yes	no	no	no
Probable peroxygenase 4	yes	no	no	yes	yes	yes
Probable LRR receptor-like serine/threonine-protein kinase At3g47570	yes	yes	yes	yes	yes	yes
E3 ubiquitin-protein ligase RHF2A {ECO:0000305}	yes	yes	yes	yes	yes	no
Homeobox-leucine zipper protein HOX6	yes	yes	yes	yes	no	yes
UPF0454 protein C12orf49 homolog	yes	yes	yes	yes	no	no
Protein IQ-DOMAIN 31	yes	yes	yes	no	no	no
Protein ALP1-like {ECO:0000305}	yes	no	no	no	no	no
Protein TITANIA {ECO:0000303|PubMed:30194869}	yes	no	yes	no	yes	no
Tuberculostearic acid methyltransferase UfaA1 {ECO:0000305}	yes	no	no	yes	yes	yes
Protochlorophyllide-dependent translocon component 52, chloroplastic	yes	yes	yes	yes	yes	yes
UDP-glucuronate:xylan alpha-glucuronosyltransferase 1	yes	no	yes	no	no	no
Splicing factor, suppressor of white-apricot homolog	yes	no	no	no	no	no
Hydroxyproline O-galactosyltransferase GALT2 {ECO:0000303|PubMed:23430255}	yes	no	no	no	yes	yes
Proteasome subunit alpha type-2	yes	no	yes	no	no	no
Protein CHROMATIN REMODELING 20 {ECO:0000303|PubMed:16547115}	yes	yes	yes	yes	yes	no
DNA-binding protein HEXBP	yes	yes	no	yes	no	no
Root phototropism protein 2	yes	no	yes	yes	yes	yes
Iron-sulfur assembly protein IscA-like 1, mitochondrial	yes	yes	yes	yes	yes	no
Basic leucine zipper and W2 domain-containing protein 2	yes	yes	yes	yes	yes	no
F-box protein SKIP31	yes	yes	yes	yes	no	no
Gamma-tubulin complex component 2	yes	yes	no	no	no	no
Protein ENHANCED DISEASE RESISTANCE 2	yes	yes	yes	yes	yes	yes
Probable GTP-binding protein OBGC1, chloroplastic	yes	yes	yes	yes	yes	yes
Sugar transport protein 11	yes	yes	no	yes	yes	yes
DNA-directed RNA polymerase 2B, chloroplastic/mitochondrial	yes	no	yes	no	yes	no
K(+) efflux antiporter 1, chloroplastic {ECO:0000303|PubMed:11500563}	yes	yes	yes	yes	yes	yes
Prolycopene isomerase, chloroplastic	yes	no	yes	yes	yes	yes
Probable transcriptional regulator SLK2	yes	yes	yes	yes	yes	no
RAN GTPase-activating protein 1	yes	yes	yes	yes	yes	yes
Protein furry homolog-like	yes	no	no	no	yes	yes
Meiotic recombination protein DMC1 homolog A {ECO:0000305}	yes	no	yes	yes	no	no
Putative yippee-like protein Os10g0369500	yes	yes	yes	no	yes	yes
Pentatricopeptide repeat-containing protein At1g10910, chloroplastic	yes	no	no	no	no	yes
CBL-interacting protein kinase 1	yes	no	no	no	no	no
Protein ACTIVITY OF BC1 COMPLEX KINASE 3, chloroplastic {ECO:0000303|PubMed:23673981}	yes	no	no	no	no	no

Ceramide kinase	yes	no	no	no	no	no
Reticulon-like protein B10	yes	yes	yes	yes	yes	yes
CASP-like protein 2C4	yes	no	no	no	no	no
B2 protein	yes	no	no	no	no	yes
NAD(P)H-quinone oxidoreductase subunit T, chloroplastic {ECO:0000305}	yes	yes	yes	yes	yes	yes
Auxin response factor 7	yes	yes	yes	yes	yes	yes
Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform	yes	yes	yes	yes	no	no
PHD and RING finger domain-containing protein 1	yes	no	yes	yes	yes	no
Linolenate hydroperoxide lyase, chloroplastic {ECO:0000305|PubMed:9701595}	yes	no	no	no	no	no
Nucleolar GTP-binding protein 1	yes	no	yes	yes	yes	yes
High mobility group B protein 15	yes	yes	no	no	no	yes
BAG family molecular chaperone regulator 6 {ECO:0000303|Ref.3}	yes	yes	yes	yes	yes	no
Polyamine oxidase 4 {ECO:0000303|PubMed:21796433}	yes	yes	yes	yes	yes	no
4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (ferredoxin), chloroplastic	yes	no	no	no	no	no
Tyrosine–tRNA ligase 1, cytoplasmic {ECO:0000305}	yes	no	yes	no	no	no
Dynamin-2B	yes	yes	yes	no	yes	yes
Serine/threonine-protein kinase SAPK5	yes	no	yes	no	no	no
6,7-dimethyl-8-ribityllumazine synthase, chloroplastic	yes	yes	yes	yes	yes	yes
Zinc finger CCCH domain-containing protein 19	yes	no	yes	yes	yes	yes
Phosphoenolpyruvate carboxylase 3	yes	yes	yes	yes	yes	yes
Acyl-coenzyme A oxidase 4, peroxisomal	yes	no	no	no	no	yes
Coatomer subunit beta'-1	yes	yes	yes	yes	yes	yes
Auxin response factor 18	yes	yes	no	yes	yes	yes
Protein STAY-GREEN LIKE, chloroplastic	yes	yes	yes	yes	yes	yes
bZIP transcription factor 60 {ECO:0000303|PubMed:18065552}	yes	yes	yes	yes	yes	yes
E3 UFM1-protein ligase 1 homolog	yes	no	yes	yes	yes	no
Protein PLASTID MOVEMENT IMPAIRED 1 {ECO:0000303|PubMed:16113226}	yes	yes	yes	no	yes	yes
Transcription factor PHYTOCHROME INTERACTING FACTOR-LIKE 13 {ECO:0000303|PubMed:17485859}	yes	no	no	no	no	no
Beta-glucosidase 5	yes	yes	yes	yes	yes	no
Cleavage stimulating factor 64 {ECO:0000303|PubMed:12379796}	yes	yes	no	no	yes	yes
Dihydrolipoyllysine-residue acetyltransferase component 4 of pyruvate dehydrogenase complex, chloroplastic	yes	no	yes	yes	yes	no
SWI/SNF complex subunit SWI3D	yes	yes	yes	no	yes	yes
Phosphoglycolate phosphatase 2	yes	no	yes	yes	yes	no
Protein SUPPRESSOR OF QUENCHING 1, chloroplastic {ECO:0000303|PubMed:23818601}	yes	no	yes	no	no	no
Sn1-specific diacylglycerol lipase alpha	yes	yes	yes	yes	yes	yes
APETALA2-like protein 3 {ECO:0000303|PubMed:28066457}	yes	no	no	no	no	no
IQ domain-containing protein IQM4 {ECO:0000305}	yes	yes	yes	yes	yes	yes
26S proteasome regulatory subunit 7	yes	no	no	no	no	no
Cytochrome P450 89A9	yes	no	no	no	no	no
OVARIAN TUMOR DOMAIN-containing deubiquitinating enzyme 11 {ECO:0000303|PubMed:24659992}	yes	no	no	no	yes	no
PLASMODESMATA CALLOSE-BINDING PROTEIN 4	yes	no	no	no	no	yes
Lysine-specific demethylase JMJ25	yes	yes	yes	yes	no	yes
Heat stress transcription factor A-2d	yes	yes	yes	yes	yes	no
Putative disease resistance protein RGA4	yes	no	no	yes	yes	yes
Myb-related protein 1 {ECO:0000303|PubMed:12008900}	yes	no	yes	yes	yes	yes

ATP-dependent zinc metalloprotease FTSH 5, mitochondrial	yes	yes	yes	yes	yes	yes
30S ribosomal protein 2, chloroplastic {ECO:0000303|PubMed:10874039}	yes	no	no	no	no	no
Replication factor C subunit 4	yes	no	no	no	no	no
Heme oxygenase 1, chloroplastic	yes	no	no	no	no	yes
Cyclin-T1-1	yes	yes	no	yes	no	no
Probable aminotransferase ACS12	yes	yes	yes	yes	yes	yes
Receptor-like protein 52 {ECO:0000303|PubMed:18434605}	yes	yes	yes	yes	no	yes
Putative disease resistance protein At3g14460	yes	no	yes	yes	yes	yes
Stress enhanced protein 1, chloroplastic {ECO:0000305}	yes	yes	yes	yes	yes	no
Myosin-2	yes	yes	yes	yes	yes	yes
Sucrose nonfermenting 4-like protein	yes	no	no	yes	no	no
Ultraviolet-B receptor UVR8	yes	yes	yes	yes	yes	yes
Probable protein arginine N-methyltransferase 6.1	yes	no	no	no	no	no
Proline-rich receptor-like protein kinase PERK9	yes	yes	yes	yes	yes	yes
DEAD-box ATP-dependent RNA helicase 25	yes	no	no	no	no	no
Probable xyloglucan endotransglucosylase/hydrolase protein 30	yes	no	no	no	no	no
F-box protein At1g55000	yes	yes	yes	yes	no	no
Embryogenesis-associated protein EMB8	yes	no	yes	yes	yes	no
Tyrosine-protein phosphatase DSP1 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Probable nucleoredoxin 2	yes	yes	no	yes	yes	yes
Calmodulin calcium-dependent NAD kinase {ECO:0000305}	yes	no	no	no	no	no
E3 ubiquitin-protein ligase HERC2	yes	yes	no	yes	yes	no
Protein SEMI-ROLLED LEAF 2 {ECO:0000303|PubMed:26873975}	yes	yes	yes	yes	yes	yes
Putative serine/threonine-protein kinase-like protein CCR3	yes	no	no	no	no	no
Serine carboxypeptidase-like 51	yes	no	no	no	no	no
Protein SRG1	yes	yes	yes	yes	no	no
CBS domain-containing protein CBSCBSPB3	yes	yes	yes	yes	yes	no
Proline transporter 1	yes	yes	no	yes	no	yes
NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial	yes	yes	yes	no	no	no
SUPPRESSOR OF GAMMA RESPONSE 1 {ECO:0000303|PubMed:19549833}	yes	yes	yes	yes	no	no
Calcium sensing receptor, chloroplastic	yes	no	yes	yes	yes	yes
Sulfite exporter TauE/SafE family protein 3 {ECO:0000312|EMBL:AEC07746.1}	yes	yes	yes	yes	yes	yes
Molybdopterin biosynthesis protein CNX1	yes	yes	yes	yes	yes	yes
Conserved oligomeric Golgi complex subunit 5 {ECO:0000303|PubMed:27448097}	yes	no	no	no	no	no
Phospholipase D delta {ECO:0000303|PubMed:11891260}	yes	no	no	no	yes	yes
Uncharacterized PKHD-type hydroxylase At1g22950	yes	no	no	no	no	no
Trimethyltridecatetraene synthase {ECO:0000303|PubMed:27662898}	yes	yes	no	yes	no	no
Protein NO VEIN {ECO:0000303|PubMed:19880797}	yes	no	yes	yes	no	yes
GDSL esterase/lipase At5g45910	yes	yes	no	yes	yes	yes
Transposon Ty3-G Gag-Pol polyprotein	yes	yes	no	no	no	yes
Disease resistance protein RGA4 {ECO:0000305}	yes	no	yes	yes	yes	yes
ETHYLENE INSENSITIVE 3-like 3 protein	yes	no	no	no	no	no
Argininosuccinate synthase, chloroplastic	yes	no	yes	no	no	no
NADH-cytochrome b5 reductase-like protein	yes	no	yes	no	yes	no
Switch 2 {ECO:0000312|EMBL:AEE27606.1}	yes	yes	yes	no	no	yes
Probable choline kinase 2	yes	yes	no	no	yes	yes
Glucose-induced degradation protein 4 homolog	yes	yes	yes	yes	yes	yes
Mitochondrial proton/calcium exchanger protein {ECO:0000305}	yes	no	no	no	no	no
Two-component response regulator-like PRR95	yes	no	no	yes	no	yes
Hypersensitive-induced reaction 1 protein {ECO:0000303|PubMed:20507517}	yes	yes	no	yes	yes	no
Aspartyl protease family protein At5g10770	yes	no	no	no	no	no

N6-mAMP deaminase {ECO:0000303|PubMed:29884623}	yes	yes	yes	no	no	yes
Binding partner of ACD11 1 {ECO:0000303|PubMed:18845362}	yes	no	yes	yes	yes	yes
XIAP-associated factor 1	yes	yes	no	yes	no	no
Calcium-transporting ATPase 10, plasma membrane-type {ECO:0000305}	yes	yes	yes	yes	yes	no
Alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase	yes	yes	no	yes	yes	yes
Splicing factor 3B subunit 2	yes	yes	yes	no	yes	yes
DNA-directed RNA polymerase II subunit RPB2	yes	yes	yes	yes	yes	yes
Protein mago nashi homolog 1 {ECO:0000305}	yes	no	no	no	no	no
Cell division cycle protein 48 homolog	yes	no	yes	no	no	no
RNA polymerase sigma factor sigA {ECO:0000303|PubMed:9421493}	yes	yes	yes	yes	no	yes
Probable mediator of RNA polymerase II transcription subunit 26b	yes	yes	yes	yes	yes	yes
Photosynthetic NDH subunit of lumenal location 4, chloroplastic {ECO:0000303|PubMed:21785130}	yes	no	no	no	yes	no
Protein indeterminate-domain 12 {ECO:0000303|PubMed:16784536}	yes	yes	yes	yes	yes	yes
RuBisCO large subunit-binding protein subunit alpha, chloroplastic	yes	yes	yes	yes	yes	yes
Probable protein phosphatase 2C 34	yes	no	yes	yes	no	yes
Transcription initiation factor TFIID subunit 6	yes	yes	yes	yes	yes	yes
Protein PHOSPHATE STARVATION RESPONSE 1 {ECO:0000250|UniProtKB:Q10LZ1}	yes	yes	yes	yes	yes	yes
Protein DETOXIFICATION 40 {ECO:0000303|PubMed:11739388}	yes	yes	no	yes	yes	yes
Protein DETOXIFICATION 27 {ECO:0000303|PubMed:11739388}	yes	yes	yes	yes	yes	yes
Importin subunit alpha-1b	yes	yes	yes	yes	yes	yes
Dual specificity protein phosphatase PHS1	yes	no	no	no	no	yes
Cytochrome P450 71A22	yes	yes	no	yes	yes	yes
Cytosolic invertase 1 {ECO:0000303|PubMed:18317796}	yes	no	no	no	no	no
Nudix hydrolase 8	yes	yes	yes	yes	yes	yes
Dehydrogenase/reductase SDR family member 7	yes	yes	yes	yes	yes	no
L-lactate dehydrogenase	yes	no	no	yes	no	yes
Receptor-like protein kinase FERONIA	yes	yes	yes	yes	yes	yes
1-aminocyclopropane-1-carboxylate oxidase homolog 1	yes	no	yes	yes	yes	no
Flowering-promoting factor 1-like protein 2	yes	yes	yes	yes	yes	no
Protein NRT1/ PTR FAMILY 2.11	yes	yes	yes	yes	yes	yes
Protein DWARF 53-LIKE {ECO:0000303|PubMed:24336200}	yes	yes	no	yes	yes	no
Apoptosis-inducing factor homolog A	yes	yes	yes	yes	yes	no
Leucine–tRNA ligase, chloroplastic/mitochondrial {ECO:0000305}	yes	no	yes	no	yes	no
Serine/threonine-protein phosphatase PP2A-2 catalytic subunit	yes	no	no	no	yes	yes
Protein MITOFERRINLIKE 1, chloroplastic	yes	yes	no	yes	no	no
3-methyl-2-oxobutanoate hydroxymethyltransferase 1, mitochondrial	yes	yes	yes	no	yes	no
Homeobox-leucine zipper protein HOX16	yes	no	no	no	no	yes
Two-component response regulator ORR1 {ECO:0000305}	yes	no	no	no	yes	yes
Probable ethylene response sensor 2 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Phosphoglycerate kinase, cytosolic	yes	no	no	no	no	no
Phosphoenolpyruvate carboxylase kinase 2	yes	yes	yes	yes	yes	no
Wall-associated receptor kinase 4	yes	no	yes	no	no	no
Post-GPI attachment to proteins factor 3	yes	no	no	yes	yes	yes
Receptor protein kinase CLAVATA1	yes	yes	yes	no	yes	no
Actin-depolymerizing factor 3	yes	no	yes	no	no	no
Protein SMG7 {ECO:0000303|PubMed:18544632}	yes	no	yes	no	yes	yes
RNA-binding protein 39	yes	no	yes	yes	yes	no
ATP synthase subunit a	yes	yes	yes	yes	yes	no
Vacuolar protein sorting-associated protein 2 homolog 3	yes	yes	yes	yes	yes	yes

Pentatricopeptide repeat-containing protein At3g29230	yes	yes	yes	yes	yes	yes
Phospholipase A1-Igamma1, chloroplastic	yes	no	no	no	yes	no
ATP synthase protein MI25	yes	yes	yes	yes	no	no
Beta-glucosidase 31	yes	yes	yes	yes	yes	yes
Mediator of RNA polymerase II transcription subunit 33A	yes	no	no	no	no	yes
Probable phospholipid hydroperoxide glutathione peroxidase	yes	yes	yes	yes	yes	yes
30S ribosomal protein S13, chloroplastic {ECO:0000303|PubMed:10874039}	yes	no	yes	no	no	no
Transcription factor ILR3	yes	yes	no	yes	no	yes
Flavonol 3-O-glucosyltransferase UGT89B1 {ECO:0000305}	yes	yes	no	yes	yes	no
Cytochrome P450 89A2	yes	yes	yes	yes	no	yes
OBERON-like protein	yes	yes	yes	yes	yes	yes
Protein OBERON 2 {ECO:0000303|PubMed:18403411}	yes	yes	yes	yes	yes	no
Uncharacterized aarF domain-containing protein kinase At5g05200, chloroplastic	yes	yes	yes	yes	yes	no
Phosphatidylinositol 4-kinase gamma 4 {ECO:0000305}	yes	yes	yes	no	no	no
DDT domain-containing protein PTM {ECO:0000305}	yes	no	no	no	no	no
UPF0014 membrane protein STAR2 {ECO:0000305}	yes	yes	no	yes	yes	yes
Uncharacterized protein At4g14100	yes	yes	yes	yes	yes	yes
Tubulin beta-4 chain	yes	no	no	no	no	no
26S proteasome non-ATPase regulatory subunit 1 homolog A	yes	yes	yes	yes	no	yes
3-hydroxy-3-methylglutaryl-coenzyme A reductase	yes	no	no	yes	no	yes
YTH domain-containing protein ECT2 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Phenylalanine–tRNA ligase, chloroplastic/mitochondrial {ECO:0000305}	yes	yes	yes	yes	yes	yes
ABC transporter G family member 36 {ECO:0000305}	yes	no	yes	no	yes	yes
Pre-mRNA-processing factor 39	yes	yes	yes	yes	yes	no
Vacuolar-sorting receptor 3	yes	yes	no	yes	no	no
Calreticulin	yes	no	yes	no	no	no
Protein REVEILLE 5	yes	no	no	no	no	no
Protein REVEILLE 6	yes	no	yes	yes	yes	no
Probable aldo-keto reductase 2	yes	yes	yes	yes	yes	yes
RNA-directed DNA methylation 4	yes	yes	yes	yes	yes	yes
Heavy metal-associated isoprenylated plant protein 39 {ECO:0000303|PubMed:21072340, ECO:0000303|PubMed:23368984}	yes	no	yes	no	yes	yes
Splicing factor U2af large subunit A	yes	no	no	no	no	no
Ubinuclein-1 {ECO:0000303|PubMed:25086063}	yes	yes	yes	yes	no	no
Phenylalanine ammonia-lyase	yes	yes	yes	yes	yes	yes
Transmembrane 9 superfamily member 7 {ECO:0000305}	yes	no	no	no	no	no
Monothiol glutaredoxin-S11	yes	yes	yes	yes	yes	yes
Probable glutathione peroxidase 2	yes	no	no	yes	no	yes
Serine/threonine protein phosphatase 2A 55 kDa regulatory subunit B beta isoform	yes	no	no	no	no	yes
Protein MALE DISCOVERER 2	yes	yes	no	no	yes	no
Uncharacterized protein At3g06530	yes	no	no	yes	no	yes
Adenine nucleotide transporter BT1, chloroplastic/mitochondrial	yes	yes	yes	no	no	no
Glutamate receptor 2.8	yes	no	yes	yes	yes	yes
Pentatricopeptide repeat-containing protein At1g50270	yes	yes	yes	yes	yes	no
Acyl-CoA-binding domain-containing protein 5 {ECO:0000305}	yes	no	no	no	no	no
ABC transporter G family member 22	yes	no	no	no	no	no
Transcription factor UNE10	yes	yes	yes	yes	yes	yes
FCS-Like Zinc finger 11 {ECO:0000303|PubMed:24901469}	yes	yes	yes	yes	no	yes
Autophagy-related protein 13b {ECO:0000303|PubMed:12114572}	yes	yes	yes	yes	yes	no
Serine/threonine-protein kinase STY8 {ECO:0000305}	yes	yes	no	no	yes	yes

SH3 domain-containing protein PJ696.02	yes	yes	yes	yes	no	yes
Protein CHROMATIN REMODELING 5 {ECO:0000303|PubMed:16547115}	yes	no	no	yes	no	no
Kinesin-like protein KIN-4A {ECO:0000305}	yes	no	no	no	no	yes
Monosaccharide-sensing protein 2	yes	no	yes	yes	yes	no
Peroxiredoxin-2F, mitochondrial	yes	no	no	no	no	no
Uncharacterized isomerase BH0283	yes	yes	no	no	no	yes
Protein SCO1 homolog 1, mitochondrial	yes	no	no	no	no	no
Protein ACTIVITY OF BC1 COMPLEX KINASE 8, chloroplastic {ECO:0000303|PubMed:22694836}	yes	no	no	no	no	no
E3 ubiquitin-protein ligase KEG	yes	no	yes	yes	no	no
K(+) efflux antiporter 3, chloroplastic {ECO:0000303|PubMed:11500563}	yes	yes	yes	yes	yes	yes
Phosphatidylinositol 4-kinase gamma 7	yes	yes	yes	yes	yes	no
Squamous cell carcinoma antigen recognized by T-cells 3 {ECO:0000305}	yes	yes	yes	yes	no	no
Probable feruloyl esterase A	yes	yes	yes	yes	yes	no
Beta-glucosidase 18	yes	no	yes	yes	yes	yes
Putative B3 domain-containing protein Os04g0346900	yes	no	no	no	no	no
Phosphoribosylamine–glycine ligase, chloroplastic	yes	yes	yes	yes	yes	no
Probable 1-acylglycerol-3-phosphate O-acyltransferase	yes	no	no	no	no	no
F-box/LRR-repeat protein 3	yes	yes	yes	no	no	no
Transcription factor GTE10	yes	no	yes	no	no	no
Fe(2+) transport protein 1	yes	yes	yes	yes	yes	yes
ABC transporter C family member 3	yes	yes	yes	yes	yes	yes
E3 ubiquitin-protein ligase UPL1	yes	no	yes	no	no	no
Transcription factor TGA2.2 {ECO:0000305}	yes	no	no	no	no	no
Rhodanese-like domain-containing protein 9, chloroplastic	yes	yes	yes	yes	yes	yes
Obtusifoliol 14-alpha demethylase	yes	yes	yes	yes	yes	yes
NAD(P)H-quinone oxidoreductase subunit L, chloroplastic	yes	no	no	no	no	yes
Aspartic proteinase	yes	yes	no	yes	yes	no
Ocs element-binding factor 1	yes	yes	yes	no	no	no
Phosphoribulokinase, chloroplastic	yes	no	no	no	yes	yes
G-type lectin S-receptor-like serine/threonine-protein kinase LECRK4 {ECO:0000305}	yes	yes	no	yes	no	no
Inositol-3-phosphate synthase	yes	yes	yes	yes	yes	no
4-coumarate–CoA ligase-like 9	yes	no	yes	yes	yes	no
Transcription factor MYB30 {ECO:0000303|PubMed:10929106}	yes	no	yes	no	no	no
DExH-box ATP-dependent RNA helicase DExH6 {ECO:0000305}	yes	no	no	no	no	no
Uncharacterized protein At2g33490	yes	yes	yes	no	no	yes
2,3-bisphosphoglycerate-independent phosphoglycerate mutase {ECO:0000255|HAMAP-Rule:MF_01402}	yes	yes	yes	no	no	no
L-aspartate oxidase, chloroplastic	yes	no	no	no	no	yes
Meiotic nuclear division protein 1 homolog	yes	yes	no	no	no	no
Probable 6-phosphogluconolactonase 4, chloroplastic	yes	yes	yes	yes	yes	no
Protein PHYTOCHROME-DEPENDENT LATE-FLOWERING {ECO:0000303|PubMed:24127609}	yes	yes	yes	yes	no	no
Peptide chain release factor PrfB3, chloroplastic {ECO:0000303|PubMed:21771930}	yes	yes	yes	yes	yes	no
Putative zinc transporter At3g08650	yes	no	yes	no	no	no
NADP-dependent malic enzyme	yes	yes	yes	yes	yes	yes
Uncharacterized methyltransferase At2g41040, chloroplastic	yes	yes	yes	yes	no	no
Lecithin-cholesterol acyltransferase-like 1	yes	no	yes	no	no	no
Phosphoinositide phospholipase C 2	yes	no	yes	no	no	no

Callose synthase 10	yes	no	no	no	yes	no
Putative disease resistance RPP13-like protein 2	yes	yes	no	no	no	no
Basic leucine zipper 23 {ECO:0000303|PubMed:11906833}	yes	no	no	no	no	no
WD repeat-containing protein 20	yes	yes	yes	yes	yes	yes
Receptor kinase-like protein Xa21 {ECO:0000303|PubMed:22735448}	yes	yes	yes	yes	yes	yes
Protein LYK5	yes	no	yes	yes	no	no
Thylakoid lumenal 15 kDa protein 1, chloroplastic	yes	yes	yes	yes	yes	yes
Nuclear/nucleolar GTPase 2 {ECO:0000303|PubMed:21205822}	yes	yes	yes	yes	yes	yes
Probable tRNA N6-adenosine threonylcarbamoyltransferase, mitochondrial {ECO:0000255|HAMAP-Rule:MF_03179}	yes	no	no	yes	yes	no
Tetratricopeptide repeat protein SKI3 {ECO:0000305}	yes	yes	no	yes	yes	no
E3 ubiquitin ligase PQT3-like	yes	yes	yes	yes	no	yes
40S ribosomal protein S13	yes	yes	yes	yes	yes	yes
Clathrin interactor EPSIN 1	yes	no	yes	no	yes	no
Probable aldehyde oxidase 2	yes	no	no	no	no	no
PsbP domain-containing protein 1, chloroplastic	yes	no	yes	yes	yes	yes
LysM and putative peptidoglycan-binding domain-containing protein 1	yes	no	no	no	yes	yes
Probable inactive serine/threonine-protein kinase scy1	yes	yes	yes	yes	yes	yes
Glucose-6-phosphate/phosphate translocator 2, chloroplastic	yes	yes	yes	yes	yes	yes
Transcription factor bHLH128	yes	yes	yes	yes	yes	yes
Cytochrome P450 98A1	yes	yes	yes	yes	yes	no
Protein BIC1 {ECO:0000303|PubMed:27846570}	yes	yes	yes	yes	yes	no
Peroxisome biogenesis protein 12	yes	no	no	yes	yes	no
Auxin transport protein BIG	yes	yes	yes	yes	no	no
Oligouridylate-binding protein 1	yes	no	no	no	no	no
U-box domain-containing protein 33	yes	no	no	no	yes	no
L-type lectin-domain containing receptor kinase SIT2 {ECO:0000305}	yes	no	no	yes	no	no
ADP-ribosylation factor	yes	no	no	no	no	no
Leucine aminopeptidase 2, chloroplastic	yes	yes	yes	yes	yes	yes
Dof zinc finger protein DOF5.8	yes	yes	yes	no	no	no
Glucose-6-phosphate 1-dehydrogenase, chloroplastic	yes	no	no	no	no	no
Probable ubiquitin-conjugating enzyme E2 25	yes	yes	yes	yes	yes	yes
Beta-1,3-galactosyltransferase 7	yes	yes	yes	yes	yes	yes
Prostaglandin reductase 3	yes	no	no	yes	yes	yes
Sucrose-phosphate synthase	yes	no	no	no	no	no
Hydroxymethylglutaryl-CoA synthase	yes	yes	yes	no	yes	yes
Transcription factor TDR {ECO:0000305}	yes	no	no	no	no	yes
Glutathione S-transferase T1	yes	yes	yes	yes	yes	no
Tetraspanin-19	yes	no	no	no	no	no
Serine/threonine-protein kinase VPS15 {ECO:0000303|PubMed:21833541}	yes	no	no	no	no	no
ATP-dependent RNA helicase DEAH12, chloroplastic	yes	yes	yes	yes	yes	no
Transcription factor TCP15 {ECO:0000305}	yes	no	no	no	yes	yes
Two pore potassium channel c	yes	yes	no	no	yes	yes
Zinc transporter ZTP29	yes	no	no	no	no	no
Serine/threonine-protein phosphatase 7 long form homolog	yes	yes	no	no	yes	yes
Guanylate-binding protein 7	yes	yes	yes	yes	yes	no
ERAD-associated E3 ubiquitin-protein ligase HRD1 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Probable inactive DNA (cytosine-5)-methyltransferase DRM3 {ECO:0000305}	yes	no	yes	no	no	no
Glutamyl-tRNA(Gln) amidotransferase subunit A {ECO:0000255|HAMAP-Rule:MF_00120}	yes	no	no	yes	no	no
Cytochrome b5 isoform E {ECO:0000303|PubMed:19054355}	yes	no	yes	no	yes	no

F-box protein FBX14	yes	no	yes	yes	yes	no
Zinc-finger homeodomain protein 9	yes	no	no	no	no	yes
E3 ubiquitin-protein ligase RZFP34 {ECO:0000305}	yes	no	no	no	no	yes
Calcium-dependent protein kinase 9 {ECO:0000305}	yes	yes	yes	yes	yes	no
Transcription factor-like protein DPB	yes	no	yes	no	no	no
Farnesyl pyrophosphate synthase 1, mitochondrial	yes	no	no	no	no	yes
Protein GrpE {ECO:0000255|HAMAP-Rule:MF_01151}	yes	yes	yes	no	yes	yes
PHD finger protein rhinoceros	yes	yes	yes	no	no	yes
DNA polymerase delta subunit 3	yes	yes	yes	no	yes	yes
Probable sulfate transporter 3.3	yes	no	no	yes	no	no
Protein TOPLESS-RELATED PROTEIN 2 {ECO:0000303|PubMed:24336200}	yes	yes	yes	yes	yes	no
2-hydroxyisoflavanone dehydratase	yes	no	no	no	no	no
Interactor of constitutive active ROPs 2, chloroplastic	yes	no	no	no	no	no
Cyclic nucleotide-gated ion channel 1	yes	no	no	yes	yes	yes
Common plant regulatory factor 1	yes	yes	yes	yes	yes	yes
Cell division cycle 5-like protein	yes	no	no	no	yes	no
MLO protein homolog 1	yes	no	no	no	no	yes
Putative ribonuclease H protein At1g65750	yes	yes	no	yes	no	no
Protein DETOXIFICATION 16 {ECO:0000303|PubMed:11739388}	yes	no	yes	yes	yes	yes
Probable serine/threonine-protein kinase PBL8 {ECO:0000305}	yes	yes	no	yes	no	no
Probable ribose-5-phosphate isomerase 4, chloroplastic	yes	yes	yes	yes	yes	yes
EIN3-binding F-box protein 1	yes	yes	yes	yes	no	no
Catalase isozyme 3	yes	no	no	yes	yes	yes
MADS-box transcription factor 47 {ECO:0000305}	yes	yes	yes	yes	yes	yes
Factor of DNA methylation 1 {ECO:0000303|PubMed:22302148}	yes	no	yes	yes	yes	no
50S ribosomal protein L18, chloroplastic	yes	no	no	no	no	no
TSET complex member tstF {ECO:0000305}	yes	yes	yes	no	yes	yes
Fra a 1-associated protein {ECO:0000303|PubMed:28656626}	yes	no	yes	no	yes	no
Pentatricopeptide repeat-containing protein At1g74850, chloroplastic	yes	yes	yes	no	yes	no
Serine carboxypeptidase 1	yes	no	no	no	yes	no
Probable inactive purple acid phosphatase 1	yes	yes	no	yes	yes	no
F-box protein SKIP5	yes	yes	yes	no	yes	yes
Mitochondrial-processing peptidase subunit alpha	yes	yes	yes	yes	yes	no
Protein TIME FOR COFFEE	yes	yes	yes	yes	yes	yes
Probable methyltransferase PMT26	yes	yes	no	yes	yes	no
Zinc finger protein ZPR1	yes	no	yes	yes	no	no
Sm-like protein LSM8 {ECO:0000305}	yes	no	no	no	no	no
Glutamine synthetase root isozyme 1	yes	no	no	no	no	no
Probable allantoinase	yes	yes	yes	yes	yes	no
Probable indole-3-acetic acid-amido synthetase GH3.8	yes	yes	yes	no	yes	yes
Myb-related protein P	yes	yes	yes	yes	yes	no
Trans-resveratrol di-O-methyltransferase	yes	no	yes	yes	yes	yes
WEB family protein At5g16730, chloroplastic	yes	no	no	yes	no	no
Transcriptional corepressor LEUNIG_HOMOLOG	yes	yes	no	no	no	no
BTB/POZ domain-containing protein At1g55760	yes	no	no	no	no	no
Chaperone protein dnaJ GFA2, mitochondrial {ECO:0000305}	yes	yes	yes	yes	no	yes
Serine/threonine-protein phosphatase 4 regulatory subunit 3	yes	no	no	no	yes	yes
FCS-Like Zinc finger 10 {ECO:0000303|PubMed:24901469}	yes	yes	no	no	no	no
Zinc finger protein ZAT18 {ECO:0000303|PubMed:28586434}	yes	no	no	no	no	no
TATA box-binding protein-associated factor RNA polymerase I subunit B	yes	no	no	no	no	no
Serine/threonine-protein kinase-like protein CCR4	yes	no	no	no	yes	no

HVA22-like protein f	yes	no	no	no	no	no
Carbamoyl-phosphate synthase large chain, chloroplastic	yes	no	no	no	no	no
40S ribosomal protein S4	yes	no	no	no	no	no
Protein argonaute 4A	yes	yes	no	no	no	no
Trigger factor {ECO:0000255|HAMAP-Rule:MF_00303}	yes	no	no	no	no	no
Aldehyde dehydrogenase family 2 member C4	yes	no	no	no	no	no
Sucrose transport protein SUT4	yes	no	no	no	no	no
DnaJ homolog subfamily C member 7	yes	no	no	no	no	no
Homeobox-leucine zipper protein HOX32	yes	no	no	no	no	no
DIBOA-glucoside dioxygenase BX6	yes	no	no	no	no	no
Bark storage protein A	yes	no	no	no	no	no
Transcription factor HEC1	yes	no	no	no	no	no
Beta-glucuronosyltransferase GlcAT14B {ECO:0000305}	yes	no	no	no	no	no
Ethylene-responsive transcription factor 8	yes	no	yes	no	no	no
Mediator of RNA polymerase II transcription subunit 12	no	yes	yes	yes	no	no
Protein STAY-GREEN, chloroplastic	no	yes	yes	yes	yes	no
NAD-dependent protein deacylase SRT2 {ECO:0000255|HAMAP-Rule:MF_03161}	no	yes	yes	yes	yes	no
Squamosa promoter-binding-like protein 12	no	yes	yes	yes	yes	yes
Target of rapamycin complex subunit LST8 {ECO:0000305}	no	yes	yes	no	no	no
Transcription factor MYB60 {ECO:0000303|PubMed:18647406}	no	yes	yes	yes	no	no
Tubby-like F-box protein 3	no	yes	no	no	yes	no
U3 small nucleolar ribonucleoprotein protein IMP3	no	yes	yes	yes	no	no
Serine/threonine/tyrosine-protein kinase HT1 {ECO:0000305}	no	yes	no	no	no	no
Alliin lyase	no	yes	no	no	yes	yes
Probable periplasmic serine protease do/HhoA-like	no	yes	yes	yes	no	no
Zinc finger CCCH domain-containing protein 44	no	yes	yes	no	no	no
Putative anthocyanidin reductase {ECO:0000303|PubMed:16399014}	no	yes	yes	yes	yes	no
DEAD-box ATP-dependent RNA helicase 22	no	yes	yes	no	yes	no
30S ribosomal protein S1, chloroplastic {ECO:0000305}	no	yes	yes	yes	yes	yes
Formin-like protein 5	no	yes	no	yes	yes	no
Magnesium/proton exchanger 1	no	yes	no	no	yes	no
Pre-mRNA-splicing factor 18	no	yes	no	yes	yes	no
26S proteasome non-ATPase regulatory subunit 11 homolog	no	yes	no	no	no	no
Protein FORGETTER 1 {ECO:0000303|PubMed:27680998}	no	yes	no	yes	yes	yes
U4/U6 small nuclear ribonucleoprotein Prp31 homolog {ECO:0000305}	no	yes	yes	yes	no	no
Stromal 70 kDa heat shock-related protein, chloroplastic	no	yes	yes	yes	yes	yes
Protein DETOXIFICATION 20 {ECO:0000303|PubMed:11739388}	no	yes	no	no	yes	no
Septin and tuftelin-interacting protein 1 homolog 1 {ECO:0000303|PubMed:23110899}	no	yes	no	yes	no	no
Ras-related protein RABA5a	no	yes	yes	no	no	no
Protein Iojap, chloroplastic	no	yes	no	no	no	no
Signal recognition particle 54 kDa protein, chloroplastic	no	yes	no	no	no	no
Fatty acid amide hydrolase	no	yes	yes	yes	yes	yes
Uncharacterized protein At5g41620	no	yes	no	yes	no	no
Cytochrome P450 84A1	no	yes	yes	yes	no	no
Probable magnesium transporter NIPA3	no	yes	no	yes	yes	no
NAC domain-containing protein 30 {ECO:0000303|PubMed:15029955}	no	yes	no	no	yes	no
GTP-binding protein At2g22870	no	yes	yes	yes	yes	yes
Ubiquitin domain-containing protein 2	no	yes	no	no	no	no
Dynamin-related protein 5A	no	yes	yes	no	no	no

ATP-dependent DNA helicase RecQ	no	yes	yes	yes	yes	no
Disease resistance protein RGA2	no	yes	yes	yes	yes	yes
eEF1A lysine and N-terminal methyltransferase {ECO:0000250|UniProtKB:Q8N6R0}	no	yes	yes	no	yes	no
Monodehydroascorbate reductase 3, cytosolic {ECO:0000305}	no	yes	yes	no	no	no
Carboxyl-terminal-processing peptidase 3, chloroplastic	no	yes	no	no	no	no
Pyrophosphate-energized membrane proton pump 3	no	yes	no	no	no	no
Probable anion transporter 1, chloroplastic	no	yes	no	no	no	no
Protein MODIFIED TRANSPORT TO THE VACUOLE 1 {ECO:0000303|PubMed:23771894}	no	yes	yes	no	no	no
E3 ubiquitin-protein ligase SINAT3 {ECO:0000305}	no	yes	no	no	no	no
Probable histone acetyltransferase HAC-like 1	no	yes	no	no	no	no
Cryptochrome-1 {ECO:0000303|PubMed:8953250}	no	yes	yes	yes	yes	yes
DNA-directed RNA polymerases II, IV and V subunit 11	no	yes	yes	no	yes	yes
Nucleolar complex protein 2 homolog	no	yes	yes	yes	no	no
E3 ubiquitin-protein ligase UPL4	no	yes	yes	yes	yes	no
Protein HIGH CHLOROPHYLL FLUORESCENCE PHENOTYPE 244, chloroplastic {ECO:0000303|PubMed:23027666}	no	yes	no	no	no	no
Lipoyl synthase, mitochondrial {ECO:0000255|HAMAP-Rule:MF_03128}	no	yes	yes	yes	no	no
Tuliposide A-converting enzyme b3, amyloplastic	no	yes	yes	no	no	no
IST1-like protein	no	yes	no	yes	yes	no
Peptidyl-prolyl cis-trans isomerase FKBP16-4, chloroplastic	no	yes	yes	no	yes	no
E3 ubiquitin-protein ligase UPL6	no	yes	no	yes	no	no
Acetylserotonin O-methyltransferase 2 {ECO:0000305}	no	yes	no	no	no	no
RNA cytidine acetyltransferase 1 {ECO:0000255|HAMAP-Rule:MF_03211}	no	yes	yes	yes	yes	no
Amino acid transporter AVT6A {ECO:0000305}	no	yes	yes	yes	no	no
Vacuolar cation/proton exchanger 3	no	yes	no	yes	no	no
Expansin-like B1	no	yes	no	no	no	no
Small RNA 2'-O-methyltransferase	no	yes	yes	no	no	no
ABC transporter C family member 13	no	yes	no	no	no	yes
SWI/SNF complex subunit SWI3A	no	yes	no	no	yes	no
PH, RCC1 and FYVE domains-containing protein 1 {ECO:0000303|PubMed:11563980}	no	yes	yes	yes	no	no
Cyclin-dependent kinase E-1	no	yes	yes	yes	yes	no
Protein ENHANCED DISEASE RESISTANCE 2-like	no	yes	no	no	no	no
Bifunctional fucokinase/fucose pyrophosphorylase	no	yes	yes	yes	no	no
Putative magnesium transporter MRS2-G	no	yes	no	no	no	no
SPX domain-containing membrane protein Os06g0129400	no	yes	yes	no	yes	no
Selenium-binding protein 2	no	yes	no	no	no	yes
Chaperone protein DnaJ {ECO:0000255|HAMAP-Rule:MF_01152}	no	yes	no	no	yes	no
E3 ubiquitin protein ligase RIN2	no	yes	yes	yes	yes	no
Probable galacturonosyltransferase 9	no	yes	no	yes	no	no
Protein MEI2-like 4	no	yes	no	no	no	yes
Probable ubiquitin-conjugating enzyme E2 26	no	yes	yes	yes	yes	no
Transcriptional corepressor SEUSS	no	yes	yes	yes	yes	yes
Protein NRT1/ PTR FAMILY 5.2	no	yes	yes	no	no	no
SPX domain-containing membrane protein Os02g45520	no	yes	yes	yes	yes	yes
Ribosomal RNA-processing protein 8	no	yes	no	yes	yes	no
Protein NRT1/ PTR FAMILY 8.1	no	yes	yes	yes	yes	no
Shaggy-related protein kinase gamma {ECO:0000303|PubMed:7509023}	no	yes	no	no	yes	no
Probable 1-acyl-sn-glycerol-3-phosphate acyltransferase 5	no	yes	no	no	yes	no

Protein indeterminate-domain 5, chloroplastic {ECO:0000303|PubMed:16784536}	no	yes	no	no	no	yes
Polyadenylate-binding protein 2-A	no	yes	no	no	no	no
3-ketoacyl-CoA synthase 6 {ECO:0000303|PubMed:18465198}	no	yes	no	no	no	no
Tubby-like F-box protein 1	no	yes	no	yes	yes	yes
Chaperone protein dnaJ 6	no	yes	yes	yes	no	yes
Auxin-responsive protein SAUR36	no	yes	no	no	no	no
Uroporphyrinogen decarboxylase 2, chloroplastic	no	yes	yes	no	yes	yes
RING-box protein 1A	no	yes	yes	no	no	yes
IQ domain-containing protein IQM3 {ECO:0000305}	no	yes	no	no	yes	yes
Bidirectional sugar transporter SWEET1 {ECO:0000303|PubMed:21107422}	no	yes	yes	no	no	no
Cyclin-B2-1	no	yes	yes	yes	yes	yes
Ubiquitin-like modifier-activating enzyme atg7	no	yes	no	no	no	no
Protein EMBRYO DEFECTIVE 1674 {ECO:0000303|PubMed:15266054}	no	yes	no	yes	no	no
VIN3-like protein 1	no	yes	yes	no	yes	no
Acyl-coenzyme A thioesterase 8	no	yes	yes	no	no	no
WD repeat-containing protein GTS1 {ECO:0000305}	no	yes	no	no	no	no
Glycine–tRNA ligase, chloroplastic/mitochondrial 2 {ECO:0000305}	no	yes	yes	yes	yes	no
Cinnamoyl-CoA reductase 1 {ECO:0000305}	no	yes	yes	yes	no	no
Microtubule-associated protein RP/EB family member 1C	no	yes	no	yes	yes	no
TBC1 domain family member 15	no	yes	yes	yes	yes	no
Protoheme IX farnesyltransferase, mitochondrial	no	yes	yes	yes	yes	no
DExH-box ATP-dependent RNA helicase DExH13 {ECO:0000305}	no	yes	no	yes	yes	no
Probable phosphoinositide phosphatase SAC9	no	yes	yes	yes	yes	yes
Splicing factor U2AF-associated protein 2	no	yes	no	no	no	yes
Eukaryotic translation initiation factor 3 subunit D {ECO:0000255|HAMAP-Rule:MF_03003}	no	yes	no	no	no	no
Methionine S-methyltransferase	no	yes	yes	yes	yes	yes
Ubiquitin-conjugating enzyme E2 2	no	yes	no	no	no	yes
ABC transporter C family member 10	no	yes	no	no	no	no
G-type lectin S-receptor-like serine/threonine-protein kinase At1g11330	no	yes	no	yes	yes	no
NAC domain-containing protein 35 {ECO:0000303|PubMed:15029955}	no	yes	no	no	no	no
Hydroxymethylglutaryl-CoA lyase, mitochondrial	no	yes	no	no	no	yes
Protein YIF1B-B	no	yes	yes	no	no	no
Probable aldo-keto reductase 1	no	yes	yes	yes	no	no
Receptor-like protein kinase 7 {ECO:0000303|PubMed:20811905}	no	yes	yes	no	no	no
UPF0481 protein At3g47200	no	yes	yes	yes	yes	no
1-phosphatidylinositol-3-phosphate 5-kinase FAB1B	no	yes	yes	no	yes	no
Putative 1-phosphatidylinositol-3-phosphate 5-kinase FAB1D	no	yes	no	no	yes	no
DENN domain-containing protein 5B	no	yes	yes	yes	yes	no
DNA-binding protein BIN4	no	yes	no	yes	no	no
Cytochrome P450 93G2 {ECO:0000303|PubMed:20647377}	no	yes	no	yes	no	no
Non-specific lipid-transfer protein 2	no	yes	no	no	yes	no
Poly [ADP-ribose] polymerase 2	no	yes	no	yes	no	no
Protein TIC 62, chloroplastic	no	yes	yes	no	yes	no
F-box protein SKIP19	no	yes	yes	yes	yes	yes
Far upstream element-binding protein 1	no	yes	yes	yes	yes	no
Activator of 90 kDa heat shock protein ATPase homolog 2	no	yes	no	no	no	no
Sphingosine kinase 1	no	yes	yes	no	yes	no
NifU-like protein 3, chloroplastic	no	yes	yes	yes	yes	yes
protein SLOW GREEN 1, chloroplastic {ECO:0000303|PubMed:24420572}	no	yes	yes	no	no	yes
Transcription factor TCP5	no	yes	no	yes	no	no

CCR4-NOT transcription complex subunit 3	no	yes	no	yes	no	no
Programmed cell death protein 2	no	yes	no	no	no	no
1-deoxy-D-xylulose-5-phosphate synthase 1, chloroplastic	no	yes	yes	yes	yes	yes
Transcription factor MYBS3 {ECO:0000303|PubMed:12172034}	no	yes	yes	yes	yes	yes
Beta-fructofuranosidase 1	no	yes	no	yes	yes	no
Double-stranded RNA-binding protein 8	no	yes	no	no	no	no
Putative disease resistance protein RGA3	no	yes	yes	yes	yes	yes
Magnesium transporter MRS2-E	no	yes	yes	yes	yes	no
Probable LRR receptor-like serine/threonine-protein kinase At1g56130	no	yes	no	no	no	no
Protein transport protein Sec24-like At3g07100	no	yes	no	no	yes	no
Serine/threonine-protein kinase PBL27 {ECO:0000303|PubMed:20413097}	no	yes	yes	yes	no	yes
Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial	no	yes	yes	yes	no	no
PHD finger protein ALFIN-LIKE 7	no	yes	no	yes	no	no
TBC1 domain family member 13	no	yes	no	yes	no	no
E3 ubiquitin-protein ligase UPL7	no	yes	no	yes	no	no
Protein transport protein Sec24-like CEF	no	yes	yes	yes	yes	no
Structural maintenance of chromosomes protein 6B	no	yes	yes	yes	yes	yes
Zeaxanthin epoxidase, chloroplastic	no	yes	yes	yes	yes	no
Protein MODIFIER OF SNC1 1	no	yes	yes	no	no	no
Chlorophyllase-2, chloroplastic	no	yes	yes	yes	yes	yes
Ethylene-responsive transcription factor 11	no	yes	no	yes	no	no
Protein CHLORORESPIRATORY REDUCTION 6, chloroplastic {ECO:0000305}	no	yes	yes	yes	yes	no
Protein S-acyltransferase 11	no	yes	no	no	yes	no
Protein transport protein SEC16B homolog {ECO:0000305}	no	yes	no	no	no	no
Probable inactive beta-glucosidase 33	no	yes	yes	yes	yes	no
Germ cell-less protein-like 1	no	yes	yes	yes	no	no
Diacylglycerol kinase 1	no	yes	no	yes	yes	no
Sperm-associated antigen 1A	no	yes	no	no	no	no
Pyridoxal reductase, chloroplastic	no	yes	no	yes	no	no
Nuclear factor related to kappa-B-binding protein	no	yes	no	no	no	no
Premnaspirodiene oxygenase	no	yes	yes	no	no	no
Serine/threonine-protein kinase CTR1 {ECO:0000303|PubMed:8431946}	no	yes	yes	yes	no	yes
Putative UDP-rhamnose:rhamnosyltransferase 1	no	yes	yes	no	no	no
Valine–tRNA ligase, chloroplastic/mitochondrial 2 {ECO:0000305}	no	yes	yes	yes	yes	yes
NAC domain-containing protein 17 {ECO:0000303|PubMed:15029955}	no	yes	yes	yes	yes	no
Uncharacterized protein sll0005	no	yes	yes	yes	yes	no
Probable acylpyruvase FAHD2, mitochondrial {ECO:0000305}	no	yes	no	no	no	no
F-box protein PP2-A13	no	yes	yes	no	no	no
Disease resistance protein Piks-2 {ECO:0000305}	no	yes	no	yes	yes	yes
PsbQ-like protein 3, chloroplastic	no	yes	no	yes	yes	yes
Probable serine/threonine-protein kinase WNK1	no	yes	yes	yes	yes	yes
Light-harvesting complex-like protein OHP2, chloroplastic {ECO:0000305}	no	yes	yes	yes	no	no
AUGMIN subunit 3 {ECO:0000303|PubMed:22505726}	no	yes	no	no	no	no
Protein ETHYLENE-INSENSITIVE 3-like 1b {ECO:0000303|PubMed:16786297}	no	yes	yes	yes	yes	no
CCR4-NOT transcription complex subunit 4	no	yes	no	no	no	no
Protein Dr1 homolog	no	yes	no	yes	no	yes
R3H domain-containing protein 2	no	yes	yes	yes	no	yes
Glutamyl-tRNA reductase, chloroplastic	no	yes	no	yes	yes	yes
Disease resistance protein PIK5-NP {ECO:0000305}	no	yes	yes	no	yes	no
Pirin-like protein	no	yes	yes	yes	yes	yes

Glutamine-dependent NAD(+) synthetase	no	yes	no	no	no	yes
Phosphoinositide phosphatase SAC1	no	yes	no	no	no	no
Protein CCA1 {ECO:0000303|PubMed:9144958}	no	yes	no	yes	no	no
Auxin-responsive protein IAA10	no	yes	no	yes	yes	no
Probable L-ascorbate peroxidase 4, peroxisomal {ECO:0000305}	no	yes	no	no	no	no
Zeta-carotene desaturase, chloroplastic/chromoplastic	no	yes	no	yes	yes	no
Serine/threonine-protein kinase TOR {ECO:0000305}	no	yes	no	yes	yes	yes
WAT1-related protein At4g01440	no	yes	no	yes	no	no
Probable serine/threonine-protein kinase PBL3 {ECO:0000305}	no	yes	yes	yes	no	yes
Photosystem I chlorophyll a/b-binding protein 5, chloroplastic {ECO:0000303|PubMed:10366881}	no	yes	no	no	yes	yes
Probable E3 ubiquitin-protein ligase ARI7	no	yes	yes	yes	no	no
DEAD-box ATP-dependent RNA helicase 7	no	yes	yes	yes	yes	no
NAD-dependent malic enzyme 59 kDa isoform, mitochondrial	no	yes	yes	yes	no	yes
Ubiquitin carboxyl-terminal hydrolase 5	no	yes	no	yes	no	yes
Serine/threonine-protein phosphatase 6 regulatory subunit 3	no	yes	no	no	yes	no
ATP-citrate synthase alpha chain protein 2	no	yes	yes	yes	no	no
Cadmium/zinc-transporting ATPase HMA2 {ECO:0000305}	no	yes	no	no	yes	no
Peroxiredoxin-2C	no	yes	yes	yes	yes	no
Transcription factor bHLH68	no	yes	yes	yes	no	yes
G-type lectin S-receptor-like serine/threonine-protein kinase B120	no	yes	yes	yes	no	no
Protein AE7-like 1 {ECO:0000303|PubMed:23104832}	no	yes	yes	no	no	no
Splicing factor U2af large subunit B	no	yes	yes	yes	no	no
Casein kinase 1-like protein HD16 {ECO:0000305}	no	yes	no	yes	no	no
E3 ubiquitin-protein ligase XB3	no	yes	yes	yes	yes	yes
Metal tolerance protein 7	no	yes	yes	yes	no	no
Nascent polypeptide-associated complex subunit beta {ECO:0000255|RuleBase:RU361272}	no	yes	yes	yes	yes	no
Eukaryotic translation initiation factor isoform 4G-2	no	yes	no	yes	no	no
UDP-glycosyltransferase 76F1	no	yes	no	no	no	no
F-box protein At2g27310	no	yes	no	no	no	no
Two-component response regulator ORR24 {ECO:0000305}	no	yes	yes	yes	yes	no
60S acidic ribosomal protein P2A	no	yes	no	no	no	no
Disease resistance protein RPM1	no	yes	no	yes	no	no
40S ribosomal protein S19	no	yes	yes	no	no	no
Cytosolic Fe-S cluster assembly factor NBP35 {ECO:0000255|HAMAP-Rule:MF_03038}	no	yes	yes	yes	no	yes
Testis-expressed protein 2	no	yes	no	no	no	no
Protein ESKIMO 1	no	yes	no	yes	no	no
Nitrate regulatory gene2 protein {ECO:0000303|PubMed:26744214}	no	yes	yes	no	yes	no
Importin subunit beta-1 {ECO:0000305}	no	yes	no	no	no	no
DExH-box ATP-dependent RNA helicase DExH3 {ECO:0000305}	no	yes	yes	yes	no	yes
Putative aconitate hydratase, cytoplasmic	no	yes	no	no	yes	no
Fumarate hydratase 2 {ECO:0000303|PubMed:20202172}	no	yes	yes	no	no	no
Diphosphomevalonate decarboxylase 2 {ECO:0000305}	no	yes	yes	yes	no	no
Probable folate-biopterin transporter 9, chloroplastic	no	yes	no	yes	yes	no
Probable potassium transporter 14	no	yes	no	yes	no	yes
Dol-P-Man:Man(7)GlcNAc(2)-PP-Dol alpha-1,6-mannosyltransferase	no	yes	no	yes	yes	no
Dihydrolipoyllysine-residue acetyltransferase component 2 of pyruvate dehydrogenase complex, mitochondrial	no	yes	yes	yes	no	no
Increased DNA methylation 1 {ECO:0000303|PubMed:22700931}	no	yes	yes	yes	no	no

Anthocyanidin reductase ((2S)-flavan-3-ol-forming) {ECO:0000303|PubMed:16806954}	no	yes	no	no	no	no
Rac-like GTP-binding protein 6	no	yes	yes	yes	yes	yes
Heavy metal-associated isoprenylated plant protein 16 {ECO:0000303|PubMed:21072340, ECO:0000303|PubMed:23368984}	no	yes	no	no	no	no
NADP-dependent D-sorbitol-6-phosphate dehydrogenase	no	yes	yes	yes	yes	no
Signal recognition particle receptor subunit beta	no	yes	yes	yes	yes	no
60S ribosomal protein L10a {ECO:0000303|PubMed:21205822}	no	yes	no	no	no	no
60S ribosomal protein L10a-1	no	yes	no	yes	yes	no
60S ribosomal protein L13-2	no	yes	yes	no	no	no
Nudix hydrolase 18, mitochondrial	no	yes	no	yes	no	no
Nucleoside diphosphate kinase IV, chloroplastic/mitochondrial	no	yes	no	no	yes	no
Golgi to ER traffic protein 4 homolog	no	yes	no	no	yes	no
Receptor-interacting serine/threonine-protein kinase 4	no	yes	yes	no	yes	yes
Nicotianamine aminotransferase 1 {ECO:0000303|PubMed:15781441}	no	yes	yes	yes	yes	yes
Exocyst complex component EXO70B1 {ECO:0000303|PubMed:23944713}	no	yes	yes	no	no	no
Exportin-7	no	yes	no	no	yes	no
MA3 DOMAIN-CONTAINING TRANSLATION REGULATORY FACTOR 1 {ECO:0000303|PubMed:29084871}	no	yes	yes	no	yes	no
Putative FBD-associated F-box protein At1g55030	no	yes	yes	no	no	no
Pre-mRNA-processing protein 40C {ECO:0000303|PubMed:19467629}	no	yes	no	no	no	no
Endoribonuclease Dicer homolog 3a	no	yes	no	no	no	no
Eukaryotic translation initiation factor 4G	no	yes	yes	no	yes	no
Cytochrome c oxidase subunit 5b-2, mitochondrial	no	yes	yes	no	no	no
Flap endonuclease 1-B {ECO:0000255|HAMAP-Rule:MF_03140}	no	yes	no	no	no	no
UDP-galactose transporter 1 {ECO:0000303|PubMed:15456736}	no	yes	yes	no	yes	no
tRNA(His) guanylyltransferase 2	no	yes	yes	no	no	no
Boron transporter 4	no	yes	yes	yes	yes	no
Protein MARD1 {ECO:0000303|PubMed:15159630}	no	yes	yes	yes	yes	yes
Putative D-cysteine desulfhydrase 1, mitochondrial	no	yes	no	no	yes	no
GDP-mannose 4,6 dehydratase 2	no	yes	yes	no	no	no
Type I inositol polyphosphate 5-phosphatase 1 {ECO:0000303|PubMed:11402208}	no	yes	yes	yes	yes	no
Elongator complex protein 2	no	yes	yes	yes	yes	yes
mRNA cap guanine-N7 methyltransferase 1	no	yes	yes	yes	no	no
2-carboxy-D-arabinitol-1-phosphatase {ECO:0000305}	no	yes	no	no	yes	no
AP2-like ethylene-responsive transcription factor At1g16060	no	yes	no	yes	no	yes
Serine/threonine-protein kinase SRPK {ECO:0000303|PubMed:19657567}	no	yes	yes	yes	yes	yes
Phospholipase A1-II 5	no	yes	yes	no	no	no
Mechanosensitive ion channel protein 1, mitochondrial	no	yes	yes	no	no	no
Tubby-like F-box protein 14	no	yes	no	no	no	yes
Exonuclease 1	no	yes	yes	no	no	no
Putative ubiquitin-conjugating enzyme E2 38	no	yes	yes	yes	yes	yes
Calcium-dependent protein kinase 3 {ECO:0000303|PubMed:12068094}	no	yes	yes	yes	no	no
Potassium transporter 23	no	yes	yes	no	yes	no
Phosphoenolpyruvate carboxykinase (ATP) 2	no	yes	yes	yes	yes	no
RNA polymerase II C-terminal domain phosphatase-like 3	no	yes	yes	no	yes	no
Mediator of RNA polymerase II transcription subunit 13	no	yes	no	no	no	no
GDP-L-galactose phosphorylase 2	no	yes	yes	no	yes	no
SNF1-related protein kinase regulatory subunit gamma-1-like	no	yes	yes	yes	yes	yes
Transcription factor UNE12	no	yes	no	yes	yes	no
E3 ubiquitin-protein ligase MIEL1 {ECO:0000303|PubMed:23403577}	no	yes	no	no	yes	no

Protein decapping 5	no	yes	yes	yes	yes	no
Replication protein A 70 kDa DNA-binding subunit B	no	yes	yes	yes	yes	no
ARM REPEAT PROTEIN INTERACTING WITH ABF2	no	yes	no	no	no	no
RNA polymerase II C-terminal domain phosphatase-like 2	no	yes	yes	yes	no	no
Pectin acetylesterase 5 {ECO:0000303|PubMed:25115560}	no	yes	yes	yes	no	no
Cinnamoyl-CoA reductase 1	no	yes	no	no	no	no
Pentatricopeptide repeat-containing protein At1g11290, chloroplastic {ECO:0000305}	no	yes	no	no	no	no
F-box protein PP2-B10	no	yes	yes	yes	yes	no
Putative BPI/LBP family protein At1g04970	no	yes	yes	yes	yes	yes
Zinc finger protein CONSTANS-LIKE 4	no	yes	yes	yes	no	no
Tryptophan aminotransferase-related protein 2 {ECO:0000303|PubMed:22582989}	no	yes	yes	yes	yes	no
Homeobox protein BEL1 homolog	no	yes	yes	yes	yes	yes
PRA1 family protein A3	no	yes	no	no	yes	no
Glycine-rich RNA-binding protein blt801 {ECO:0000250|UniProtKB:Q03250, ECO:0000303|PubMed:8639753}	no	yes	yes	no	no	no
Zinc finger protein BRUTUS {ECO:0000303|PubMed:20675571}	no	yes	yes	yes	no	no
Phosphoribosylaminoimidazole carboxylase, chloroplastic	no	yes	no	yes	no	no
Glutamate synthase 1 [NADH], chloroplastic	no	yes	yes	no	yes	no
Zinc finger protein CONSTANS-LIKE 10	no	yes	no	no	no	no
Protein BUNDLE SHEATH DEFECTIVE 2, chloroplastic {ECO:0000303|PubMed:10330470}	no	yes	no	no	no	no
Polypyrimidine tract-binding protein homolog 1	no	yes	no	no	no	no
Outer envelope pore protein 16-2, chloroplastic	no	yes	no	no	no	no
Gibberellin 2-beta-dioxygenase 1	no	yes	no	no	no	no
Protein REDUCED WALL ACETYLATION 2 {ECO:0000303|PubMed:21212300, ECO:0000303|PubMed:21673009}	no	yes	no	no	no	no
Tubby-like F-box protein 5	no	yes	no	no	no	no
Retrovirus-related Pol polyprotein from transposon RE1	no	yes	no	no	no	no
Mitogen-activated protein kinase kinase kinase 3 {ECO:0000303|PubMed:27679653}	no	no	yes	yes	no	no
Probable E3 ubiquitin-protein ligase LUL2	no	no	yes	no	no	no
Caffeoylshikimate esterase	no	no	yes	no	no	no
30S ribosomal protein S6 alpha, chloroplastic	no	no	yes	yes	no	no
Cytochrome P450 99A2	no	no	yes	yes	yes	no
Probable GTP diphosphokinase CRSH1, chloroplastic	no	no	yes	no	yes	no
La-related protein 6B	no	no	yes	yes	no	no
Proline iminopeptidase	no	no	yes	no	no	no
Probable LRR receptor-like serine/threonine-protein kinase MEE39	no	no	yes	yes	no	no
Transcription factor TFIIIB component B" {ECO:0000250|UniProtKB:P46678}	no	no	yes	no	no	no
DNA-directed RNA polymerases II, IV and V subunit 9A	no	no	yes	no	yes	no
TNF receptor-associated factor homolog 1a {ECO:0000305}	no	no	yes	no	no	no
Prosaposin	no	no	yes	yes	yes	no
Mitogen-activated protein kinase kinase 2	no	no	yes	yes	no	no
Mediator of RNA polymerase II transcription subunit 21	no	no	yes	yes	no	no
Heat shock 70 kDa protein 14	no	no	yes	yes	yes	yes
ABC transporter C family member 2	no	no	yes	yes	yes	no
NAC domain-containing protein 21/22	no	no	yes	no	no	no
Receptor-like protein kinase HSL1	no	no	yes	no	no	no
Phosphatidate phosphatase PAH2	no	no	yes	yes	yes	no
Putative cyclin-dependent kinase F-2	no	no	yes	no	yes	no

Protein PGR {ECO:0000303|Ref.6}	no	no	yes	no	no	no
Ubiquitin-activating enzyme E1 3	no	no	yes	no	no	no
Uncharacterized protein At2g34460, chloroplastic	no	no	yes	no	no	yes
Probable serine/threonine-protein kinase SIS8 {ECO:0000305}	no	no	yes	yes	no	yes
Alpha-mannosidase {ECO:0000303|PubMed:4973951}	no	no	yes	no	no	no
E3 ubiquitin-protein ligase RNF144B	no	no	yes	no	no	no
Pentatricopeptide repeat-containing protein At5g55840	no	no	yes	yes	no	no
Nucleobase-ascorbate transporter 7	no	no	yes	no	no	no
Putative phospholipid-transporting ATPase 9 {ECO:0000303|PubMed:11402198}	no	no	yes	no	no	no
Adenine/guanine permease AZG1	no	no	yes	no	no	no
Choline/ethanolaminephosphotransferase 1	no	no	yes	no	yes	yes
Alpha-aminoadipic semialdehyde synthase	no	no	yes	no	no	no
Protein STICHEL-like 2	no	no	yes	yes	no	no
Pleckstrin homology domain-containing protein 1	no	no	yes	yes	no	no
WD repeat-containing protein 13	no	no	yes	no	no	no
Protein MATERNALLY EXPRESSED GENE 5	no	no	yes	no	no	no
Sugar transporter ERD6-like 5	no	no	yes	no	no	no
E3 ubiquitin-protein ligase UPL2	no	no	yes	yes	yes	no
Endonuclease MutS2 {ECO:0000255|HAMAP-Rule:MF_00092}	no	no	yes	no	no	no
Long chain acyl-CoA synthetase 9, chloroplastic	no	no	yes	yes	yes	yes
Probable metal-nicotianamine transporter YSL13	no	no	yes	yes	no	yes
DNA topoisomerase 1 alpha {ECO:0000305}	no	no	yes	no	no	no
Protein PHLOEM PROTEIN 2-LIKE A1	no	no	yes	yes	yes	no
Transcription factor MTB1 {ECO:0000303|PubMed:30610166}	no	no	yes	no	no	no
Sedoheptulose-1,7-bisphosphatase, chloroplastic	no	no	yes	no	yes	no
Branched-chain-amino-acid aminotransferase 5, chloroplastic	no	no	yes	no	no	no
GPN-loop GTPase 3 {ECO:0000250|UniProtKB:Q06543}	no	no	yes	no	no	yes
Probable serine protease EDA2	no	no	yes	yes	no	no
Cystathionine beta-lyase, chloroplastic	no	no	yes	no	yes	yes
Protein CASP	no	no	yes	no	no	no
DExH-box ATP-dependent RNA helicase DExH11 {ECO:0000305}	no	no	yes	no	yes	yes
Calreticulin-3	no	no	yes	no	no	no
E3 ubiquitin-protein ligase HOS1	no	no	yes	no	yes	no
Pentatricopeptide repeat-containing protein At5g13770, chloroplastic	no	no	yes	no	yes	no
Equilibrative nucleotide transporter 1	no	no	yes	yes	no	no
TOM1-like protein 8 {ECO:0000305}	no	no	yes	no	no	no
Probable E3 ubiquitin-protein ligase ZFP1 {ECO:0000305}	no	no	yes	no	no	no
Protein MID1-COMPLEMENTING ACTIVITY 1	no	no	yes	no	yes	no
Asparagine synthetase domain-containing protein 1	no	no	yes	yes	yes	yes
Uncharacterized WD repeat-containing protein C2A9.03	no	no	yes	no	yes	no
Protein THYLAKOID RHODANESE-LIKE, chloroplastic {ECO:0000303|PubMed:26941088}	no	no	yes	no	yes	yes
Probable serine/threonine-protein kinase PBL23 {ECO:0000305}	no	no	yes	no	no	no
Protein FAR1-RELATED SEQUENCE 5	no	no	yes	yes	yes	yes
Metal tolerance protein C4	no	no	yes	yes	no	no
Transposon TX1 uncharacterized 149 kDa protein	no	no	yes	yes	no	no
Protein DEHYDRATION-INDUCED 19 homolog 5	no	no	yes	no	no	no
Ubiquinone biosynthesis O-methyltransferase, mitochondrial {ECO:0000255|HAMAP-Rule:MF_03190}	no	no	yes	no	no	no
Reticulon-like protein B23	no	no	yes	no	no	no
Protein TRANSPARENT TESTA 9 {ECO:0000303|PubMed:8528278}	no	no	yes	no	yes	yes

Bax inhibitor 1	no	no	yes	no	no	no
Very-long-chain aldehyde decarbonylase GL1-4 {ECO:0000305}	no	no	yes	yes	no	no
Pentatricopeptide repeat-containing protein At5g21222	no	no	yes	no	yes	no
KH domain-containing protein SPIN1	no	no	yes	no	no	no
Trafficking protein particle complex subunit 5	no	no	yes	no	no	yes
Nicastrin	no	no	yes	no	no	no
BTB/POZ domain and ankyrin repeat-containing protein NPR2 {ECO:0000305}	no	no	yes	no	yes	yes
Transcription factor bHLH140	no	no	yes	no	no	no
Auxin response factor 21	no	no	yes	yes	yes	no
Polygalacturonase	no	no	yes	no	no	no
Scarecrow-like protein 9	no	no	yes	no	no	no
Glycine-rich RNA-binding protein 2, mitochondrial	no	no	yes	no	no	yes
ARF guanine-nucleotide exchange factor GNOM	no	no	yes	no	no	no
Internal alternative NAD(P)H-ubiquinone oxidoreductase A1, mitochondrial	no	no	yes	no	no	no
Ubiquitin-conjugating enzyme E2 28	no	no	yes	yes	yes	no
Putative UPF0481 protein At3g02645	no	no	yes	yes	no	yes
DEAD-box ATP-dependent RNA helicase 52A	no	no	yes	no	no	no
Single myb histone 6	no	no	yes	yes	no	no
Brassinosteroid-responsive RING protein 1 {ECO:0000303|PubMed:12012249}	no	no	yes	no	no	no
CRC domain-containing protein TSO1	no	no	yes	no	no	no
Protein IN2-1 homolog B	no	no	yes	no	yes	no
GABA transporter 1	no	no	yes	no	no	no
Syntaxin-121	no	no	yes	no	no	no
3-isopropylmalate dehydrogenase 2, chloroplastic {ECO:0000303|PubMed:15849421}	no	no	yes	yes	no	no
Wall-associated receptor kinase 3	no	no	yes	no	no	no
30S ribosomal protein S15 {ECO:0000255|HAMAP-Rule:MF_01343}	no	no	yes	no	no	no
Mitotic spindle checkpoint protein BUBR1	no	no	yes	no	no	no
SH3 domain-containing protein 2 {ECO:0000312|EMBL:AAL32439.1}	no	no	yes	no	yes	no
Pentatricopeptide repeat-containing protein At3g48810	no	no	yes	yes	yes	no
Polypyrimidine tract-binding protein homolog 3	no	no	yes	no	yes	no
Calmodulin binding protein PICBP {ECO:0000305}	no	no	yes	no	no	no
AT-hook motif nuclear-localized protein 11 {ECO:0000312|EMBL:FAA00282.1}	no	no	yes	no	yes	yes
Tubby-like F-box protein 7	no	no	yes	no	no	no
General transcription factor IIH subunit 2 {ECO:0000305}	no	no	yes	yes	yes	no
Chaperone protein dnaJ 49	no	no	yes	no	no	no
Aspartic proteinase oryzasin-1	no	no	yes	no	no	no
Dynamin-like protein ARC5	no	no	yes	no	no	no
Flowering time control protein FY	no	no	yes	no	no	no
Zinc finger BED domain-containing protein RICESLEEPER 2	no	no	yes	no	yes	no
UDP-rhamnose/UDP-galactose transporter 5 {ECO:0000312|EMBL:AKA88218.1}	no	no	yes	yes	yes	no
Anoctamin-like protein Os01g0706700	no	no	yes	yes	no	no
Probable pinoresinol-lariciresinol reductase 3	no	no	yes	no	no	yes
Protein trichome birefringence-like 18	no	no	yes	no	no	no
Allene oxide synthase 2	no	no	yes	no	no	no
AP-2 complex subunit alpha-1	no	no	yes	no	no	no
Histone-lysine N-methyltransferase ATX4	no	no	yes	no	no	yes
Xanthine dehydrogenase	no	no	yes	yes	yes	no
Solanesyl-diphosphate synthase 2, chloroplastic {ECO:0000303|PubMed:20421194}	no	no	yes	yes	yes	no
Zinc finger CCCH domain-containing protein 31	no	no	yes	no	no	yes

Alpha-mannosidase I MNS5	no	no	yes	no	yes	yes
Patatin-like protein 2	no	no	yes	no	yes	no
Protein SEH1 {ECO:0000303|PubMed:21189294}	no	no	yes	no	no	no
Hydroxyproline O-galactosyltransferase GALT4 {ECO:0000303|PubMed:26690932}	no	no	yes	no	yes	yes
Gluconokinase {ECO:0000303|PubMed:12447540}	no	no	yes	yes	no	yes
Transcription termination factor MTEF1, chloroplastic {ECO:0000305}	no	no	yes	yes	no	no
CRS2-associated factor 1, chloroplastic	no	no	yes	no	no	no
Two-component response regulator-like PRR1	no	no	yes	no	no	no
LOB domain-containing protein 12	no	no	yes	no	no	no
Glycosyltransferase BC10 {ECO:0000305}	no	no	yes	no	no	no
WPP domain-associated protein	no	no	yes	yes	no	yes
Zinc finger protein ZOP1 {ECO:0000305}	no	no	yes	no	no	no
Glycine dehydrogenase (decarboxylating), mitochondrial	no	no	yes	yes	yes	yes
Amino acid transporter AVT1I {ECO:0000305}	no	no	yes	no	no	no
Peroxidase 17	no	no	yes	no	no	no
Protein NRT1/ PTR FAMILY 8.5	no	no	yes	no	yes	yes
Regulation of nuclear pre-mRNA domain-containing protein 1A	no	no	yes	no	no	no
5'-3' exonuclease	no	no	yes	no	no	no
Cullin-1 {ECO:0000305}	no	no	yes	no	no	no
ATPase family AAA domain-containing protein 3-B	no	no	yes	no	no	no
Pyrophosphate–fructose 6-phosphate 1-phosphotransferase subunit alpha {ECO:0000255|HAMAP-Rule:MF_03185}	no	no	yes	no	no	no
Serine/threonine-protein kinase STY13 {ECO:0000305}	no	no	yes	no	no	no
Protein ALWAYS EARLY 2	no	no	yes	no	no	no
Leucine-rich repeat receptor-like tyrosine-protein kinase PXC3 {ECO:0000305}	no	no	no	yes	yes	yes
1-aminocyclopropane-1-carboxylate oxidase	no	no	no	yes	no	yes
Protein farnesyltransferase subunit beta	no	no	no	yes	no	no
Cysteine synthase, chloroplastic/chromoplastic	no	no	no	yes	no	no
Transcription factor MYB124 {ECO:0000303|PubMed:11597504}	no	no	no	yes	no	no
Nuclear pore complex protein NUP133 {ECO:0000303|PubMed:21189294}	no	no	no	yes	no	no
Serine/threonine-protein kinase KIPK1 {ECO:0000305}	no	no	no	yes	no	no
NDR1/HIN1-like protein 1 {ECO:0000303|Ref.1}	no	no	no	yes	no	no
Zinc finger CCCH domain-containing protein 54	no	no	no	yes	no	no
Scarecrow-like protein 6	no	no	no	yes	yes	yes
Zinc finger CCCH domain-containing protein 36	no	no	no	yes	no	no
ATP-dependent DNA helicase MER3 homolog {ECO:0000305}	no	no	no	yes	no	no
Autophagy-related protein 101 {ECO:0000303|PubMed:24563201}	no	no	no	yes	yes	no
Protein REVEILLE 1	no	no	no	yes	yes	no
Mitochondrial import inner membrane translocase subunit TIM14-2	no	no	no	yes	no	no
Peroxisomal acyl-coenzyme A oxidase 1	no	no	no	yes	yes	yes
Nuclear pore complex protein NUP98A {ECO:0000303|PubMed:21189294}	no	no	no	yes	no	no
Protein real-time	no	no	no	yes	no	no
Extra-large guanine nucleotide-binding protein 1	no	no	no	yes	yes	no
Werner Syndrome-like exonuclease	no	no	no	yes	no	no
Activating signal cointegrator 1 {ECO:0000303|PubMed:12077347}	no	no	no	yes	no	no
LRR receptor-like serine/threonine-protein kinase SIK1 {ECO:0000305}	no	no	no	yes	no	no
BEL1-like homeodomain protein 4	no	no	no	yes	yes	no
Pyruvate kinase isozyme A, chloroplastic	no	no	no	yes	no	no
Shaggy-related protein kinase theta	no	no	no	yes	no	no
Putative E3 ubiquitin-protein ligase UBR7	no	no	no	yes	no	no
Wall-associated receptor kinase-like 10	no	no	no	yes	no	no

Probable protein phosphatase 2C 10	no	no	no	yes	no	no
Nuclear poly(A) polymerase 4 {ECO:0000303|PubMed:18479511}	no	no	no	yes	no	no
Casein kinase 1-like protein 2 {ECO:0000305}	no	no	no	yes	no	no
Mediator of RNA polymerase II transcription subunit 25 {ECO:0000305}	no	no	no	yes	no	no
Auxin response factor 23	no	no	no	yes	no	no
NAD(P)H-quinone oxidoreductase subunit N, chloroplastic {ECO:0000305}	no	no	no	yes	no	no
Protein ENHANCED DISEASE RESISTANCE 4 {ECO:0000303|PubMed:25747881}	no	no	no	yes	no	yes
Pyruvate, phosphate dikinase 1, chloroplastic {ECO:0000303|PubMed:1668653}	no	no	no	yes	no	no
Magnesium-chelatase subunit ChlH, chloroplastic	no	no	no	yes	no	no
Polyprotein of EF-Ts, chloroplastic {ECO:0000303|PubMed:15548736}	no	no	no	yes	no	no
E3 ubiquitin-protein ligase SIRP1 {ECO:0000305}	no	no	no	yes	no	no
Serine/threonine-protein phosphatase BSL1 homolog	no	no	no	yes	no	no
L-threonate dehydrogenase {ECO:0000303|PubMed:27402745}	no	no	no	yes	no	no
Protein root UVB sensitive 6 {ECO:0000303|PubMed:19515790}	no	no	no	yes	no	no
Kinesin-like protein KIN-14M {ECO:0000305}	no	no	no	yes	no	no
Probable leucine-rich repeat receptor-like protein kinase At5g49770	no	no	no	yes	no	no
Tyrosine-sulfated glycopeptide receptor 1 {ECO:0000305}	no	no	no	yes	no	yes
Inositol-tetrakisphosphate 1-kinase 3 {ECO:0000305}	no	no	no	yes	no	no
Transcription factor TGAL6 {ECO:0000305}	no	no	no	yes	no	no
F-box protein At1g30200	no	no	no	yes	yes	no
Protein NRT1/ PTR FAMILY 4.6	no	no	no	yes	yes	no
2-alkenal reductase (NADP(+)-dependent)	no	no	no	yes	no	no
O-fucosyltransferase 15 {ECO:0000305}	no	no	no	yes	no	no
Myb-related protein Zm38	no	no	no	yes	no	no
Serine/threonine-protein kinase EDR1	no	no	no	yes	no	no
Probable folate-biopterin transporter 7	no	no	no	yes	yes	yes
Isopentenyl phosphate kinase {ECO:0000303|PubMed:24327557}	no	no	no	yes	no	no
Serine hydroxymethyltransferase 4	no	no	no	yes	no	no
Glutathione S-transferase T3	no	no	no	yes	no	no
Cysteine-rich receptor-like protein kinase 11	no	no	no	yes	no	no
Spatacsin	no	no	no	yes	no	no
Benzoate–CoA ligase, peroxisomal	no	no	no	yes	no	yes
Replication protein A 70 kDa DNA-binding subunit	no	no	no	yes	yes	no
Phosphatidylinositol 4-kinase alpha 1	no	no	no	yes	no	no
NAC domain-containing protein 2 {ECO:0000303|PubMed:15029955}	no	no	no	yes	no	no
NRR repressor homolog 1 {ECO:0000305}	no	no	no	yes	no	no
3-ketoacyl-CoA synthase 1 {ECO:0000303|PubMed:10074711}	no	no	no	yes	no	no
F-box protein SKIP8	no	no	no	yes	no	no
Ubiquitin carboxyl-terminal hydrolase 24	no	no	no	yes	no	no
Polyubiquitin 11	no	no	no	yes	no	no
Ran-binding protein M homolog {ECO:0000305}	no	no	no	yes	yes	no
TVP38/TMEM64 family membrane protein slr0305	no	no	no	yes	no	yes
Protein EMSY-LIKE 3 {ECO:0000303|PubMed:21830950}	no	no	no	yes	no	yes
NADH dehydrogenase [ubiquinone] iron-sulfur protein 1, mitochondrial	no	no	no	yes	no	no
F-box protein FBW2	no	no	no	yes	no	no
Protein-L-isoaspartate O-methyltransferase	no	no	no	yes	no	no
Receptor-like cytoplasmic kinase 185 {ECO:0000303|PubMed:19825577}	no	no	no	yes	yes	no
Probable LRR receptor-like serine/threonine-protein kinase At4g31250	no	no	no	yes	no	no
DNA repair protein UVH3	no	no	no	yes	no	no
Organellar oligopeptidase A, chloroplastic/mitochondrial	no	no	no	yes	no	no

Putative alpha-L-fucosidase 1	no	no	no	yes	no	no
Light-regulated protein, chloroplastic	no	no	no	yes	no	no
Cyclin-L1-1	no	no	no	yes	no	no
ATP-dependent DNA helicase At3g02060, chloroplastic	no	no	no	yes	no	no
Pentatricopeptide repeat-containing protein At5g52850, chloroplastic	no	no	no	yes	no	no
Golgi SNAP receptor complex member 1-2	no	no	no	yes	no	no
Transcription factor GTE4	no	no	no	yes	yes	no
Probable mitochondrial saccharopine dehydrogenase-like oxidoreductase At5g39410	no	no	no	yes	no	no
ABC transporter G family member 11	no	no	no	yes	no	no
50S ribosomal protein L10, chloroplastic	no	no	no	yes	no	yes
Protein transport protein sec23-1	no	no	no	yes	yes	no
ATP-dependent DNA helicase 2 subunit KU80	no	no	no	yes	no	no
Probable histone H2A variant 3	no	no	no	yes	no	yes
Squamosa promoter-binding-like protein 6	no	no	no	yes	no	no
ABC transporter G family member 53 {ECO:0000305}	no	no	no	yes	no	no
NADH–cytochrome b5 reductase 1	no	no	no	yes	yes	no
Serine/threonine-protein kinase SAPK1	no	no	no	yes	no	no
Polyadenylate-binding protein RBP47	no	no	no	yes	no	no
La-related protein 1B	no	no	no	yes	no	yes
F-box protein At1g70590	no	no	no	yes	no	no
Zinc finger CCCH domain-containing protein 32	no	no	no	yes	yes	no
Auxin efflux carrier component 1a {ECO:0000305}	no	no	no	yes	no	yes
Disease resistance protein PIK6-NP {ECO:0000305}	no	no	no	yes	yes	yes
Protein EDS1L {ECO:0000305}	no	no	no	yes	no	no
Transcription factor bHLH63	no	no	no	yes	no	no
Probable ubiquitin-conjugating enzyme E2 33	no	no	no	yes	no	no
Replication protein A 70 kDa DNA-binding subunit C	no	no	no	yes	no	no
Type IV inositol polyphosphate 5-phosphatase 3 {ECO:0000305}	no	no	no	yes	yes	no
Receptor-like serine/threonine-protein kinase NCRK	no	no	no	yes	no	no
Myosin-binding protein 2 {ECO:0000303|PubMed:23995081}	no	no	no	yes	yes	yes
Monothiol glutaredoxin-S7, chloroplastic	no	no	no	yes	no	no
F-box protein At4g18380	no	no	no	yes	yes	yes
Pentatricopeptide repeat-containing protein At4g14850	no	no	no	yes	yes	no
Endoplasmic reticulum oxidoreductin-1	no	no	no	yes	no	no
AFG1-like ATPase {ECO:0000312|MGI:MGI:2148801}	no	no	no	yes	yes	no
SUMO-conjugating enzyme SCE1	no	no	no	yes	no	no
Acyl-CoA-binding domain-containing protein 4	no	no	no	yes	no	no
Serine/threonine-protein kinase GRIK2	no	no	no	yes	no	no
Serine/threonine-protein kinase ATM {ECO:0000303|PubMed:10734187}	no	no	no	yes	yes	no
Methionine adenosyltransferase 2 subunit beta	no	no	no	yes	no	no
Vacuolar protein sorting-associated protein 32 homolog 1	no	no	no	yes	yes	no
Dynein light chain, cytoplasmic	no	no	no	yes	no	no
Probable cytokinin riboside 5'-monophosphate phosphoribohydrolase LOGL9	no	no	no	yes	no	no
Non-symbiotic hemoglobin 1	no	no	no	yes	no	no
Heat shock 70 kDa protein 17	no	no	no	no	yes	no
Two-component response regulator ORR6 {ECO:0000305}	no	no	no	no	yes	yes
Receptor-like cytosolic serine/threonine-protein kinase RBK1	no	no	no	no	yes	no
Myosin-15	no	no	no	no	yes	no
CASP-like protein 4U1	no	no	no	no	yes	no
4-hydroxyphenylacetaldehyde oxime monooxygenase	no	no	no	no	yes	no

Syntaxin-52	no	no	no	no	yes	yes
Splicing factor 3B subunit 4	no	no	no	no	yes	no
Ninja-family protein 6	no	no	no	no	yes	no
Two-component response regulator ORR9 {ECO:0000305}	no	no	no	no	yes	no
Peroxidase 52	no	no	no	no	yes	no
Mitochondrial metalloendopeptidase OMA1	no	no	no	no	yes	no
Nucleoprotein TPR	no	no	no	no	yes	no
Cationic amino acid transporter 1	no	no	no	no	yes	no
60S ribosomal protein L17	no	no	no	no	yes	no
Monoacylglycerol lipase {ECO:0000303|PubMed:17784850}	no	no	no	no	yes	no
F-box/FBD/LRR-repeat protein At1g13570	no	no	no	no	yes	no
Probable E3 ubiquitin-protein ligase RHY1A {ECO:0000305}	no	no	no	no	yes	no
MAG2-interacting protein 2 {ECO:0000303|PubMed:24118572}	no	no	no	no	yes	no
Ribulose bisphosphate carboxylase large chain {ECO:0000255|HAMAP-Rule:MF_01338}	no	no	no	no	yes	yes
Receptor-like protein 44 {ECO:0000303|PubMed:18434605}	no	no	no	no	yes	yes
Bifunctional purine biosynthesis protein PurH {ECO:0000255|HAMAP-Rule:MF_00139}	no	no	no	no	yes	no
Tudor domain-containing protein 3	no	no	no	no	yes	no
Myb family transcription factor PHL5 {ECO:0000305}	no	no	no	no	yes	no
Acyl carrier protein 1, chloroplastic	no	no	no	no	yes	no
DNA topoisomerase 2	no	no	no	no	yes	no
Transcriptional repressor ILP1 {ECO:0000303|PubMed:17012601}	no	no	no	no	yes	no
Protein NETWORKED 4A {ECO:0000303|PubMed:22840520}	no	no	no	no	yes	no
Delta-1-pyrroline-5-carboxylate synthase 1 {ECO:0000305}	no	no	no	no	yes	yes
Aquaporin TIP4-2	no	no	no	no	yes	no
BTB/POZ domain-containing protein At2g30600	no	no	no	no	yes	no
Metal tolerance protein 5	no	no	no	no	yes	no
Homeobox-leucine zipper protein HOX19	no	no	no	no	yes	yes
Probable pre-mRNA-splicing factor ATP-dependent RNA helicase DEAH5 {ECO:0000305}	no	no	no	no	yes	no
Zinc finger CCCH domain-containing protein 49	no	no	no	no	yes	no
Nuclear transcription factor Y subunit C-6 {ECO:0000305}	no	no	no	no	yes	no
Protein HLB1 {ECO:0000303|PubMed:26941089}	no	no	no	no	yes	no
O-fucosyltransferase 31 {ECO:0000305}	no	no	no	no	yes	no
Peroxisomal membrane protein PEX14	no	no	no	no	yes	yes
Histidinol-phosphate aminotransferase, chloroplastic	no	no	no	no	yes	no
Conserved oligomeric Golgi complex subunit 1 {ECO:0000303|PubMed:27448097}	no	no	no	no	yes	no
PsbP domain-containing protein 3, chloroplastic	no	no	no	no	yes	no
Cysteine-rich receptor-like protein kinase 10 {ECO:0000305}	no	no	no	no	yes	yes
Glutamate receptor 3.4	no	no	no	no	yes	no
Protection of telomeres protein 1a {ECO:0000303|PubMed:17627276}	no	no	no	no	yes	no
Protein UPSTREAM OF FLC	no	no	no	no	yes	no
DnaJ homolog subfamily B member 5	no	no	no	no	yes	no
Putative methylesterase 12, chloroplastic	no	no	no	no	yes	no
Histidine protein methyltransferase 1 homolog	no	no	no	no	yes	no
Cryptochrome DASH, chloroplastic/mitochondrial	no	no	no	no	yes	yes
Zeamatin	no	no	no	no	yes	no
Nuclear pore complex protein NUP96 {ECO:0000303|PubMed:21189294}	no	no	no	no	yes	yes
Katanin p80 WD40 repeat-containing subunit B1 homolog {ECO:0000255|HAMAP-Rule:MF_03022}	no	no	no	no	yes	no

Transposon Tf2-9 polyprotein	no	no	no	no	yes	yes
DNA replication licensing factor MCM5	no	no	no	no	yes	no
Transcription factor TCP20	no	no	no	no	yes	no
Kinesin-like protein KIN-5A {ECO:0000305}	no	no	no	no	yes	no
Protein CHUP1, chloroplastic	no	no	no	no	yes	no
Zinc finger MYM-type protein 1	no	no	no	no	yes	no
Methylcrotonoyl-CoA carboxylase subunit alpha, mitochondrial	no	no	no	no	yes	no
Protein CHLORORESPIRATORY REDUCTION 7, chloroplastic {ECO:0000305}	no	no	no	no	yes	no
Polyadenylate-binding protein RBP45B	no	no	no	no	yes	no
ATPase family AAA domain-containing protein 1	no	no	no	no	yes	yes
Glucan endo-1,3-beta-glucosidase 13	no	no	no	no	yes	no
ABC transporter C family member 8	no	no	no	no	yes	yes
8-amino-7-oxononanoate synthase	no	no	no	no	yes	no
Bromodomain and PHD finger-containing protein 3	no	no	no	no	yes	no
Probable staphylococcal-like nuclease CAN1	no	no	no	no	yes	no
WAT1-related protein At5g07050	no	no	no	no	yes	no
Inactive poly [ADP-ribose] polymerase RCD1	no	no	no	no	yes	no
E3 ubiquitin-protein ligase Os04g0590900	no	no	no	no	yes	no
Type I inositol polyphosphate 5-phosphatase 10 {ECO:0000305}	no	no	no	no	yes	no
Transmembrane 9 superfamily member 9 {ECO:0000305}	no	no	no	no	yes	no
Chlorophyll a-b binding protein 1D	no	no	no	no	yes	no
E3 ubiquitin-protein ligase SPL11	no	no	no	no	yes	no
Protein CANDIDATE G-PROTEIN COUPLED RECEPTOR 7 {ECO:0000303|PubMed:18671868}	no	no	no	no	yes	no
Clp protease adapter protein ClpF, chloroplastic {ECO:0000303|PubMed:26419670}	no	no	no	no	yes	no
Acylamino-acid-releasing enzyme 1 {ECO:0000305}	no	no	no	no	yes	no
50S ribosomal protein L15, chloroplastic	no	no	no	no	yes	no
F-box protein SKIP22	no	no	no	no	yes	no
Respiratory burst oxidase homolog protein B	no	no	no	no	yes	no
SH3 domain-containing protein 3 {ECO:0000312|EMBL:AEE83989.1}	no	no	no	no	yes	no
ALA-interacting subunit 1	no	no	no	no	yes	no
Temperature-induced lipocalin-1 {ECO:0000303|PubMed:18671872}	no	no	no	no	yes	no
Omega-amidase, chloroplastic {ECO:0000305}	no	no	no	no	yes	no
Basic leucine zipper 24 {ECO:0000303|PubMed:11906833}	no	no	no	no	yes	yes
Aspartic proteinase CDR1	no	no	no	no	yes	no
ABC transporter C family member 14	no	no	no	no	yes	yes
Dual specificity protein kinase splA	no	no	no	no	yes	no
Zinc finger CCCH domain-containing protein 5	no	no	no	no	yes	yes
Putative NAD kinase 3	no	no	no	no	yes	no
Transcription factor bHLH77	no	no	no	no	yes	no
Lipoxygenase 2.3, chloroplastic	no	no	no	no	yes	no
Uncharacterized protein At5g50100, chloroplastic	no	no	no	no	yes	no
Probable homogentisate phytyltransferase 2, chloroplastic	no	no	no	no	yes	no
Kinesin-like protein KIN-10A {ECO:0000305}	no	no	no	no	yes	no
Pentatricopeptide repeat-containing protein At2g17033	no	no	no	no	yes	no
F-box protein At1g47056	no	no	no	no	yes	no
Putative dihydroflavonol 4-reductase	no	no	no	no	yes	no
Armadillo repeat-containing protein 8	no	no	no	no	yes	no
PHD finger protein EHD3 {ECO:0000305}	no	no	no	no	yes	no

70 kDa peptidyl-prolyl isomerase	no	no	no	no	yes	yes
CTP synthase	no	no	no	no	yes	no
Mitochondrial amidoxime reducing component 2	no	no	no	no	yes	no
Protein RIK	no	no	no	no	yes	no
Magnesium-protoporphyrin IX monomethyl ester [oxidative] cyclase, chloroplastic	no	no	no	no	yes	yes
Protein NETWORKED 3A {ECO:0000303|PubMed:22840520}	no	no	no	no	yes	yes
Signal recognition particle 54 kDa protein 3	no	no	no	no	yes	no
Probable protein S-acyltransferase 7	no	no	no	no	yes	no
U-box domain-containing protein 4	no	no	no	no	yes	no
Probable BOI-related E3 ubiquitin-protein ligase 3	no	no	no	no	yes	no
Uncharacterized membrane protein At3g27390	no	no	no	no	yes	no
Salicylic acid-binding protein 2	no	no	no	no	yes	no
Putative E3 ubiquitin-protein ligase LIN-2 {ECO:0000303|PubMed:19776163}	no	no	no	no	yes	no
Acyl-acyl carrier protein thioesterase ATL3, chloroplastic {ECO:0000305}	no	no	no	no	yes	no
Phosphoglycerate mutase-like protein 1 {ECO:0000305}	no	no	no	no	yes	no
E3 ubiquitin-protein ligase makorin	no	no	no	no	yes	no
Protein FLUORESCENT IN BLUE LIGHT, chloroplastic	no	no	no	no	yes	no
GDSL esterase/lipase At2g04570	no	no	no	no	yes	no
Gibberellin 2-beta-dioxygenase 3 {ECO:0000305}	no	no	no	no	yes	no
Aspartic proteinase nepenthesin-1	no	no	no	no	yes	yes
Testis-expressed protein 10	no	no	no	no	yes	no
Thaumatin-like protein	no	no	no	no	yes	no
Leucine-rich repeat receptor-like kinase protein THICK TASSEL DWARF1	no	no	no	no	yes	no
G patch domain-containing protein TGH homolog	no	no	no	no	yes	no
Protein GRIP	no	no	no	no	no	yes
Structural maintenance of chromosomes protein 5	no	no	no	no	no	yes
Persulfide dioxygenase ETHE1 homolog, mitochondrial	no	no	no	no	no	yes
Bifunctional lysine-specific demethylase and histidyl-hydroxylase NO66	no	no	no	no	no	yes
GDSL esterase/lipase At1g28570	no	no	no	no	no	yes
Villin-5 {ECO:0000303|PubMed:20807878}	no	no	no	no	no	yes
Calcium-dependent protein kinase 8 {ECO:0000305}	no	no	no	no	no	yes
Probable mediator of RNA polymerase II transcription subunit 26c	no	no	no	no	no	yes
PI-PLC X domain-containing protein At5g67130	no	no	no	no	no	yes
5-oxoprolinase	no	no	no	no	no	yes
Soluble inorganic pyrophosphatase	no	no	no	no	no	yes
Probable acyl-CoA dehydrogenase IBR3 {ECO:0000305}	no	no	no	no	no	yes
Sugar transporter ERD6-like 6	no	no	no	no	no	yes
1-acyl-sn-glycerol-3-phosphate acyltransferase	no	no	no	no	no	yes
Casein kinase 1-like protein 10 {ECO:0000305}	no	no	no	no	no	yes
Secretory carrier-associated membrane protein 4	no	no	no	no	no	yes
Endoglucanase 13	no	no	no	no	no	yes
ATP-dependent DNA helicase PIF1 {ECO:0000255|HAMAP-Rule:MF_03176}	no	no	no	no	no	yes
bZIP transcription factor RISBZ4 {ECO:0000305}	no	no	no	no	no	yes
Cytochrome P450 714C3	no	no	no	no	no	yes
Putative L-ascorbate peroxidase 6	no	no	no	no	no	yes
Protein MET1, chloroplastic {ECO:0000303|PubMed:25587003}	no	no	no	no	no	yes
Protein GRAVITROPIC IN THE LIGHT 1 {ECO:0000303|PubMed:16640600}	no	no	no	no	no	yes
Glutathione S-transferase	no	no	no	no	no	yes
Methionine aminopeptidase 1B, chloroplastic {ECO:0000255|HAMAP-Rule:MF_03174}	no	no	no	no	no	yes
5-pentadecatrienyl resorcinol O-methyltransferase	no	no	no	no	no	yes

Zinc finger CCCH domain-containing protein 53	no	no	no	no	no	yes
E3 ubiquitin-protein ligase UPL5	no	no	no	no	no	yes
Fasciclin-like arabinogalactan protein 1	no	no	no	no	no	yes
Protein DGS1, mitochondrial {ECO:0000305}	no	no	no	no	no	yes
E3 ubiquitin-protein ligase At1g12760	no	no	no	no	no	yes
Gamma-glutamylcyclotransferase 2-2 {ECO:0000305}	no	no	no	no	no	yes
Putative ABC transporter C family member 15	no	no	no	no	no	yes
Probable LRR receptor-like serine/threonine-protein kinase At1g56140	no	no	no	no	no	yes
Peroxidase 5 {ECO:0000250|UniProtKB:P22195}	no	no	no	no	no	yes
Protein ILITYHIA {ECO:0000312|EMBL:AEE34290.1}	no	no	no	no	no	yes
Homeobox-DDT domain protein RLT1 {ECO:0000305}	no	no	no	no	no	yes
Ribonuclease II, chloroplastic/mitochondrial	no	no	no	no	no	yes
Cell division control protein 48 homolog E	no	no	no	no	no	yes
Protein DETOXIFICATION 42 {ECO:0000303|PubMed:11739388}	no	no	no	no	no	yes
Retinoblastoma-related protein 1	no	no	no	no	no	yes
Calmodulin-5/6/7/8	no	no	no	no	no	yes
Molybdopterin synthase catalytic subunit {ECO:0000255|HAMAP-Rule:MF_03052}	no	no	no	no	no	yes
Transcription initiation factor IIE subunit alpha	no	no	no	no	no	yes
Probable RNA helicase SDE3	no	no	no	no	no	yes
Potassium channel AKT1	no	no	no	no	no	yes
Rhodanese-like domain-containing protein 6	no	no	no	no	no	yes
7-ethoxycoumarin O-deethylase	no	no	no	no	no	yes
Photosystem I subunit O	no	no	no	no	no	yes
Probable isoprenylcysteine alpha-carbonyl methylesterase ICMEL2	no	no	no	no	no	yes
Transcription factor bHLH35	no	no	no	no	no	yes
Transmembrane emp24 domain-containing protein p24delta5	no	no	no	no	no	yes
Protein NEN1 {ECO:0000303|PubMed:25081480}	no	no	no	no	no	yes
Pentatricopeptide repeat-containing protein At4g19220, mitochondrial	no	no	no	no	no	yes
Threonine dehydratase 1 biosynthetic, chloroplastic {ECO:0000305}	no	no	no	no	no	yes
Organic cation/carnitine transporter 7	no	no	no	no	no	yes
DNA-directed RNA polymerase II subunit RPB7	no	no	no	no	no	yes
Beta-glucosidase 22	no	no	no	no	no	yes
Double-stranded RNA-binding protein 2	no	no	no	no	no	yes
Sister chromatid cohesion protein PDS5 homolog A	no	no	no	no	no	yes
Putative disease resistance RPP13-like protein 1	no	no	no	no	no	yes

Table 5: Genes with allele-specific expression shared among accessions. This table contains the intersection of all the genotypes (All), the low-fiber genotypes (LF), LF and IN84-58 (LF_IN), LF and US85-1008 (LF_US), and LF with IN84-58 and US85-1008 (LF_IN_US).

Description	All	LF	LF_IN	LF_US	LF_IN_US
Diacylglycerol O-acyltransferase 3 {ECO:0000305}	yes	no	no	no	no
Histone deacetylase 14	yes	no	no	no	no
Probable disease resistance protein RPP1 {ECO:0000305}	yes	no	no	no	no
bZIP transcription factor TRAB1 {ECO:0000305}	yes	no	no	no	no
Stress-related protein	yes	no	no	no	no
Stomatal closure-related actin-binding protein 1 {ECO:0000303|PubMed:21719691}	yes	no	no	no	no

Transcription factor PHYTOCHROME INTERACTING FACTOR-LIKE 15 {ECO:0000303|PubMed:17485859}	yes	no	no	no	no
ADP-ribosylation factor GTPase-activating protein AGD12	yes	no	no	no	no
G-type lectin S-receptor-like serine/threonine-protein kinase At2g19130	yes	no	no	no	yes
3-ketoacyl-CoA synthase 4 {ECO:0000303|PubMed:18465198}	yes	no	no	no	no
Glucose-1-phosphate adenylyltransferase small subunit 2, chloroplastic/amyloplastic/cytosolic {ECO:0000305}	yes	no	no	no	no
Uncharacterized protein YKR070W	yes	no	no	no	no
Metallothionein-like protein 1B	yes	no	no	no	no
Protein NRT1/ PTR FAMILY 8.3	yes	no	no	no	no
Protein NRT1/ PTR FAMILY 5.10	yes	no	no	no	no
Auxin response factor 1	yes	no	no	no	no
Ferredoxin–nitrite reductase, chloroplastic	yes	no	no	no	no
Reactive Intermediate Deaminase A, chloroplastic {ECO:0000303|PubMed:25070638}	yes	no	no	no	no
BTB/POZ and MATH domain-containing protein 1	yes	no	no	no	no
RNA-binding protein 25	yes	no	no	no	no
Lipase	yes	no	no	no	no
Calmodulin-binding protein 60 C {ECO:0000303|PubMed:11782485}	yes	no	no	no	no
Non-lysosomal glucosylceramidase {ECO:0000305}	yes	no	no	no	no
Non-functional NADPH-dependent codeinone reductase 2	yes	no	no	no	no
Putative disease resistance protein RGA1	yes	no	no	no	no
Serine carboxypeptidase-like 50	yes	no	no	no	no
BTB/POZ domain-containing protein At1g30440	yes	no	no	no	no
Triacylglycerol lipase SDP1	yes	no	no	no	no
Transcription-associated protein 1	yes	no	no	no	no
GEM-like protein 1	yes	no	no	no	no
Phototropin-2	yes	no	no	no	no
Probable pterin-4-alpha-carbinolamine dehydratase, chloroplastic {ECO:0000305}	yes	no	no	no	no
Probable inactive ATP-dependent zinc metalloprotease FTSHI 3, chloroplastic	yes	no	no	no	no
NADP-dependent malic enzyme, chloroplastic {ECO:0000305}	yes	no	no	no	no
5'-adenylylsulfate reductase-like 4	yes	no	no	no	no
Protochlorophyllide-dependent translocon component 52, chloroplastic	yes	no	no	no	no
Probable GTP-binding protein OBGC1, chloroplastic	yes	no	no	no	no
K(+) efflux antiporter 1, chloroplastic {ECO:0000303|PubMed:11500563}	yes	no	no	no	no
RAN GTPase-activating protein 1	yes	no	no	no	no
Reticulon-like protein B10	yes	no	no	no	no
NAD(P)H-quinone oxidoreductase subunit T, chloroplastic {ECO:0000305}	yes	no	no	no	no
Auxin response factor 7	yes	no	no	no	no
6,7-dimethyl-8-ribityllumazine synthase, chloroplastic	yes	no	no	no	no
Phosphoenolpyruvate carboxylase 3	yes	no	no	no	no
Coatomer subunit beta'-1	yes	no	no	no	no
Protein STAY-GREEN LIKE, chloroplastic	yes	no	no	no	no
bZIP transcription factor 60 {ECO:0000303|PubMed:18065552}	yes	no	no	no	no
Sn1-specific diacylglycerol lipase alpha	yes	no	no	no	no
IQ domain-containing protein IQM4 {ECO:0000305}	yes	no	no	no	no
ATP-dependent zinc metalloprotease FTSH 5, mitochondrial	yes	no	no	no	no
Probable aminotransferase ACS12	yes	no	no	no	no
Myosin-2	yes	no	no	no	no
Proline-rich receptor-like protein kinase PERK9	yes	no	no	no	no
Tyrosine-protein phosphatase DSP1 {ECO:0000305}	yes	no	no	no	no
Protein SEMI-ROLLED LEAF 2 {ECO:0000303|PubMed:26873975}	yes	no	no	no	no

Sulfite exporter TauE/SafE family protein 3 {ECO:0000312|EMBL:AEC07746.1}	yes	no	no	no	no
Molybdopterin biosynthesis protein CNX1	yes	no	no	no	no
Glucose-induced degradation protein 4 homolog	yes	no	no	no	no
DNA-directed RNA polymerase II subunit RPB2	yes	no	no	no	no
Protein ENHANCED DISEASE RESISTANCE 2	yes	no	no	no	no
Protein indeterminate-domain 12 {ECO:0000303|PubMed:16784536}	yes	no	no	no	no
RuBisCO large subunit-binding protein subunit alpha, chloroplastic	yes	no	no	no	no
Transcription initiation factor TFIID subunit 6	yes	no	no	no	no
Protein PHOSPHATE STARVATION RESPONSE 1 {ECO:0000250|UniProtKB:Q10LZ1}	yes	no	no	no	no
Protein DETOXIFICATION 33 {ECO:0000303|PubMed:11739388}	yes	no	no	no	no
Protein DETOXIFICATION 27 {ECO:0000303|PubMed:11739388}	yes	no	no	no	no
Importin subunit alpha-1b	yes	no	no	no	no
Nudix hydrolase 8	yes	no	no	no	no
Receptor-like protein kinase FERONIA	yes	no	no	no	no
Protein NRT1/ PTR FAMILY 2.11	yes	no	no	no	no
Probable ethylene response sensor 2 {ECO:0000305}	yes	no	no	no	no
Vacuolar protein sorting-associated protein 2 homolog 3	yes	no	no	no	no
Pentatricopeptide repeat-containing protein At3g29230	yes	no	no	no	no
Beta-glucosidase 31	yes	no	no	no	no
Probable phospholipid hydroperoxide glutathione peroxidase	yes	no	no	no	no
OBERON-like protein	yes	no	no	no	no
Uncharacterized protein At4g14100	yes	no	no	no	no
YTH domain-containing protein ECT2 {ECO:0000305}	yes	no	no	no	no
Phenylalanine–tRNA ligase, chloroplastic/mitochondrial {ECO:0000305}	yes	no	no	no	no
Probable aldo-keto reductase 2	yes	no	no	no	no
RNA-directed DNA methylation 4	yes	no	no	no	no
Phenylalanine ammonia-lyase	yes	no	no	no	no
Monothiol glutaredoxin-S11	yes	no	no	no	no
Transcription factor UNE10	yes	no	no	no	no
K(+) efflux antiporter 3, chloroplastic {ECO:0000303|PubMed:11500563}	yes	no	no	no	no
Fe(2+) transport protein 1	yes	no	no	no	no
ABC transporter C family member 3	yes	no	no	no	no
Rhodanese-like domain-containing protein 9, chloroplastic	yes	no	no	no	no
Obtusifoliol 14-alpha demethylase	yes	no	no	no	no
NADP-dependent malic enzyme	yes	no	no	no	no
WD repeat-containing protein 20	yes	no	no	no	no
Receptor kinase-like protein Xa21 {ECO:0000303|PubMed:22735448}	yes	no	no	no	no
Thylakoid lumenal 15 kDa protein 1, chloroplastic	yes	no	no	no	no
Nuclear/nucleolar GTPase 2 {ECO:0000303|PubMed:21205822}	yes	no	no	no	no
40S ribosomal protein S13	yes	no	no	no	no
Probable inactive serine/threonine-protein kinase scy1	yes	no	no	no	no
Glucose-6-phosphate/phosphate translocator 2, chloroplastic	yes	no	no	no	no
Transcription factor bHLH128	yes	no	no	no	no
Leucine aminopeptidase 2, chloroplastic	yes	no	no	no	no
Probable ubiquitin-conjugating enzyme E2 25	yes	no	no	no	no
Beta-1,3-galactosyltransferase 7	yes	no	no	no	no
ERAD-associated E3 ubiquitin-protein ligase HRD1 {ECO:0000305}	yes	no	no	no	no
Common plant regulatory factor 1	yes	no	no	no	no
Probable ribose-5-phosphate isomerase 4, chloroplastic	yes	no	no	no	no
MADS-box transcription factor 47 {ECO:0000305}	yes	no	no	no	no

Protein TIME FOR COFFEE	yes	no	no	no	no
Mediator of RNA polymerase II transcription subunit 12	no	yes	no	no	no
Transcription factor MYB60 {ECO:0000303|PubMed:18647406}	no	yes	no	no	no
U3 small nucleolar ribonucleoprotein protein IMP3	no	yes	no	no	no
Probable periplasmic serine protease do/HhoA-like	no	yes	no	no	no
U4/U6 small nuclear ribonucleoprotein Prp31 homolog {ECO:0000305}	no	yes	no	no	no
Cytochrome P450 84A1	no	yes	no	no	no
Nucleolar complex protein 2 homolog	no	yes	no	no	no
Lipoyl synthase, mitochondrial {ECO:0000255|HAMAP-Rule:MF_03128}	no	yes	no	no	no
Amino acid transporter AVT6A {ECO:0000305}	no	yes	no	no	no
PH, RCC1 and FYVE domains-containing protein 1 {ECO:0000303|PubMed:11563980}	no	yes	no	no	no
Bifunctional fucokinase/fucose pyrophosphorylase	no	yes	no	no	no
Cinnamoyl-CoA reductase 1 {ECO:0000305}	no	yes	no	no	no
Receptor-like serine/threonine-protein kinase SD1-8	no	yes	no	no	no
Probable aldo-keto reductase 1	no	yes	no	no	no
Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial	no	yes	no	no	no
Germ cell-less protein-like 1	no	yes	no	no	no
Light-harvesting complex-like protein OHP2, chloroplastic {ECO:0000305}	no	yes	no	no	no
Probable E3 ubiquitin-protein ligase ARI7	no	yes	no	no	no
ATP-citrate synthase alpha chain protein 2	no	yes	no	no	no
Transcription factor bHLH68	no	yes	no	no	no
Splicing factor U2af large subunit B	no	yes	no	no	no
Metal tolerance protein 7	no	yes	no	no	no
Disease resistance protein RGA2	no	yes	no	no	no
Diphosphomevalonate decarboxylase 2 {ECO:0000305}	no	yes	no	no	no
Dihydrolipoyllysine-residue acetyltransferase component 2 of pyruvate dehydrogenase complex, mitochondrial	no	yes	no	no	no
Increased DNA methylation 1 {ECO:0000303|PubMed:22700931}	no	yes	no	no	no
Probable LRR receptor-like serine/threonine-protein kinase At3g47570	no	yes	no	no	no
mRNA cap guanine-N7 methyltransferase 1	no	yes	no	no	no
Calcium-dependent protein kinase 3 {ECO:0000303|PubMed:12068094}	no	yes	no	no	no
Protein decapping 5	no	yes	no	yes	no
RNA polymerase II C-terminal domain phosphatase-like 2	no	yes	no	no	no
Pectin acetylesterase 5 {ECO:0000303|PubMed:25115560}	no	yes	no	no	no
Zinc finger protein CONSTANS-LIKE 4	no	yes	no	no	no
Zinc finger protein BRUTUS {ECO:0000303|PubMed:20675571}	no	yes	no	no	no
Calcium-transporting ATPase 5, plasma membrane-type {ECO:0000305}	no	no	yes	no	no
Suppressor of mec-8 and unc-52 protein homolog 1	no	no	yes	no	no
Protein CHLOROPLAST ENHANCING STRESS TOLERANCE, chloroplastic {ECO:0000305}	no	no	yes	no	no
Protein WRKY1	no	no	yes	no	no
BTB/POZ and TAZ domain-containing protein 3	no	no	yes	no	no
UPF0454 protein C12orf49 homolog	no	no	yes	no	no
F-box protein SKIP31	no	no	yes	no	no
Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform	no	no	yes	no	no
F-box protein At1g55000	no	no	yes	no	no
Protein SRG1	no	no	yes	no	no
SUPPRESSOR OF GAMMA RESPONSE 1 {ECO:0000303|PubMed:19549833}	no	no	yes	no	no
ATP synthase protein MI25	no	no	yes	no	no
Ubinuclein-1 {ECO:0000303|PubMed:25086063}	no	no	yes	no	no

Squamous cell carcinoma antigen recognized by T-cells 3 {ECO:0000305}	no	no	yes	no	no
Protein PHYTOCHROME-DEPENDENT LATE-FLOWERING {ECO:0000303|PubMed:24127609}	no	no	yes	no	no
Uncharacterized methyltransferase At2g41040, chloroplastic	no	no	yes	no	no
E3 ubiquitin ligase PQT3-like	no	no	yes	no	no
Auxin transport protein BIG	no	no	yes	no	no
Protein STAY-GREEN, chloroplastic	no	no	no	yes	no
NAD-dependent protein deacylase SRT2 {ECO:0000255|HAMAP-Rule:MF_03161}	no	no	no	yes	no
Putative anthocyanidin reductase {ECO:0000303|PubMed:16399014}	no	no	no	yes	no
ATP-dependent DNA helicase RecQ	no	no	no	yes	no
E3 ubiquitin-protein ligase UPL4	no	no	no	yes	no
RNA cytidine acetyltransferase 1 {ECO:0000255|HAMAP-Rule:MF_03211}	no	no	no	yes	no
Cyclin-dependent kinase E-1	no	no	no	yes	no
E3 ubiquitin protein ligase RIN2	no	no	no	yes	no
Probable ubiquitin-conjugating enzyme E2 26	no	no	no	yes	no
Protein NRT1/ PTR FAMILY 8.1	no	no	no	yes	no
Glycine–tRNA ligase, chloroplastic/mitochondrial 2 {ECO:0000305}	no	no	no	yes	no
TBC1 domain family member 15	no	no	no	yes	no
Protoheme IX farnesyltransferase, mitochondrial	no	no	no	yes	no
UPF0481 protein At3g47200	no	no	no	yes	no
DENN domain-containing protein 5B	no	no	no	yes	no
Far upstream element-binding protein 1	no	no	no	yes	no
Polyadenylate-binding protein RBP45	no	no	no	yes	no
Magnesium transporter MRS2-E	no	no	no	yes	no
Protein transport protein Sec24-like CEF	no	no	no	yes	no
Zeaxanthin epoxidase, chloroplastic	no	no	no	yes	no
Protein CHLORORESPIRATORY REDUCTION 6, chloroplastic {ECO:0000305}	no	no	no	yes	no
Probable inactive beta-glucosidase 33	no	no	no	yes	no
NAC domain-containing protein 17 {ECO:0000303|PubMed:15029955}	no	no	no	yes	no
Uncharacterized protein sll0005	no	no	no	yes	no
Protein ETHYLENE-INSENSITIVE 3-like 1b {ECO:0000303|PubMed:16786297}	no	no	no	yes	no
DEAD-box ATP-dependent RNA helicase 7	no	no	no	yes	no
Peroxiredoxin-2C	no	no	no	yes	no
Nascent polypeptide-associated complex subunit beta {ECO:0000255|RuleBase:RU361272}	no	no	no	yes	no
Two-component response regulator ORR24 {ECO:0000305}	no	no	no	yes	no
NADP-dependent D-sorbitol-6-phosphate dehydrogenase	no	no	no	yes	no
Signal recognition particle receptor subunit beta	no	no	no	yes	no
Boron transporter 4	no	no	no	yes	no
Type I inositol polyphosphate 5-phosphatase 1 {ECO:0000303|PubMed:11402208}	no	no	no	yes	no
Phosphoenolpyruvate carboxykinase (ATP) 2	no	no	no	yes	no
Protein TOPLESS-RELATED PROTEIN 2 {ECO:0000303|PubMed:24336200}	no	no	no	yes	no
Replication protein A 70 kDa DNA-binding subunit B	no	no	no	yes	no
F-box protein PP2-B10	no	no	no	yes	no
Tryptophan aminotransferase-related protein 2 {ECO:0000303|PubMed:22582989}	no	no	no	yes	no
Putative vacuolar protein sorting-associated protein 13A	no	no	no	yes	no
Transcription factor TGAL7 {ECO:0000305}	no	no	no	no	yes
Calmodulin-binding protein 60 B {ECO:0000303|PubMed:11782485}	no	no	no	no	yes
Aldehyde dehydrogenase family 7 member A1	no	no	no	no	yes
Protein ANTHESIS POMOTING FACTOR 1 {ECO:0000303|PubMed:27968983}	no	no	no	no	yes
Probable protein phosphatase 2C 39	no	no	no	no	yes

UTP–glucose-1-phosphate uridylyltransferase	no	no	no	no	yes
Auxilin-related protein 2	no	no	no	no	yes
Craniofacial development protein 2	no	no	no	no	yes
RNA-binding protein 42	no	no	no	no	yes
Probable calcium-binding protein CML27	no	no	no	no	yes
Uncharacterized protein MJ1408	no	no	no	no	yes
Putative adagio-like protein 2	no	no	no	no	yes
Probable protein phosphatase 2C 55	no	no	no	no	yes
Glucose-6-phosphate isomerase 1, chloroplastic	no	no	no	no	yes
Arginine–tRNA ligase, chloroplastic/mitochondrial {ECO:0000305}	no	no	no	no	yes
Trans-cinnamate 4-monooxygenase	no	no	no	no	yes
RNA polymerase sigma factor sigE, chloroplastic/mitochondrial	no	no	no	no	yes
Glutamate receptor 3.1	no	no	no	no	yes
Pentatricopeptide repeat-containing protein At5g67570, chloroplastic	no	no	no	no	yes
E3 ubiquitin-protein ligase RHF2A {ECO:0000305}	no	no	no	no	yes
Protein CHROMATIN REMODELING 20 {ECO:0000303|PubMed:16547115}	no	no	no	no	yes
Iron-sulfur assembly protein IscA-like 1, mitochondrial	no	no	no	no	yes
Basic leucine zipper and W2 domain-containing protein 2	no	no	no	no	yes
Probable transcriptional regulator SLK2	no	no	no	no	yes
BAG family molecular chaperone regulator 6 {ECO:0000303|Ref.3}	no	no	no	no	yes
Polyamine oxidase 4 {ECO:0000303|PubMed:21796433}	no	no	no	no	yes
Beta-glucosidase 5	no	no	no	no	yes
Heat stress transcription factor A-2d	no	no	no	no	yes
Stress enhanced protein 1, chloroplastic {ECO:0000305}	no	no	no	no	yes
CBS domain-containing protein CBSCBSPB3	no	no	no	no	yes
Calcium-transporting ATPase 10, plasma membrane-type {ECO:0000305}	no	no	no	no	yes
Dehydrogenase/reductase SDR family member 7	no	no	no	no	yes
Flowering-promoting factor 1-like protein 2	no	no	no	no	yes
Apoptosis-inducing factor homolog A	no	no	no	no	yes
Phosphoenolpyruvate carboxylase kinase 2	no	no	no	no	yes
ATP synthase subunit a	no	no	no	no	yes
Protein OBERON 2 {ECO:0000303|PubMed:18403411}	no	no	no	no	yes
Uncharacterized aarF domain-containing protein kinase At5g05200, chloroplastic	no	no	no	no	yes
Pre-mRNA-processing factor 39	no	no	no	no	yes
Pentatricopeptide repeat-containing protein At1g50270	no	no	no	no	yes
Autophagy-related protein 13b {ECO:0000303|PubMed:12114572}	no	no	no	no	yes
Probable feruloyl esterase A	no	no	no	no	yes
Phosphoribosylamine–glycine ligase, chloroplastic	no	no	no	no	yes
Inositol-3-phosphate synthase	no	no	no	no	yes
Probable 6-phosphogluconolactonase 4, chloroplastic	no	no	no	no	yes
Peptide chain release factor PrfB3, chloroplastic {ECO:0000303|PubMed:21771930}	no	no	no	no	yes
Cytochrome P450 98A1	no	no	no	no	yes
Protein BIC1 {ECO:0000303|PubMed:27846570}	no	no	no	no	yes
Glutathione S-transferase T1	no	no	no	no	yes
ATP-dependent RNA helicase DEAH12, chloroplastic	no	no	no	no	yes
Guanylate-binding protein 7	no	no	no	no	yes
Calcium-dependent protein kinase 9 {ECO:0000305}	no	no	no	no	yes
Mitochondrial-processing peptidase subunit alpha	no	no	no	no	yes
Probable allantoinase	no	no	no	no	yes
Myb-related protein P	no	no	no	no	yes

Table 6: Number of genes with allele-specific expression for each functional term, for all accessions. The 300 most frequent GO terms are shown.

Term ID	Frequency	Description
GO:0003674	1299	molecular_function
GO:0008150	1238	biological_process
GO:0005575	1233	cellular_component
GO:0044464	1161	cell part
GO:0044424	1043	intracellular part
GO:0005488	1033	binding
GO:0009987	909	cellular process
GO:0043226	795	organelle
GO:0043229	794	intracellular organelle
GO:0008152	775	metabolic process
GO:0043227	765	membrane-bounded organelle
GO:0043231	749	intracellular membrane-bounded organelle
GO:0003824	724	catalytic activity
GO:0071704	709	organic substance metabolic process
GO:0044237	674	cellular metabolic process
GO:0097159	670	organic cyclic compound binding
GO:1901363	670	heterocyclic compound binding
GO:0044238	639	primary metabolic process
GO:0043167	637	ion binding
GO:0044444	628	cytoplasmic part
GO:0006807	552	nitrogen compound metabolic process
GO:0065007	481	biological regulation
GO:0043170	468	macromolecule metabolic process
GO:0044422	467	organelle part
GO:0044446	464	intracellular organelle part
GO:0050789	447	regulation of biological process
GO:0044260	445	cellular macromolecule metabolic process
GO:0016020	432	membrane
GO:0005634	411	nucleus
GO:0050896	404	response to stimulus
GO:0050794	388	regulation of cellular process
GO:0036094	383	small molecule binding
GO:0043168	383	anion binding
GO:0005515	381	protein binding
GO:1901564	378	organonitrogen compound metabolic process
GO:0044425	368	membrane part
GO:0000166	361	nucleotide binding
GO:1901265	361	nucleoside phosphate binding
GO:0043169	342	cation binding
GO:0097367	342	carbohydrate derivative binding
GO:0046872	340	metal ion binding
GO:0032553	337	ribonucleotide binding
GO:0017076	336	purine nucleotide binding
GO:0032555	335	purine ribonucleotide binding
GO:0035639	330	purine ribonucleoside triphosphate binding
GO:0016740	327	transferase activity
GO:0003676	324	nucleic acid binding

GO:0031224	319	intrinsic component of membrane
GO:0016021	312	integral component of membrane
GO:0030554	312	adenyl nucleotide binding
GO:0032559	311	adenyl ribonucleotide binding
GO:0005524	306	ATP binding
GO:0006950	302	response to stress
GO:0034641	278	cellular nitrogen compound metabolic process
GO:0019538	277	protein metabolic process
GO:1901360	272	organic cyclic compound metabolic process
GO:0006725	263	cellular aromatic compound metabolic process
GO:0016787	256	hydrolase activity
GO:0032991	256	protein-containing complex
GO:0046483	252	heterocycle metabolic process
GO:0009058	242	biosynthetic process
GO:0019222	242	regulation of metabolic process
GO:0044267	235	cellular protein metabolic process
GO:0043412	234	macromolecule modification
GO:0005737	232	cytoplasm
GO:0006139	226	nucleobase-containing compound metabolic process
GO:1901576	224	organic substance biosynthetic process
GO:0031323	219	regulation of cellular metabolic process
GO:0051179	218	localization
GO:0044249	217	cellular biosynthetic process
GO:0060255	217	regulation of macromolecule metabolic process
GO:0005886	214	plasma membrane
GO:0006464	213	cellular protein modification process
GO:0036211	213	protein modification process
GO:0051234	210	establishment of localization
GO:0071840	207	cellular component organization or biogenesis
GO:0003677	206	DNA binding
GO:0006810	205	transport
GO:0080090	202	regulation of primary metabolic process
GO:0016043	201	cellular component organization
GO:0051171	201	regulation of nitrogen compound metabolic process
GO:0005829	197	cytosol
GO:0090304	195	nucleic acid metabolic process
GO:0010468	189	regulation of gene expression
GO:0032502	181	developmental process
GO:0009536	178	plastid
GO:0009889	175	regulation of biosynthetic process
GO:0031326	173	regulation of cellular biosynthetic process
GO:0019219	171	regulation of nucleobase-containing compound metabolic process
GO:0006793	170	phosphorus metabolic process
GO:0016772	169	transferase activity, transferring phosphorus-containing groups
GO:0010556	168	regulation of macromolecule biosynthetic process
GO:2000112	168	regulation of cellular macromolecule biosynthetic process
GO:0006796	166	phosphate-containing compound metabolic process
GO:0051252	162	regulation of RNA metabolic process
GO:0031090	159	organelle membrane
GO:0042221	158	response to chemical

GO:0006355	152	regulation of transcription, DNA-templated
GO:1903506	152	regulation of nucleic acid-templated transcription
GO:2001141	152	regulation of RNA biosynthetic process
GO:0044281	147	small molecule metabolic process
GO:0009507	144	chloroplast
GO:0016301	143	kinase activity
GO:0007165	139	signal transduction
GO:0009628	139	response to abiotic stimulus
GO:0051716	139	cellular response to stimulus
GO:0044428	138	nuclear part
GO:0009056	136	catabolic process
GO:0006952	135	defense response
GO:0016491	134	oxidoreductase activity
GO:0016773	132	phosphotransferase activity, alcohol group as acceptor
GO:0032501	127	multicellular organismal process
GO:0048518	127	positive regulation of biological process
GO:0003723	125	RNA binding
GO:0016070	125	RNA metabolic process
GO:0098588	122	bounding membrane of organelle
GO:0048519	120	negative regulation of biological process
GO:1901575	118	organic substance catabolic process
GO:0048856	117	anatomical structure development
GO:0009605	116	response to external stimulus
GO:0022414	113	reproductive process
GO:0004672	109	protein kinase activity
GO:0016310	109	phosphorylation
GO:0033554	109	cellular response to stress
GO:0005215	108	transporter activity
GO:0010033	108	response to organic substance
GO:0022857	108	transmembrane transporter activity
GO:0006629	106	lipid metabolic process
GO:0048583	106	regulation of response to stimulus
GO:0044248	104	cellular catabolic process
GO:0071702	104	organic substance transport
GO:0016788	101	hydrolase activity, acting on ester bonds
GO:1901700	101	response to oxygen-containing compound
GO:0044434	100	chloroplast part
GO:0044435	100	plastid part
GO:0055114	99	oxidation-reduction process
GO:0004674	98	protein serine/threonine kinase activity
GO:0006996	98	organelle organization
GO:0006082	97	organic acid metabolic process
GO:0043436	97	oxoacid metabolic process
GO:0046914	97	transition metal ion binding
GO:0048522	97	positive regulation of cellular process
GO:0006468	96	protein phosphorylation
GO:0051704	95	multi-organism process
GO:1901362	95	organic cyclic compound biosynthetic process
GO:0005739	94	mitochondrion
GO:0044271	94	cellular nitrogen compound biosynthetic process
GO:0098805	93	whole membrane

GO:0019752	89	carboxylic acid metabolic process
GO:0009059	88	macromolecule biosynthetic process
GO:0003700	87	DNA-binding transcription factor activity
GO:0071705	87	nitrogen compound transport
GO:0009607	86	response to biotic stimulus
GO:0043228	86	non-membrane-bounded organelle
GO:0043232	86	intracellular non-membrane-bounded organelle
GO:0016817	85	hydrolase activity, acting on acid anhydrides
GO:0019438	85	aromatic compound biosynthetic process
GO:0043207	85	response to external biotic stimulus
GO:0006259	84	DNA metabolic process
GO:0016462	84	pyrophosphatase activity
GO:0016818	84	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
GO:0048523	83	negative regulation of cellular process
GO:0051707	83	response to other organism
GO:0017111	82	nucleoside-triphosphatase activity
GO:0034645	81	cellular macromolecule biosynthetic process
GO:0065008	81	regulation of biological quality
GO:0009893	80	positive regulation of metabolic process
GO:0033036	80	macromolecule localization
GO:0009719	79	response to endogenous stimulus
GO:0044255	79	cellular lipid metabolic process
GO:0055085	79	transmembrane transport
GO:0008104	78	protein localization
GO:0018130	78	heterocycle biosynthetic process
GO:0051641	78	cellular localization
GO:1902494	76	catalytic complex
GO:0001101	74	response to acid chemical
GO:0003006	72	developmental process involved in reproduction
GO:0031325	72	positive regulation of cellular metabolic process
GO:0045184	72	establishment of protein localization
GO:0070647	72	protein modification by small protein conjugation or removal
GO:0006396	71	RNA processing
GO:0009725	71	response to hormone
GO:0015833	71	peptide transport
GO:0042886	71	amide transport
GO:0015031	70	protein transport
GO:0032446	69	protein modification by small protein conjugation
GO:0046983	69	protein dimerization activity
GO:0008270	68	zinc ion binding
GO:0010604	68	positive regulation of macromolecule metabolic process
GO:0005975	67	carbohydrate metabolic process
GO:0006811	67	ion transport
GO:0050793	67	regulation of developmental process
GO:1901566	67	organonitrogen compound biosynthetic process
GO:0016567	66	protein ubiquitination
GO:0048869	66	cellular developmental process
GO:0005794	65	Golgi apparatus
GO:0051649	65	establishment of localization in cell
GO:0006508	64	proteolysis

GO:0009755	64	hormone-mediated signaling pathway
GO:0048037	64	cofactor binding
GO:0051173	64	positive regulation of nitrogen compound metabolic process
GO:0080134	64	regulation of response to stress
GO:0010035	63	response to inorganic substance
GO:0098542	63	defense response to other organism
GO:0015075	62	ion transmembrane transporter activity
GO:1990904	62	ribonucleoprotein complex
GO:0019637	61	organophosphate metabolic process
GO:0046907	60	intracellular transport
GO:0007275	59	multicellular organism development
GO:0009057	59	macromolecule catabolic process
GO:0016071	59	mRNA metabolic process
GO:0035556	58	intracellular signal transduction
GO:0009892	57	negative regulation of metabolic process
GO:0022804	57	active transmembrane transporter activity
GO:0034654	57	nucleobase-containing compound biosynthetic process
GO:0044432	57	endoplasmic reticulum part
GO:0051239	57	regulation of multicellular organismal process
GO:0043565	56	sequence-specific DNA binding
GO:0005773	55	vacuole
GO:0009314	55	response to radiation
GO:0044283	55	small molecule biosynthetic process
GO:0044437	55	vacuolar part
GO:0005774	54	vacuolar membrane
GO:0005576	53	extracellular region
GO:0008610	53	lipid biosynthetic process
GO:0010605	53	negative regulation of macromolecule metabolic process
GO:0030054	53	cell junction
GO:0044265	53	cellular macromolecule catabolic process
GO:1901565	52	organonitrogen compound catabolic process
GO:2000026	52	regulation of multicellular organismal development
GO:0005783	51	endoplasmic reticulum
GO:0009416	51	response to light stimulus
GO:0043531	51	ADP binding
GO:0005911	50	cell-cell junction
GO:0009506	50	plasmodesma
GO:0009891	50	positive regulation of biosynthetic process
GO:0019787	50	ubiquitin-like protein transferase activity
GO:0031982	50	vesicle
GO:0005789	49	endoplasmic reticulum membrane
GO:0006397	49	mRNA processing
GO:0006886	49	intracellular protein transport
GO:0006974	49	cellular response to DNA damage stimulus
GO:0019899	49	enzyme binding
GO:0022607	49	cellular component assembly
GO:0032787	49	monocarboxylic acid metabolic process
GO:0044431	49	Golgi apparatus part
GO:0051186	49	cofactor metabolic process
GO:1901135	49	carbohydrate derivative metabolic process
GO:1990234	49	transferase complex

GO:0002376	48	immune system process
GO:0004842	47	ubiquitin-protein transferase activity
GO:0010628	47	positive regulation of gene expression
GO:0010646	47	regulation of cell communication
GO:0031328	47	positive regulation of cellular biosynthetic process
GO:0050662	47	coenzyme binding
GO:0010557	46	positive regulation of macromolecule biosynthetic process
GO:0023051	46	regulation of signaling
GO:0031410	46	cytoplasmic vesicle
GO:0045935	46	positive regulation of nucleobase-containing compound metabolic process
GO:0051603	46	proteolysis involved in cellular protein catabolic process
GO:0097708	46	intracellular vesicle
GO:0010629	45	negative regulation of gene expression
GO:0042802	45	identical protein binding
GO:0044723	45	
GO:0051128	45	regulation of cellular component organization
GO:0070887	45	cellular response to chemical stimulus
GO:0009966	44	regulation of signal transduction
GO:0042592	44	homeostatic process
GO:0051254	44	positive regulation of RNA metabolic process
GO:0008324	43	cation transmembrane transporter activity
GO:0009617	43	response to bacterium
GO:0022402	43	cell cycle process
GO:0033993	43	response to lipid
GO:0043933	43	protein-containing complex subunit organization
GO:0044429	43	mitochondrial part
GO:0044451	43	nucleoplasm part
GO:0045893	43	positive regulation of transcription, DNA-templated
GO:1902680	43	positive regulation of RNA biosynthetic process
GO:1903508	43	positive regulation of nucleic acid-templated transcription
GO:0009532	42	plastid stroma
GO:0009570	42	chloroplast stroma
GO:0031347	42	regulation of defense response
GO:0006281	41	DNA repair
GO:0006970	41	response to osmotic stress
GO:0016192	41	vesicle-mediated transport
GO:0048585	41	negative regulation of response to stimulus
GO:0006511	40	ubiquitin-dependent protein catabolic process
GO:0009791	40	post-embryonic development
GO:0016053	40	organic acid biosynthetic process
GO:0019941	40	modification-dependent protein catabolic process
GO:0043632	40	modification-dependent macromolecule catabolic process
GO:0046394	40	carboxylic acid biosynthetic process
GO:0051246	40	regulation of protein metabolic process
GO:0000975	39	
GO:0001067	39	regulatory region nucleic acid binding
GO:0006310	39	DNA recombination
GO:0006325	39	chromatin organization
GO:0008233	39	peptidase activity
GO:0016829	39	lyase activity

GO:0044212	39	transcription regulatory region DNA binding

Table 7: Frequency of genes with allele-specific expression in Gene Ontology terms for each genotype. The 300 most frequent GO terms are shown in this table.

Term ID	Description	IN84-58	White Transparent	RB72454	SP80-3280	US85-1008	SES205.A
GO:0003674	molecular_function	545	560	581	611	598	414
GO:0005575	cellular_component	528	529	550	577	561	397
GO:0008150	biological_process	515	533	552	583	560	399
GO:0044464	cell part	493	496	517	542	525	369
GO:0044424	intracellular part	438	456	469	489	478	331
GO:0005488	binding	429	445	460	482	465	337
GO:0009987	cellular process	380	397	408	431	404	279
GO:0043226	organelle	346	355	372	393	375	259
GO:0043229	intracellular organelle	345	355	372	393	375	259
GO:0043227	membrane-bounded organelle	332	339	354	379	359	247
GO:0043231	intracellular membrane-bounded organelle	325	332	345	369	353	244
GO:0008152	metabolic process	323	329	345	363	347	234
GO:0003824	catalytic activity	302	304	314	333	332	223
GO:0071704	organic substance metabolic process	293	298	313	330	314	208
GO:0044237	cellular metabolic process	278	289	298	320	299	205
GO:0097159	organic cyclic compound binding	265	280	301	315	292	223
GO:1901363	heterocyclic compound binding	265	280	301	315	292	223
GO:0044238	primary metabolic process	263	271	287	300	282	183
GO:0044444	cytoplasmic part	263	276	283	290	309	202
GO:0043167	ion binding	256	258	282	303	283	212
GO:0006807	nitrogen compound metabolic process	229	238	252	263	245	163
GO:0065007	biological regulation	208	206	220	234	206	152
GO:0043170	macromolecule metabolic process	197	205	213	223	197	129
GO:0050789	regulation of biological process	196	195	209	216	192	142
GO:0044422	organelle part	188	207	218	214	228	140
GO:0044446	intracellular organelle part	185	204	216	211	225	138
GO:0044260	cellular macromolecule metabolic process	184	196	203	213	185	124
GO:0016020	membrane	182	173	189	199	206	145
GO:0005634	nucleus	174	191	186	211	167	130
GO:0050794	regulation of cellular process	172	168	183	189	172	127
GO:1901564	organonitrogen compound metabolic process	157	159	168	182	167	113
GO:0050896	response to stimulus	155	163	180	202	176	135
GO:0044425	membrane part	153	148	162	177	173	124

GO:0005515	protein binding	151	178	171	180	183	122
GO:0036094	small molecule binding	147	152	173	189	176	132
GO:0043169	cation binding	147	151	160	163	155	112
GO:0046872	metal ion binding	146	151	159	162	154	111
GO:0043168	anion binding	144	151	170	189	169	130
GO:0000166	nucleotide binding	140	143	162	175	160	121
GO:1901265	nucleoside phosphate binding	140	143	162	175	160	121
GO:0003676	nucleic acid binding	135	141	147	145	134	102
GO:0031224	intrinsic component of membrane	135	131	141	149	151	112
GO:0016021	integral component of membrane	133	130	139	148	150	108
GO:0016740	transferase activity	131	144	143	153	146	87
GO:0017076	purine nucleotide binding	129	131	148	159	146	112
GO:0097367	carbohydrate derivative binding	129	134	153	163	152	113
GO:0032553	ribonucleotide binding	128	132	149	159	147	111
GO:0032555	purine ribonucleotide binding	128	130	147	159	145	111
GO:0035639	purine ribonucleoside triphosphate binding	124	129	145	156	142	108
GO:0034641	cellular nitrogen compound metabolic process	117	126	133	126	126	84
GO:0030554	adenyl nucleotide binding	115	118	133	148	136	105
GO:0032559	adenyl ribonucleotide binding	114	117	132	148	135	104
GO:0019538	protein metabolic process	113	115	122	136	114	77
GO:1901360	organic cyclic compound metabolic process	113	124	131	123	123	82
GO:0016787	hydrolase activity	111	98	109	111	122	83
GO:0005524	ATP binding	110	116	130	145	132	101
GO:0006725	cellular aromatic compound metabolic process	110	118	125	121	117	77
GO:0019222	regulation of metabolic process	106	124	118	120	107	83
GO:0046483	heterocycle metabolic process	106	116	120	117	115	75
GO:0006950	response to stress	105	117	134	140	128	104
GO:0005737	cytoplasm	98	108	103	104	98	68
GO:0006139	nucleobase-containing compound metabolic process	98	106	112	105	100	64
GO:0043412	macromolecule modification	98	105	106	121	94	68
GO:0031323	regulation of cellular metabolic process	97	113	107	112	99	77
GO:0009058	biosynthetic process	96	107	113	113	113	76
GO:0044267	cellular protein metabolic process	95	102	103	119	93	68
GO:0060255	regulation of macromolecule metabolic process	95	111	105	107	90	72
GO:0032991	protein-containing complex	92	108	117	117	109	65
GO:1901576	organic substance biosynthetic process	91	96	104	104	102	67
GO:0051179	localization	89	96	101	104	112	72
GO:0003677	DNA binding	88	86	91	98	90	68
GO:0051171	regulation of nitrogen compound metabolic process	88	103	97	101	84	69

GO:0080090	regulation of primary metabolic process	88	105	98	102	84	69
GO:0006464	cellular protein modification process	87	94	94	109	83	63
GO:0010468	regulation of gene expression	87	101	92	98	81	62
GO:0036211	protein modification process	87	94	94	109	83	63
GO:0044249	cellular biosynthetic process	87	92	103	104	97	66
GO:0005886	plasma membrane	86	77	96	99	92	70
GO:0051234	establishment of localization	86	93	100	99	107	70
GO:0071840	cellular component organization or biogenesis	86	96	94	82	86	61
GO:0006810	transport	85	90	98	98	103	70
GO:0016043	cellular component organization	84	93	90	79	84	59
GO:0090304	nucleic acid metabolic process	83	90	95	90	87	59
GO:0009536	plastid	82	74	85	87	102	64
GO:0009889	regulation of biosynthetic process	81	94	83	90	75	60
GO:0005829	cytosol	80	94	93	86	102	69
GO:0031326	regulation of cellular biosynthetic process	80	93	82	89	73	59
GO:0019219	regulation of nucleobase-containing compound metabolic process	79	90	81	90	74	60
GO:0010556	regulation of macromolecule biosynthetic process	77	91	78	87	70	58
GO:2000112	regulation of cellular macromolecule biosynthetic process	77	91	78	87	70	58
GO:0051252	regulation of RNA metabolic process	75	87	77	88	70	57
GO:0032502	developmental process	74	88	78	82	81	54
GO:0006355	regulation of transcription, DNA-templated	72	82	71	83	64	52
GO:1903506	regulation of nucleic acid-templated transcription	72	82	71	83	64	52
GO:2001141	regulation of RNA biosynthetic process	72	82	71	83	64	52
GO:0009507	chloroplast	70	64	70	74	81	53
GO:0006793	phosphorus metabolic process	68	72	76	86	66	45
GO:0016772	transferase activity, transferring phosphorus-containing groups	68	70	73	80	68	48
GO:0042221	response to chemical	68	69	79	80	66	51
GO:0009628	response to abiotic stimulus	67	69	65	73	63	47
GO:0006796	phosphate-containing compound metabolic process	66	70	72	84	64	45
GO:0016491	oxidoreductase activity	61	57	57	66	59	49
GO:0031090	organelle membrane	61	61	60	66	83	49
GO:0044281	small molecule metabolic process	60	61	74	74	75	48
GO:0009056	catabolic process	57	54	57	60	53	41
GO:0016070	RNA metabolic process	57	65	66	62	59	35

GO:0032501	multicellular organismal process	57	68	65	57	54	36
GO:0016301	kinase activity	56	55	59	70	55	36
GO:0044428	nuclear part	56	70	68	68	67	34
GO:0048518	positive regulation of biological process	55	61	63	61	56	39
GO:0051716	cellular response to stimulus	55	53	68	68	61	33
GO:0007165	signal transduction	54	50	65	73	56	41
GO:0016773	phosphotransferase activity, alcohol group as acceptor	52	51	56	66	49	34
GO:0048856	anatomical structure development	52	60	49	55	50	33
GO:0022414	reproductive process	51	44	47	56	48	28
GO:0005215	transporter activity	49	53	56	56	56	40
GO:0022857	transmembrane transporter activity	49	53	56	56	56	40
GO:0044434	chloroplast part	49	47	54	52	61	36
GO:0044435	plastid part	49	47	54	52	61	36
GO:1901575	organic substance catabolic process	49	46	51	54	46	35
GO:0003700	DNA-binding transcription factor activity	47	45	40	47	41	31
GO:0003723	RNA binding	47	53	59	46	49	34
GO:0055114	oxidation-reduction process	46	38	42	47	46	34
GO:0016788	hydrolase activity, acting on ester bonds	45	41	44	41	52	29
GO:0010033	response to organic substance	44	45	51	56	46	37
GO:0048522	positive regulation of cellular process	44	52	49	47	44	29
GO:0006996	organelle organization	43	38	42	40	46	33
GO:0044271	cellular nitrogen compound biosynthetic process	43	43	47	43	46	30
GO:0098588	bounding membrane of organelle	43	44	45	48	60	39
GO:0004672	protein kinase activity	42	42	46	57	40	31
GO:0009059	macromolecule biosynthetic process	42	37	45	41	36	26
GO:1901700	response to oxygen-containing compound	42	40	48	48	35	36
GO:0006629	lipid metabolic process	41	45	40	51	53	33
GO:0009893	positive regulation of metabolic process	41	41	41	39	35	27
GO:0016310	phosphorylation	41	43	46	59	35	27
GO:0033554	cellular response to stress	41	38	52	50	45	25
GO:0046914	transition metal ion binding	41	46	43	45	43	26
GO:1901362	organic cyclic compound biosynthetic process	41	44	48	45	46	33
GO:0009605	response to external stimulus	40	42	50	57	50	36
GO:0044248	cellular catabolic process	40	42	44	49	41	33
GO:0048583	regulation of response to stimulus	40	41	52	53	51	33
GO:0006082	organic acid metabolic process	39	37	47	48	52	35

GO:0043228	non-membrane-bounded organelle	39	40	43	46	40	22
GO:0043232	intracellular non-membrane-bounded organelle	39	40	43	46	40	22
GO:0043436	oxoacid metabolic process	39	37	47	48	52	35
GO:0048519	negative regulation of biological process	39	53	61	51	48	34
GO:0016462	pyrophosphatase activity	38	32	35	35	39	26
GO:0016817	hydrolase activity, acting on acid anhydrides	38	32	36	35	39	26
GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	38	32	35	35	39	26
GO:0019438	aromatic compound biosynthetic process	38	38	43	41	41	27
GO:0031325	positive regulation of cellular metabolic process	38	37	38	36	31	23
GO:0004674	protein serine/threonine kinase activity	37	38	39	51	34	26
GO:0006468	protein phosphorylation	37	36	38	52	30	26
GO:0017111	nucleoside-triphosphatase activity	37	30	34	34	38	25
GO:0034645	cellular macromolecule biosynthetic process	37	32	41	37	34	23
GO:0005739	mitochondrion	36	40	45	43	46	34
GO:0006952	defense response	36	45	56	62	54	47
GO:0009719	response to endogenous stimulus	36	33	37	39	31	29
GO:1902494	catalytic complex	36	34	42	38	37	17
GO:0010604	positive regulation of macromolecule metabolic process	35	37	37	34	27	22
GO:0018130	heterocycle biosynthetic process	35	37	39	38	41	28
GO:0019752	carboxylic acid metabolic process	35	34	45	44	47	29
GO:0055085	transmembrane transport	35	38	41	40	40	38
GO:0005975	carbohydrate metabolic process	34	27	30	31	28	17
GO:0051173	positive regulation of nitrogen compound metabolic process	34	34	35	32	25	19
GO:0065008	regulation of biological quality	34	28	33	33	39	24
GO:0071702	organic substance transport	34	43	49	51	53	28
GO:0001101	response to acid chemical	33	27	34	34	26	24
GO:0006259	DNA metabolic process	33	30	39	36	36	26
GO:0071705	nitrogen compound transport	33	39	41	42	45	24
GO:0003006	developmental process involved in reproduction	32	29	34	32	32	21
GO:0006811	ion transport	32	31	34	29	32	25
GO:0044255	cellular lipid metabolic process	32	32	29	41	42	26
GO:0048869	cellular developmental process	32	37	33	33	29	22
GO:0005794	Golgi apparatus	31	23	24	24	36	22
GO:0009725	response to hormone	31	32	32	33	29	26

GO:0009891	positive regulation of biosynthetic process	31	30	28	27	23	16
GO:0015075	ion transmembrane transporter activity	31	34	37	32	34	22
GO:0033036	macromolecule localization	31	36	33	40	43	20
GO:0008104	protein localization	30	35	32	38	41	19
GO:0010628	positive regulation of gene expression	30	29	27	26	21	15
GO:0031328	positive regulation of cellular biosynthetic process	30	28	27	25	21	14
GO:0051704	multi-organism process	30	30	39	47	41	35
GO:0070647	protein modification by small protein conjugation or removal	30	34	36	36	30	16
GO:0010035	response to inorganic substance	29	25	30	33	23	22
GO:0010557	positive regulation of macromolecule biosynthetic process	29	29	26	26	20	15
GO:0022804	active transmembrane transporter activity	29	28	30	31	33	25
GO:0032446	protein modification by small protein conjugation	29	32	35	33	29	14
GO:0034654	nucleobase-containing compound biosynthetic process	29	27	32	29	28	19
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	29	27	25	25	21	14
GO:0051641	cellular localization	29	34	36	37	40	20
GO:0098805	whole membrane	29	33	35	33	46	29
GO:0007275	multicellular organism development	28	35	28	28	25	15
GO:0016567	protein ubiquitination	28	31	34	31	28	14
GO:0045893	positive regulation of transcription, DNA-templated	28	27	25	25	19	14
GO:0048523	negative regulation of cellular process	28	33	43	35	33	24
GO:0051254	positive regulation of RNA metabolic process	28	27	25	25	20	14
GO:1902680	positive regulation of RNA biosynthetic process	28	27	25	25	19	14
GO:1903508	positive regulation of nucleic acid-templated transcription	28	27	25	25	19	14
GO:0043565	sequence-specific DNA binding	27	20	26	28	28	18
GO:0045184	establishment of protein localization	27	33	32	35	38	19
GO:0006396	RNA processing	26	40	36	33	29	15
GO:0009314	response to radiation	26	31	24	31	27	15
GO:0015031	protein transport	26	31	31	34	37	19
GO:0015833	peptide transport	26	32	32	34	37	19
GO:0019637	organophosphate metabolic process	26	29	29	28	31	16
GO:0042886	amide transport	26	32	32	34	37	19
GO:0046983	protein dimerization activity	26	31	31	37	32	25

GO:0008270	zinc ion binding	25	34	28	33	30	15
GO:0009607	response to biotic stimulus	25	28	34	42	35	28
GO:0009755	hormone-mediated signaling pathway	25	28	32	39	29	22
GO:1901566	organonitrogen compound biosynthetic process	25	27	28	27	32	21
GO:0009416	response to light stimulus	24	28	22	28	26	14
GO:0043207	response to external biotic stimulus	24	28	34	42	35	27
GO:0050793	regulation of developmental process	24	30	28	28	25	18
GO:0051649	establishment of localization in cell	24	26	32	28	33	18
GO:0051707	response to other organism	24	27	34	42	34	27
GO:1990234	transferase complex	24	21	31	24	27	9
GO:0000975		23	23	17	18	16	15
GO:0001067	regulatory region nucleic acid binding	23	23	17	18	16	15
GO:0006970	response to osmotic stress	23	17	22	16	15	14
GO:0044212	transcription regulatory region DNA binding	23	23	17	18	16	15
GO:0044723		23	18	20	19	19	11
GO:0006508	proteolysis	22	26	29	30	26	13
GO:0016071	mRNA metabolic process	22	31	30	26	26	14
GO:0030054	cell junction	22	21	26	26	26	25
GO:0031982	vesicle	22	21	24	27	22	10
GO:0032787	monocarboxylic acid metabolic process	22	19	24	27	26	16
GO:0046907	intracellular transport	22	25	30	28	30	18
GO:1901135	carbohydrate derivative metabolic process	22	21	24	27	23	13
GO:0005911	cell-cell junction	21	20	25	24	24	23
GO:0008324	cation transmembrane transporter activity	21	24	27	20	23	15
GO:0009057	macromolecule catabolic process	21	24	25	27	17	11
GO:0009506	plasmodesma	21	20	25	24	24	23
GO:0015291	secondary active transmembrane transporter activity	21	19	21	23	22	17
GO:0044432	endoplasmic reticulum part	21	22	25	26	26	22
GO:0051128	regulation of cellular component organization	21	23	22	23	23	16
GO:0080134	regulation of response to stress	21	22	30	36	33	19
GO:0005773	vacuole	20	21	28	25	31	24
GO:0008610	lipid biosynthetic process	20	24	20	28	24	18
GO:0009966	regulation of signal transduction	20	16	26	21	18	16
GO:0010646	regulation of cell communication	20	16	26	22	19	18
GO:0022402	cell cycle process	20	14	19	15	13	14
GO:0023051	regulation of signaling	20	16	26	22	19	17

GO:0032774	RNA biosynthetic process	20	15	20	18	17	14
GO:0035556	intracellular signal transduction	20	23	26	30	20	18
GO:0044431	Golgi apparatus part	20	16	17	23	21	17
GO:0048037	cofactor binding	20	33	32	35	32	23
GO:0051186	cofactor metabolic process	20	24	28	26	23	16
GO:1901565	organonitrogen compound catabolic process	20	20	25	24	25	17
GO:1990904	ribonucleoprotein complex	20	31	26	27	23	15
GO:0005516	calmodulin binding	19	17	18	20	19	11
GO:0006886	intracellular protein transport	19	22	23	24	25	13
GO:0008233	peptidase activity	19	13	17	16	19	13
GO:0009532	plastid stroma	19	19	24	24	24	15
GO:0009570	chloroplast stroma	19	19	24	24	24	15
GO:0009791	post-embryonic development	19	18	17	15	14	12
GO:0031410	cytoplasmic vesicle	19	18	22	25	20	10
GO:0033993	response to lipid	19	19	21	21	16	16
GO:0044283	small molecule biosynthetic process	19	25	30	28	29	14
GO:0044451	nucleoplasm part	19	23	27	25	22	13
GO:0051239	regulation of multicellular organismal process	19	29	26	24	22	17
GO:0070011	peptidase activity, acting on L-amino acid peptides	19	13	17	16	19	13
GO:0097708	intracellular vesicle	19	18	22	25	20	10
GO:0006325	chromatin organization	18	25	23	22	17	14
GO:0006397	mRNA processing	18	26	26	22	21	10
GO:0009892	negative regulation of metabolic process	18	30	28	22	27	17
GO:0019787	ubiquitin-like protein transferase activity	18	22	24	27	25	9
GO:0030246	carbohydrate binding	18	10	14	18	10	10
GO:0031967	organelle envelope	18	14	19	17	18	10
GO:0031975	envelope	18	14	19	17	18	10
GO:0042592	homeostatic process	18	15	19	19	19	12
GO:0044430	cytoskeletal part	18	15	12	13	15	8
GO:0070887	cellular response to chemical stimulus	18	16	22	19	19	9
GO:2000026	regulation of multicellular organismal development	18	26	24	21	21	16
GO:0005576	extracellular region	17	16	19	21	27	16
GO:0006310	DNA recombination	17	15	15	18	14	15
GO:0006357	regulation of transcription by RNA polymerase II	17	17	19	17	14	7
GO:0006812	cation transport	17	16	18	13	13	12
GO:0006974	cellular response to DNA damage stimulus	17	20	22	23	19	14
GO:0008092	cytoskeletal protein binding	17	15	11	13	16	13
GO:0010605	negative regulation of macromolecule metabolic process	17	28	27	19	24	14
GO:0016757	transferase activity, transferring glycosyl groups	17	12	11	10	14	9

GO:0032259	methylation	17	16	17	16	16	10
GO:0044265	cellular macromolecule catabolic process	17	22	23	25	15	10
GO:0097305	response to alcohol	17	18	17	19	16	15
GO:0000139	Golgi membrane	16	13	11	16	17	12
GO:0004842	ubiquitin-protein transferase activity	16	22	23	24	24	9
GO:0005774	vacuolar membrane	16	20	24	21	30	20
GO:0005789	endoplasmic reticulum membrane	16	19	21	22	22	19
GO:0006351	transcription, DNA-templated	16	11	16	14	15	11
GO:0006732	coenzyme metabolic process	16	16	20	18	13	8
GO:0009651	response to salt stress	16	13	17	13	10	10
GO:0010629	negative regulation of gene expression	16	24	23	19	20	13
GO:0016887	ATPase activity	16	10	14	14	20	13
GO:0019899	enzyme binding	16	25	21	25	19	12
GO:0022607	cellular component assembly	16	32	25	24	17	10
GO:0042578	phosphoric ester hydrolase activity	16	15	14	17	19	11
GO:0042802	identical protein binding	16	20	18	27	18	12
GO:0043603	cellular amide metabolic process	16	12	13	12	13	11

4 CONCLUSIONS

In this thesis, we aimed to investigate differences among *Saccharum* genotypes phenotypically contrasting in their biomass content. In the first chapter we assessed gene expression profiles of twelve sugarcane genotypes grouped into high and low biomass groups. The gene expression data correctly represented the difference between the groups and revealed substantial variability among the high biomass accessions. The groups showed significant differences in the expression of genes involved in carbon partitioning, mostly sucrose synthesis and degradation. Within the groups we could identify the enrichment of defense and carbohydrate-related terms. In addition, we explored the expression and co-expression profiles of groups of genes that were members of pathways of interest. Finally, we also showed how expression profiles at the transcript level can bring new insights when assessing differences between the biomass groups.

We devoted the second chapter to investigate if genes showing allele imbalance could be related to distinct functional processes. As we aimed to investigate whether alleles were expressed accordingly to their estimated dosages, we proposed a model to account for prior knowledge of this information. We used a hierarchical Bayesian approach to go from a prior distribution of the allele proportion, based on genotyping information, to a posterior considering the relative expression of the allele. Our results reveal that allele-specific expression affects part of the investigated loci in *Saccharum* genotypes. However, we could not find clear functional patterns among genes showing allele-specific expression. Despite the innate limitations of the genotyping-by-sequencing approach, we successfully developed and applied a model to drive insights about allele-specific expression in the complex polyploid sugarcane.