Mol Breeding (2016) 36:9
DOI 10.1007/s11032-015-0430-2

Orphan Crops Browser: a bridge between model and orphan crops

Claire Lessa Alvim Kamei · Edouard I. Severing · Annemarie Dechesne · Heleen Furrer · Oene Dolstra · Luisa M. Trindade

Received: 3 September 2015 / Accepted: 23 December 2015 / Published online: 12 January 2016

Abstract Many important crops have received little attention by the scientific community, either because they are not considered economically important or due to their large and complex genomes. De novo transcriptome assembly, using next-generation sequencing data, is an attractive option for the study of these orphan crops. In spite of the large amount of sequencing data that can be generated, there is currently a lack of tools which can effectively help molecular breeders and biologists to mine this type of information. Our goal was to develop a tool that enables molecular breeders, without extensive bioinformatics knowledge, to efficiently study de novo transcriptome data from any orphan crop (http://www.bioinformatics.nl/denovobrowser/db/species/index). The Orphan Crops Browser has been designed to facilitate the following tasks (1) search and identification of candidate transcripts based on phylogenetic relationships between orthologous sequence data from a set of related species and (2) design specific and degenerate primers for expression studies in the orphan crop of interest. To demonstrate the usability and reliability of the browser, it was used to identify the putative orthologues of 17 known lignin biosynthetic genes from maize and sugarcane in the orphan crop *Miscanthus sinensis*. Expression studies in miscanthus stem internode tissue differing in maturation were subsequently carried out, to follow the expression of these genes during lignification. Our results showed a negative correlation between lignin content and gene expression. The present data are in agreement with recent findings in maize and other crops, and it is further discussed in this paper.

Claire Lessa Alvim Kamei and Edouard I. Severing have contributed equally to this article.

Electronic supplementary material The online version of this article (doi:10.1007/s11032-015-0430-2) contains supplementary material, which is available to authorized users.

C. L. A. Kamei · A. Dechesne · H. Furrer · O. Dolstra · L. M. Trindade (✉)
Wageningen UR Plant Breeding, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands
e-mail: luisa.trindade@wur.nl

E. I. Severing
Laboratory of Genetics, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands

Present Address:
C. L. A. Kamei
Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

E. I. Severing
Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

Keywords Orthologous genes · Bioinformatics tool · Breeding targets · De novo transcriptome · Orphan crops

Introduction

In developing countries, several indigenous plant species form the basis of subsistence to many local and regional communities, providing food, animal feed and other non-food products. These crops are well accepted and preferred by farmers and consumers and well adapted to the local conditions. However, their cultural and agricultural importance is undervalued. Given their lack of representation on the global markets, the research investment by public and private sectors is just a very small fraction of what is invested in major arable crops, such as maize, rice and wheat, more economically important to Europe and the USA (Jonkers 2010). In addition, the commonly large, complex and polyploidy genomes of these "orphan crops" also discourage further research. Given these challenges, one alternative to progress in orphan crop research would be to make use of knowledge available from model species, translating findings from models to orphans. To facilitate this task, breeding and biotechnological tools to study complex genomes and the transcriptomes of many orphan crops are currently being developed. Transcriptome datasets are valuable sources of information and useful to unravel plant pathways during different developmental stages and in response to biotic or abiotic stresses. The use of transcriptome datasets from closely related species will allow deepening our current knowledge on processes fundamental to the development of better and more competitive crops. For example, the orphan crop miscanthus is a new and promising C4 grass suitable for the production of second-generation biofuels. The crop comprises different polyploid species, including *Miscanthus sinensis* and *M. x giganteus*, and still lacks a reference assembled genome. Several researchers, however, have recently explored their transcriptional profile as an alternative to whole-genome sequencing (Barling et al. 2013; Chouvarine et al. 2012; Kim et al. 2014; Straub et al. 2013a; Swaminathan et al. 2012).

Commonly, transcriptome profiling consists of several steps: first, transcript fragments are sequenced using next-generation sequencing technologies such as Illumina and 454. Second, the transcriptome is reconstructed by performing *de novo* assemblies using dedicated software such as Trinity (Grabherr et al. 2011). Third, the de novo transcriptome is annotated using tools such as Blast2GO (Conesa et al. 2005). These steps are followed by the search of interesting genes and confirmation steps using wet-lab approaches. To this end, the knowledge gained from other organisms is often used to identify genes of interest in their *de novo* transcriptome. There are tools that provide user-friendly interfaces for gene discovery in de novo transcriptomes such as Trapid (Van Bel et al. 2013) and Trinotate-web (http://trinotate.github.io/TrinotateWeb.html). Although these tools are very powerful in helping to explore *de novo* transcriptome, they stop at the point where the user has a potential target gene and they do not provide the means of producing primers to confirm their presence.

Here we present the Orphan Crop Browser: a novel molecular breeding tool specifically designed to search for genes in species where little genome sequencing information is available using data available from other species (http://www.bioinformatics.nl/denovobrowser/db/species/index). To demonstrate the applicability of this newly developed tool, we focused on the search of miscanthus orthologues from maize and sugarcane genes known to operate in the lignin biosynthetic pathway.

Results

A browser to facilitate the study and to assist molecular breeding of orphan crops is presented. This tool enables to quickly identify target genes in their de novo transcriptome through a user-friendly graphical interface. In addition, the browser aids in the design and testing of primers that can be used for confirming the existence and quantifying the abundance of the target genes molecularly. This feature is included in the browser to avoid unnecessary copying and pasting between programs and Web sites. The result section is structured as follows: First, we describe the information that can be used as input for the browser; secondly, an overview of the various functionalities of the browser is given, and finally the usefulness and potentialities of this browser are illustrated using a biological example focusing on the orphan crop *Miscanthus sinensis*.

Browser setup

Before the browser can be used, a database has to be constructed that contains all information available from the novel-(orphan crop) and several annotated species. A graphical interface has specifically been developed to simplify the creation of the database. Different types of data can be uploaded into the browser including sequences, annotations, gene function information and homologous sequence clusters. A brief description of the input data is provided below.

Sequence data are the most basic type of information required before the browser can be used. The browser has been designed to handle full-length mRNA, coding and protein sequences. Users are required to provide both coding sequences and the corresponding protein translations for each species to the browser, since orthologous inferences in the browser are performed using predicted proteins and their corresponding coding sequences. Full-length mRNAs have no effect on the browser performance because they are not directly used for orthology inference. As a result, their inclusion in the browser is optional. The browser also includes the possibility to integrate exon/intron structure information. Information about the gene structure can help users to detect crop-specific alternative splicing events.

Different types of annotations can be uploaded into the browser. Although all sequences that are uploaded into the database receive an internal name, their original identifier is preserved as part of its annotation. Function descriptions and Gene Ontology term annotations, obtained using programs such as Blast2GO, enable users to search the database for specific key words. Users can upload tables containing the sequence IDs, descriptions and gene ontology (GO) (Ashburner et al. 2000) terms into the database. The most advanced type of annotation that can be uploaded into the database is protein domain annotations such as derived from PFAM (Punta et al. 2012). Protein domains correspond to conserved regions of proteins for which several have a known function, and they can be searched on both their ID or description.

Predefined clusters of homologous sequence can also be uploaded into the database, representing groups of orthologous proteins or gene families. Construction of these clusters can be done using programs such as OrthoMCL (Li et al. 2003). Although the inclusion of predefined sequence clusters enables users to quickly find related sequences, they are not strictly necessary for efficient use of the browser.

Functionalities

The functionalities of the browser are based on a specific work flow (Fig. 1a) which can be subdivided into five phases: (1) Initial search: users will search sequences within the database using the text or similarity search modules; (2) Creation of a set of related sequences: a set of related sequences is constructed for phylogenetic analysis; (3) Phylogenetic analysis: here users generate multiple sequence alignments and construct phylogenetic trees for identifying the target sequences; (4) Primer-design phase: primers are designed for wet-lab confirmation of sequences; (5) Primer testing phase: in this last phase, users test whether the primer pairs are adequate for finding the intended targets. Here, we provide a brief description of the different modules of the browser.

Text search

In the initial phase, users can use the text-search module to search for sequences in the database using the IDs, names, annotation key words, PFAM and GO-terms. In case the user knows beforehand that a particular sequence is a member of a homologous sequence cluster, the cluster can immediately be retrieved using its ID in the cluster-search module. In a successful search, a list of sequences is generated which can be selected for further actions (Fig. 2a).

Similarity search

As an alternative for the text search, it is possible to perform blast searches with the blast-search module. In this module, users can set the e-value threshold and the maximum number of hits to return. Furthermore, it is possible to specify a limiting blast search to a specific set of species in the database. The browser supports the following blast searches: nucleotide search (BlastN), protein search (BlastP) and nucleotide-to-protein search (BlastX).

Primer match

The primer-search module can be used both for the initial search and for testing whether primers are

specific for the intended target (see below). In the primer-match module, primer pairs can be provided that will be searched for a user-defined set of sequence datasets. The primer-match module first uses blast to identify candidate primer matches and then performs a Needleman Wunsch alignment (Needleman and Wunsch 1970) to refine the primer/target alignment. Users can influence the number of initial blast hits to consider and the number of mismatches in the refined alignment. Furthermore, it is possible to determine the maximum and minimum lengths of the products. The primer-match module will also search for possible products that can be amplified by the combination of forward and reverse primers or by each of them independently.

Fig. 2 Screenshots of the orphan crop browser. **a** Sequence search results table has dropdown menu through which the user is capable of performing several actions on (a subset of) the retrieved sequences. **b** The interactive alignment module enables users to edit alignments. **c** Tree view module enables user to select sequences and perform several actions. **d** The interface to the primer3 program enables users to design primers on selected sequences

Gathering related sequences

In order to perform phylogenetic analysis, users need to gather a set of related sequences if no initial blast search was performed. One way to identify these sequences is by using the pre-defined homologous sequence cluster information (when available).

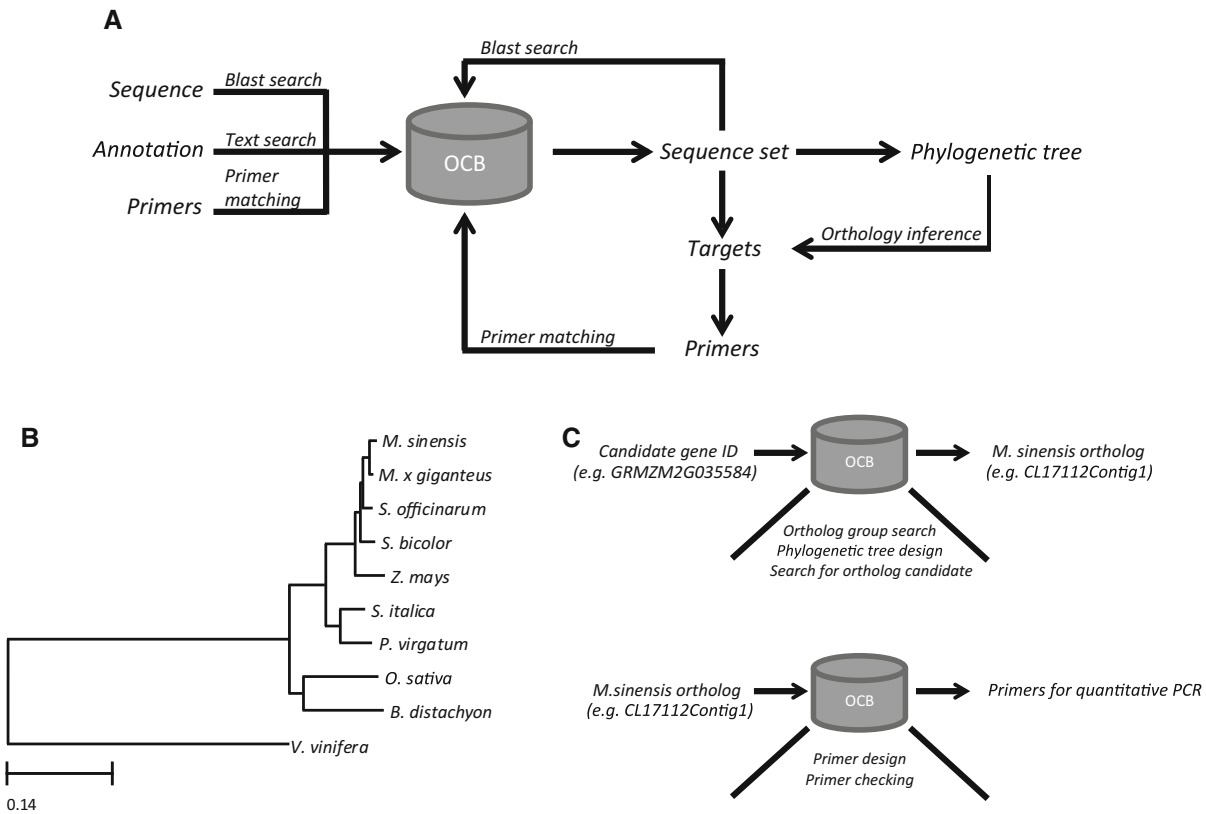
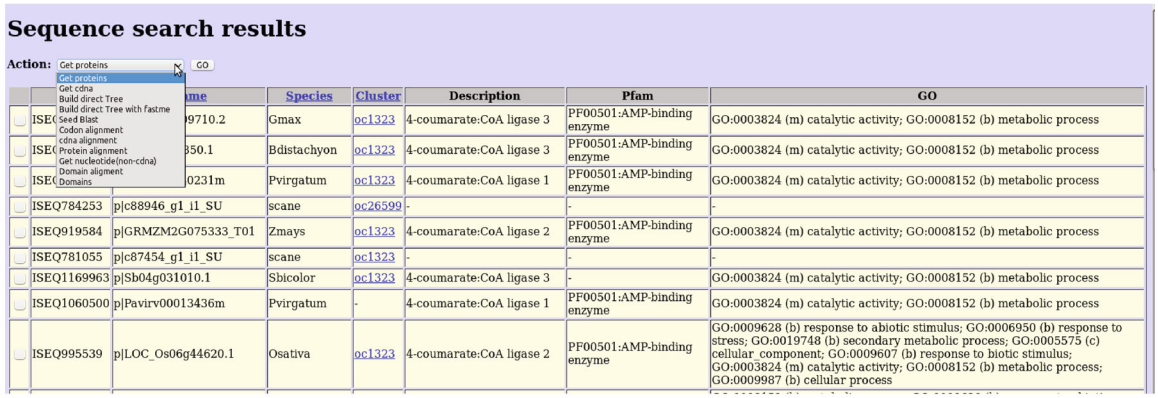


Fig. 1 Schematic work flow and use of the Orphan Crop Browser. **a** Main work flow of the Orphan Crops Browser. Using blast-, text- or primer-search sequences are identified in the browser. The initial search results can subsequently be used to create a set of related sequences, which can be used for constructing phylogenetic trees. Target genes can be identified by visually inferring orthologous relationships in the trees. After identification of targets, primers can be designed and tested for uniqueness by matching them against the database sequences. **b** Maximum likelihood tree of monocot species used in this study, including *Vitis vinifera* as an out-group. **c** Schematic representation of adopted strategy to identify miscanthus orthologous genes with the aid of the Orphan Crops Browser (OCB) and design of primers for qPCR

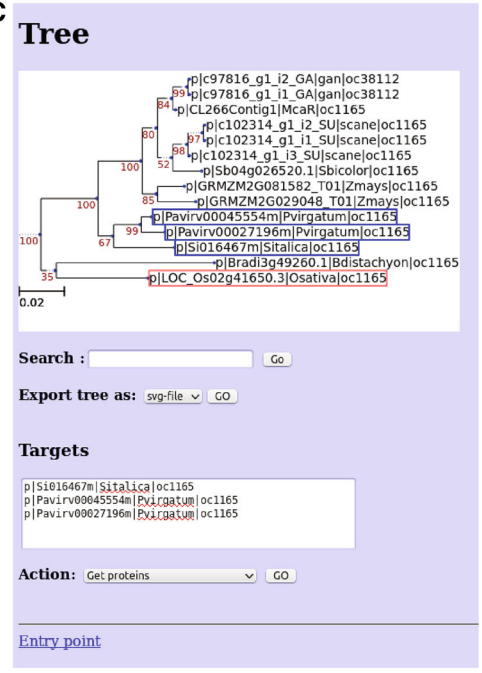
A


Sequence search results
Action: Get proteins GO
Get proteins
Get cdna
Build direct Tree
Build direct Tree with fastme
Seed Blast
Codon alignment
cdna alignment
Protein alignment
Get nucleotide(non-cdna)
Domain aligment
Domains
Species
Cluster
Description
Pfam
GO
ISEQ784253 p|c88946_g1_i1_SU scane oc26599 - - -
ISEQ919584 p|GRMZM2G075333_T01 Zmays oc1323 4-coumarate:CoA ligase 2 PF00501:AMP-binding enzyme GO:0003824 (m) catalytic activity; GO:0008152 (b) metabolic process
ISEQ781055 p|c87454_g1_i1_SU scane oc1323 - - -
ISEQ1169963 p|Sb04g031010.1 Sbicolor oc1323 4-coumarate:CoA ligase 3 - GO:0003824 (m) catalytic activity; GO:0008152 (b) metabolic process
ISEQ1060500 p|Pavirv00013436m Pvirgatum - 4-coumarate:CoA ligase 1 PF00501:AMP-binding enzyme GO:0003824 (m) catalytic activity; GO:0008152 (b) metabolic process
ISEQ995539 p|LOC_Os06g44620.1 Osativa oc1323 4-coumarate:CoA ligase 2 PF00501:AMP-binding enzyme GO:0009628 (b) response to abiotic stimulus; GO:0006950 (b) response to stress; GO:0019748 (b) secondary metabolic process; GO:0005575 (c) cellular_component; GO:0009607 (b) response to biotic stimulus; GO:0003824 (m) catalytic activity; GO:0008152 (b) metabolic process; GO:0009987 (b) cellular process

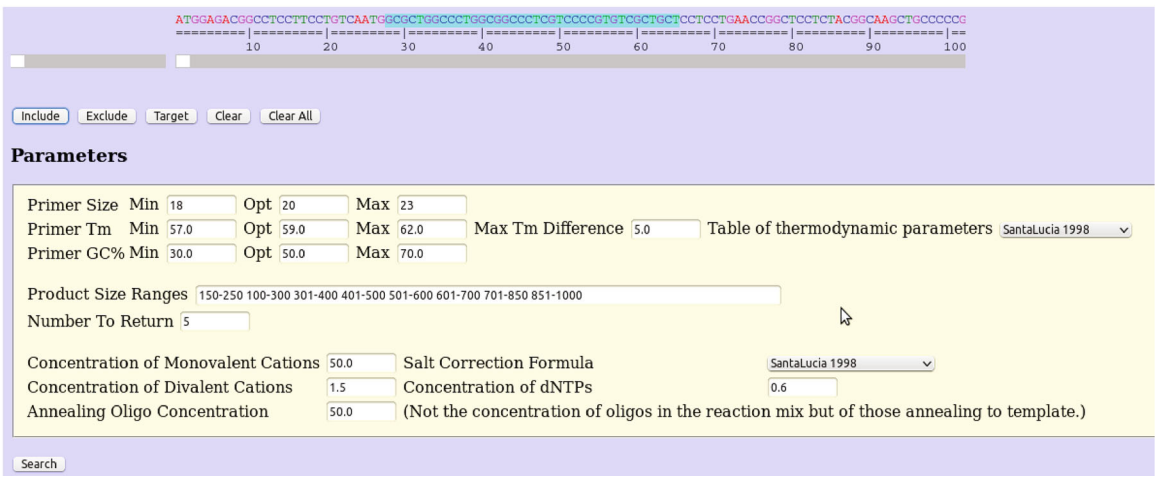
B


Alignment
p|Glyma13g44950.1|Gm
p|CL4569Contig3|McaR
p|CL4569Contig2|McaR
p|CL12078Contig1|Mca
p|Si016767m|Sitalica
p|Solyc06g068650.2.1
p|Potri.019G049500.1
p|Sb04g005210.1|Sbic
p|Bradi3g18960.1|Bdi
p|Glyma01g44270.1|Gm
p|Solyc03g097030.2.1
p|Pavirv00055097m|Pv
p|Glyma17g07190.2|Gm
p|Sb07g022040.1|Sbic
p|LOC_Os02g08100.1|O
p|GSVIVT01029182001|
p|GSVIVT01029183001|
p|Pavirv00040231m|Pv
p|c88946_g1_i1_SU|sc
230 240 250 260
Editing
Delete sequences
Delete blocks
Auto trim (TrimAl)
Clear structure
Trim all left
Trim all right
Crop to selected seqs
Export
Alignment
Consensus
Tree
NJ Clustal
FastME
Primers
Specific
Alignment
Consensus

C


Tree
p|c97816_g1_i2_GA|gan|oc38112
p|c97816_g1_i1_GA|gan|oc38112
p|CL266Contig1|McaR|oc1165
p|c102314_g1_i2_SU|scane|oc1165
p|c102314_g1_i1_SU|scane|oc1165
p|c102314_g1_i3_SU|scane|oc1165
p|Sb04g026520.1|Sbicolor|oc1165
p|GRMZM2G081582_T01|Zmays|oc1165
p|GRMZM2G029048_T01|Zmays|oc1165
p|Pavirv00045554m|Pvirgatum|oc1165
p|Pavirv00027196m|Pvirgatum|oc1165
p|Si016467m|Sitalica|oc1165
p|Bradi3g49260.1|Bdistachyon|oc1165
p|LOC_Os02g41650.3|Osativa|oc1165
0.02
Search : Go
Export tree as: svg-file GO
Targets
p|Si016467m|Sitalica|oc1165
p|Pavirv00045554m|Pvirgatum|oc1165
p|Pavirv00027196m|Pvirgatum|oc1165
Action: Get proteins GO
Entry point

D


10 20 30 40 50 60 70 80 90 100
Include
Exclude
Target
Clear
Clear All
Parameters
Primer Size Min 18 Opt 20 Max 23
Primer Tm Min 57.0 Opt 59.0 Max 62.0 Max Tm Difference 5.0 Table of thermodynamic parameters SantaLucia 1998
Primer GC% Min 30.0 Opt 50.0 Max 70.0
Product Size Ranges 150-250 100-300 301-400 401-500 501-600 601-700 701-850 851-1000
Number To Return 5
Concentration of Monovalent Cations 50.0 Salt Correction Formula SantaLucia 1998
Concentration of Divalent Cations 1.5 Concentration of dNTPs 0.6
Annealing Oligo Concentration 50.0 (Not the concentration of oligos in the reaction mix but of those annealing to template.)
Search

Alternatively, the user can perform a sequence similarity search (blast) against the database with a selected set of sequences using the seed-blast module. In the seed-blast module, users can restrict the search to a specific set of species and determine the maximum number of hits in total and per species. In addition, users can adjust the minimum identity, alignment coverage and e-value thresholds for hits. The availability of the seed-blast module provides a solution for gathering related sequences to users that do not have the resources for large-scale clustering of highly similar sequences.

Alignments

In the browser, users can construct protein, nucleotide and codon alignments. Codon alignments are generated by first creating protein alignments and then replacing the amino acids by the corresponding codons on the coding sequences. Alignments are displayed in a custom-built interactive alignment viewer (Fig. 2b) written in JavaScript. For nucleotide alignments, the viewer will display the consensus sequences at the bottom.

If available, the gene structure underlying the coding regions is superimposed onto the alignment. In protein alignments, the intron positions are indicated by placing the phase of the intron (Long et al. 1995) before the amino acid of which the corresponding codon is preceded or interrupted by the intron. The phase of intron indicates whether it resides between two consecutive codons, between the first and second or between the second or third nucleotide of a codon. In nucleotide alignments, intron positions are indicated by "^" characters. Previous studies have shown intron positions within the coding regions of genes are highly conserved, even between distantly related species as Arabidopsis and rice. Therefore, users can make use of the superimposed gene structure information for detecting potential alternative splicing events in an orphan crop.

The accuracy of phylogenetic tree reconstruction can be improved by removing regions of low conservation from the multiple sequence alignment (Talavera and Castresana 2007). Therefore, the alignment viewer has the possibility to remove low conserved regions from the alignment. Low conserved regions can be removed either manually or automatically by invoking the TrimAl program (Capella-Gutierrez et al. 2009).

Phylogenetic tree module

Phylogenetic trees can be generated either by invoking ClustalW2 (Larkin et al. 2007) with a single click or by first configuring and then launching FastME (Desper and Gascuel 2002). The generated trees are rendered on the background by ETE2 (Huerta-Cepas et al. 2010) and displayed in a custom viewer that allows users to select individual or groups of sequences by clicking on the respective terminal or internal nodes (Fig. 2c). In this way, users can easily select potential targets through manual inspection of the phylogenetic tree and perform further analysis. The phylogenetic tree can be exported in PNG and SVG image formats. Additionally, users can export the tree in a Newick format, which can be opened by tree-rendering programs.

Primer design

In order to aid researchers in confirming or quantifying transcripts, the browser provides the possibility of designing primers. There are modules for three different primer-design modes in the browsers: (1) the specific primer mode can be used to design primers for individual sequences (Fig. 2d). Intuitively, the specific mode should be used once a target gene has been identified. The specific primer-design module is simply a custom-build interface to the primer3 (Untergasser et al. 2012) program. Users can use the mouse for selecting which regions must be included in or excluded from the primer search. (2) The consensus mode is an option of the same interface as the specific mode but should be used to design primes based on the consensus sequence of an alignment. The purpose of this mode is to construct primers for perfectly conserved regions on the alignment. (3) The alignment mode can be used for designing degenerate primers. The degenerate primers can be used in the case no target sequences were identified after inspecting the phylogenetic tree. The gene might still be present in the species of interest but overlooked due to insufficient sequencing or incorrect assembly. By using the alignment of several species, users can design degenerate primers for an in-depth search for the presence the target gene in the orphan crop of interest.

Predicted primer pairs are presented in a sortable table, and the amplified gene fragments can be visualized on the sequences/alignment. This allows users to make a selection and to export it for further use. The primer-match module also enables users to

check the specificity of the selected primer pairs for the desired target sequence, by checking whether they may recognize other regions and consequently result in (less specific) amplification of fragments different from the target sequence. This is an important feature to guarantee the best selection of primer pairs for amplification of your target.

Orphan Crops Browser in action

To evaluate the applicability and accuracy of our developed browser in the search of candidate orthologues in orphan crops, a case study is presented. Given the importance of miscanthus for second-generation biofuels research and the limited amount of sequence information currently available, we chose this crop as the orphan crop of interest. The advance of second-generation biofuels mainly depends on overcoming the natural recalcitrance of the cell wall matrix to deconstruction, to directly improve the processability of lignocellulose feedstocks. Taking this feature into account, we decided to focus on finding genes involved in this process. Identifying the key genes involved in lignin biosynthesis, and correlating their expression to the content and composition on the secondary cell walls is an essential step to improve saccharification efficiency. For example, molecular engineering of key genes in lignin biosynthesis has already shown to improve the release of glucose from cell walls of switchgrass and sugarcane, without affecting plant growth, fertility or biomass yield (Jung et al. 2012; Saathoff et al. 2011). Lignin is a structural component of secondary cell walls, providing firmness and safeguard an operational vascular system, while hemicelluloses are polysaccharides able to covalently link with cellulose, lignin and some pectins, granting strength to the cell walls (Samuel et al. 2011). The accumulation of all cell wall constituents throughout plant growth and development hinders the accessibility of cellulose to hydrolyzing enzymes, thus compromising the release of sugars for cellulosic ethanol production. Finally, we choose to use sequence information from the plant species most closely related to miscanthus for the search of orthologous genes.

Selection of *M. sinensis* genotypes

In order to identify key genes in the lignin biosynthesis in *Miscanthus sinensis*, genotypes contrasting to secondary cell wall composition were selected from a collection of approximately 120 *M. sinensis* accessions, located at Wageningen University and Research Centre (WUR). Ten stems were harvested at mature stage, from two individual plants per plot and pulled together for quantification of lignin, hemicellulose and cellulose contents (data not shown). Out of this set, four genotypes differing in cell wall properties (Table 1) were selected for a gene expression study using stem materials collected at two stages of development. Shoots were taken during the vegetative stage (<1 m height) and the generative stage (early flowering). Three internodes (i.e., internodes 2, 3 and 4 above the soil) were collected per shoot and divided into three sections, which were used for further analyses. They are referred to as up, middle or low section. The two extreme sections of each internode (up and low sections) were used to measure lignin content. Three stem samples widely differing in maturation were finally chosen for gene expression analyses for each of the four genotypes (Supplemental Table S1; Fig. 3a). The lowest lignin content was found as expected in young plants, at the meristematic region of the youngest internode (YIL4), while the highest content was observed in mature stems, in the top section of the first formed internodes, a region enriched of cells containing a mature cell wall (MIU2). In order to evaluate gene expression during stem lignification, we also included the most lignified segment from young plants (YIU2).

Selection of candidate plant species and lignin genes

Based on our orthology predictions, the best candidate species to base the prediction of miscanthus orthologous genes were selected. Using *Vitis vinifera* as outgroup species, a total of 180 OrthoMCL clusters were identified, in which all monocot species and *V.*

Table 1 Stem biochemical composition of four selected *Miscanthus sinensis* genotypes

Genotype	%LIG	%HEM	%CEL
H0116	13.8	28.5	40.5
H0117	13.8	30.1	42.2
H0119	12.8	26.7	46.6
H0120	15.2	29.0	39.8

%LIG, lignin; %HEM, hemicellulose; %CEL, cellulose

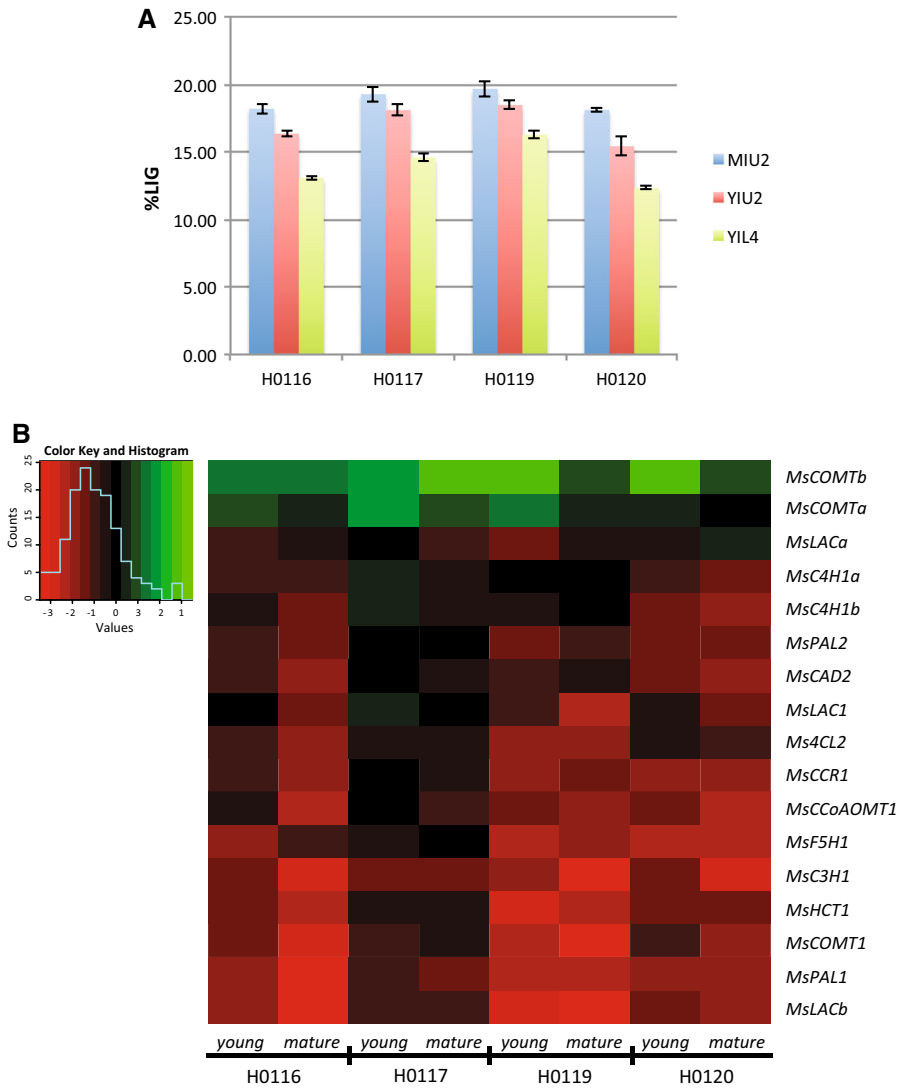


Fig. 3 Lignin content and gene expression of studied internode sections of miscanthus. **a** Lignin cell wall content (%LIG) of selected internode segments from four *Miscanthus sinensis* genotypes (H0116, H0117, H0119 and H0120). MIU, up section of mature internode; YIU, up section of young internode; YIL, low section of young internode. **b** Heat map of cross-comparison of miscanthus lignin genes expressed in tested internode sections based on hierarchical clustering. *Each pair of columns* represents one of the studied genotypes, and each row represents a lignin gene. Down-regulated genes are indicated in *red*, while up-regulated genes in *green*. The young and mature denominations in the map correspond to the gene expression measured on YIU2 and MIU2 sections relative to the expression on YIL4, respectively. (Color figure online)

vinifera were represented only one time. The concatenated codon alignment of the sequences in these clusters consisted out of 146,199 sites. Using the jModelTest (Posada and Crandall 1998), we determined (based on the BIC criterion) that the most appropriate substitution model to calculate a maximum likelihood tree with the alignment was GTR+G with four rate categories. The generated maximum likelihood tree, displaying 100 % bootstrap support at all nodes, confirmed sugarcane (*Saccharum officinarum*) as the closest evolutionary-related species to miscanthus (Fig. 1b). Sugarcane and miscanthus belong to the *Saccharinae* group, and their close phylogenetically relationship has even instigated introgression of miscanthus germplasm into sugarcane to confer better growth under sub-optimal temperature (Lam et al. 2009). The high similarity between the genomes of both species has also been shown through the assembly of the first genetic map of miscanthus, when genetic markers from sugarcane were efficiently used to target specific regions in the *M. sinensis* genome (Swaminathan et al. 2012). Considering the recent works from Mazzafera's group (Bottcher et al. 2013; Cesarino et al. 2013), we searched for miscanthus orthologues using key genes known to operate during the lignification of sugarcane stems (*SofLAC*,

Sh4CL2, *ShC3H1*, *ShC4H1*, *ShCAD2*, *ShCCoAOMT1*, *ShCOMT1*, *ShF5H1* and *ShPAL1*). Sorghum shares a high level of collinearity with miscanthus (Kim et al. 2012; Ma et al. 2012; Swaminathan et al. 2012) and sugarcane (Wang et al. 2010) genomes, which makes the sorghum genome an ideal template for comparative genomic studies with these species. However, since currently most knowledge on lignification is described for maize, we chose to perform a translational research analysis from this model crop to miscanthus (Lawrence and Walbot 2007). Analysis of a maize dataset (Sekhon et al. 2013) resulted in the identification of differentially expressed genes possibly involved in lignification in the two analyzed internode fractions (V5_first internode and V9_fourth internode). Considering that the higher expression of these genes was mainly observed in the younger internode, we screened for genes involved in the lignin biosynthesis pathway showing higher expression in this sample (V5_first internode). In this way, eight candidates from maize (*ZmCCR1*, GRMZM2G141026, GRMZM2G140996, GRMZM2G035584, GRMZM5G842071, GRMZM2G447271, and both GRMZM2G081582 and GRMZM2G029048) were selected, two of them that showed homology to sugarcane genes (*ShHCT1* and *ShPAL2*). These two sugarcane genes were not initially added in our analysis, since they were considered not to have a role on stem lignification in sugarcane; however, we still decided to include them in our study since they are up-regulated in young maize internodes. A list with all used candidate genes in our analysis is summarized in Table 2.

Identification and evaluation of *Miscanthus sinensis* orthologous genes

After selecting the lignin candidate genes from sugarcane and maize, a two-step approach was adopted for the identification of miscanthus orthologues (Fig. 1c). First, a S*equence Search* was performed for each candidate gene ID separately, resulting in the identification of its correspondent orthologous group. All sequences from each group were used to construct a phylogenetic tree, which allowed the identification of single or multiple miscanthus orthologous candidate genes. Overall, multiple miscanthus orthologous sequences were generally found for each candidate gene, with a few exceptions. Orthologous sequences from *M. x giganteus* were found for seven candidate genes and were kept in the analysis to support the alignments (data not shown). All generated phylogenetic trees are shown in Supplemental Figures S1A-P. As second step, the selected single or consensus miscanthus sequences were used as templates for the design of specific primers and further checked in silico for efficiency. Preceding the use of primers for quantitative gene expression analyses, the specificity of each pair of primers was also confirmed by sequencing of the amplified gene fragment (for primers list, see Supplemental Table S2).

To evaluate how lignin accumulation would be correlated with gene expression, the first step was the analysis of how gene expression variation associates with distinguishable developmental stages in early plant growth. A cross-comparison was performed within the young plant, focusing on the most mature section of the oldest internode (YIU2) and the meristematic section of the youngest internode (YIU2 × YIL4). Lignification was more prominent in YIU2 (Fig. 3a), whereas the lignin genes showed higher expression levels in YIL4 for all four genotypes (Fig. 3b). Although the down-regulation of lignin genes was less obvious in the genotype H0117, for the other three genotypes only two to three genes showed up-regulation. In all genotypes, the *MsCOMTa* and *MsCOMTb* genes were the only genes with increased expression following tissue lignification. Since lignification is a process linked to tissue maturity, we decided to confirm these results in a more contrasting scenario, by carrying out a second comparison on plants in different developmental stages (MIU2 × YIL4). As expected, the down-regulation of lignin genes was even more pronounced in this stage enriched of cells containing mature cell walls. In addition, the expression of the *MsLACa* gene slightly increased in genotype H0120 (Supplemental Table S3).

Discussion

Here we presented the Orphan Crops Browser, a novel user-friendly tool for the identification of orthologous genes in orphan crops and to assist molecular breeding of these species. We have illustrated its capabilities by identifying and studying key genes involved in secondary cell wall lignification in *M. sinensis*, an important bioenergy feedstock candidate, using knowledge from closely related C4 plants. Some

Table 2 Lignin candidate genes from sugarcane, maize and their putative *Miscanthus sinensis* orthologues used in this work

Family; description	Maize gene ID	Sugarcane gene ID	Sugarcane orthologues in OCB[a]	*Miscanthus* gene ID	*Miscanthus* orthologues in OCB	References
4CL; 4-coumarate:CoA ligase 2	GRMZM2G055320	*Sh4CL2*	c88946_g1_i1_SU c88946_g1_i2_SU	*Ms4CL2*	CL12078Contig1	Bottcher et al. (2013) Shangguan et al. (2013) Zhang et al. (2014)
C3H; coumarate-3-hydroxylase	GRMZM2G140817	*ShC3H1*	c70890_g1_i1_SU	*MsC3H1*	CL5399Contig1 CL5399Contig2	Bottcher et al. (2013)
C4H; cinnamate-4-hydroxylase	GRMZM2G139874 1st cluster	*ShC4H1*	c94596_g1_i1_SU	*MsC4H1a*	CL585Contig3 m454_isotig05388 m454_isotig05386 MU_comp33105_c1_seq10	Bottcher et al. (2013) Courtial et al. (2013) Shangguan et al. (2013)
	2nd cluster	NA	NA	*MsC4H1b*	CL585Contig2 MF_comp34800_c0_seq1 MF_comp34800_co_seq3	
CAD; cinnamyl alcohol dehydrogenase	GRMZM5G844562	*ShCAD2*	c91614_g1_i2_SU c91614_g1_i1_SU	*MsCAD2*	CL3287Contig1 CL3287Contig2 CL3287Contig3 CL3287Contig4	Courtial et al. (2013) Tanaka et al. (2014) Zhang et al. (2014)
CCoAOMT; S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	GRMZM2G099363 GRMZM2G127948	*ShCCoAOMT1*	c107456_g2_i1_SU	*MsCCoAOMT1*	CL4519Contig1 CL24843Contig1	Bosch et al. (2011) Bottcher et al. (2013) Courtial et al. (2013) Li et al. (2013) Wen et al. (2014) Zhang et al. (2014)

Table 2 continued

Family; description	Maize gene ID	Sugarcane gene ID	Sugarcane orthologues in OCB[a]	*Miscanthus* gene ID	*Miscanthus* orthologues in OCB	References
CCR; cinnamoyl coa reductase	GRMZM2G131205 (*ZmCCR1*)	NA	NA	*MsCCR1*	CL5621Contig1 CL5621Contig2 CL5621Contig3 CL23415Contig1	Khan et al. (2010) Bosch et al. (2011) Tamasloukht et al. (2011) Courtial et al. (2013) Liseron-Monfils et al. (2013) Tanaka et al. (2014) Zhang et al. (2014)
COMT; O-methyltransferase family protein	AC196475.3_FGT004	*ShCOMT1*	c96304_g1_i2_SU c96304_g1_i1_SU	*MsCOMT1*	CL3733Contig1 CL3733Contig2 m454_isotig10910 m454_isotig10911	Bottcher et al. (2013)
	GRMZM2G141026	NA	NA	*MsCOMTa*	MU_comp25105_c0_seq2 MF_comp27777_c0_seq2 MF_comp27777_c0_seq4	Bosch et al. (2011) Courtial et al. (2013) Meihls et al. (2013)
	GRMZM2G140996	NA	NA	*MsCOMTb*	MU_comp504097_c0_seq1	Bosch et al. (2011) Courtial et al. (2013) Meihls et al. (2013)
F5H; ferulic acid 5-hydroxylase	AC210173.4_FGT005	*ShF5H1*	c101416_g2_i1_SU	*MsF5H1*	CL8875Contig1	Bottcher et al. (2013)

Table 2 continued

Family; description	Maize gene ID	Sugarcane gene ID	Sugarcane orthologues in OCB[a]	*Miscanthus* gene ID	*Miscanthus* orthologues in OCB	References
HCT; hydroxycinnamoyl-CoA shikimate/ quinate hydroxycinnamoyl transferase	GRMZM2G035584	*ShHCT1*	c93886_g2_i1_SU	*MsHCT1*	CL17112Contig1	Courtial et al. (2013)
LAC; Laccase	GRMZM2G305526	*SofLAC*	NA	*MsLAC1*	CL12102Contig1	Cesarino et al. (2013)
	GRMZM5G842071	NA	NA	*MsLACa*	CL15280Contig1	Courtial et al. (2013)
	GRMZM2G447271	NA	NA	*MsLACb*	CL3689Contig1	Courtial et al. (2013)
PAL; PHE ammonia lyase	GRMZM2G074604	*ShPAL1*	c100670_g1_i1_SU	*MsPAL1*	CL266Contig4 MF_comp35858_c1_seq3 m454_isotig15917	Bosch et al. (2011) Bottcher et al. (2013) Courtial et al. (2013) Shangguan et al. (2013) Tanaka et al. (2014) Zhang et al. (2014)
	GRMZM2G081582 GRMZM2G029048	*ShPAL2*	c102314_g1_i3_SU	*MsPAL2*	CL266Contig3	Bosch et al. (2011) Bottcher et al. (2013) Courtial et al. (2013) Liseron-Monfils et al. (2013) Shangguan et al. (2013)

The miscanthus gene ID was given according to the nomenclature adopted in sugarcane whenever possible

[a] Sugarcane orthologous are only provided if they were used as the starting point for tree construction

online tools exist that can perform the tasks of some of the individual modules in the browser. However, the browser enables the users to perform all those tasks without constantly needing to transfer data between Web pages. In contrast to Trinotate-web, the Orphan Crops Browser provides a graphical user interface that simplifies the database construction, which is needed for exploring the data. As the Orphan Crops Browser was built for the purpose of target gene identification, it is sufficient to import a de novo transcript together with a set of species with annotated genomes. Therefore, the Orphan Crops Browser does not perform large-scale annotations of de novo transcriptomes in the same manner as Trapid.

In contrast to the Orphan Crops Browser, neither Trapid nor Trinotate-web has the possibility to design and test specific and degenerate primers. These tools do not provide a solution for the situation in which a target sequence cannot be found due to, for instance, insufficient sequencing depth. Our tool overcomes this problem, by providing users with the possibility of designing degenerated primers for confirming the presence or absence of a particular gene in the orphan crop. This is particularly useful in those cases where closely related species contain certain genes that were not assembled in the orphan crop due to, for instance, low expression levels. The user friendliness of the Orphan Crop Browser will especially be useful for researches with limited bioinformatics knowledge.

In the miscanthus case study, we were able to identify 17 orthologous genes in *M. sinensis* having a putative role in lignification with the help of the Orphan Crops Browser. In general, our results showed a down-regulation of lignin genes in stem tissues in parallel with lignin accumulation, suggesting that gene expression decreases as cell wall matures. Recent works support our findings, such as the research of Sekhon et al. (2013), which showed through a microarray analysis that in maize tissues the highest expression for the majority of lignin genes is found in the most immature organs. This indicates the most active developmental processes on secondary cell wall formation and lignification starts when cells are young. Results of a detailed transcriptome analysis made by Zhang and collaborators (Zhang et al. 2014) using elongating maize internodes showed that the peak of lignin genes expression is found in the internode section showing cell division and active elongation, with expression being no longer detected when elongation stops and lignin continues to accumulate. In sugarcane, with the exception of *SofLAC*, the determinant criterion used to select genes to have a key role in the lignin pathway was the association of gene expression in stem tissues with lignin deposition (Bottcher et al. 2013; Cesarino et al. 2013). Analysis of two maize internodes with contrasting levels of lignification showed that expression of *MsCCR1*, *MsCOMT1*, *MsCOMTa*, *MsCCoAOMT1*, *MsPAL1* and *MsPAL2* was higher in the most lignified internode, except for the *MsCOMTb* gene (Bosch et al. 2011). Most probably, meristematic and elongating internode sections produce more lignin monomer components, more important in the first steps of lignification. At later stages, lignin genes would reduce their expression before a cell reached maturity, taking advantage of the stability of enzymes and storage monolignols to continue the stem lignification. Increased lignification during post-harvest has been already observed in asparagus spears (Hennion et al. 1992), which validates the idea of lignification occurring in the absence of an active gene expression, possibly due to high stability of lignin biosynthetic enzymes. The formation of the lignin polymer requires the synthesis of different monolignols, mainly p-coumaryl, coniferyl and sinapyl alcohols, which form the basis of the phenylpropanoids p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units, respectively. Understanding their biosynthesis is essential to be able to improve plant biomass composition (Boerjan et al. 2003; Li and Chapple 2010). In the case of *M. sinensis*, only *MsCOMT* genes were up-regulated in mature tissues, and that may indicate that the need for syringyl (S) units is higher during the last stages of stem lignification. Cell walls rich in S-units are known to be less recalcitrant to saccharification, in comparison with G-rich cell walls (Huntley et al. 2003). This finding makes *MsCOMT* genes interesting genetic targets to improve intrinsic properties of cell walls and to develop better feedstocks for second-generation biofuels. Altogether, the browser revealed to be versatile and efficient in the identification lignin genes in miscanthus and primer design for expression studies. We have shown that the OCB browser is an effective tool to identify specific breeding targets in orphan crops and assist molecular breeding.

Materials and methods

Web browser

The Web browser was constructed using the Django framework for Python, in combination with HTML and JavaScript. All biological data are stored in MySQL databases. Several features of the browser use external programs (see below).

Primer design

Specific primer pairs are predicted using the Primer3 (Untergasser et al. 2012) command line program. Degenerate primer pair predictions start by searching the nucleotide alignment of all potential primers with predefined lengths and maximum degeneracy lying with specific distance. For each degenerate primer pair, the following procedure is respected: First, all possible primer sequences are determined from the degenerate sequences. Next, the melting temperature is calculated for each primer sequence using the oligotm function (oligotm.h from the Primer3 source code). All primers with melting temperatures outside the user-defined range are removed, including the ones predicted to form homo-dimers according to the dpal function (dpal.h) from Primer3 or that have high or low GC content. From the remaining set of primers, pairs that potentially form hetero-dimers according to the dpal function are discarded. Finally, primer pairs that can amplify all sequences are reported to the user.

Primer matching

All possible forward and reverse primer sequences are determined for primers containing degenerate nucleotides. Initial matches to the database sequence are obtained through BlastN (Altschul et al. 1997) searches with user-defined parameters. Primers with significant blast hits are realigned to the corresponding target sequences using the Needleman Wunsch alignment (Needleman and Wunsch 1970). Primers pairs for which each of the NW alignments meets user-defined requirements and is separated by a predefined distance are reported back to the user.

Alignments

All alignments are made using the MUSCLE version 3.5 (Edgar 2004). Codon alignments are generated from the protein alignments by substituting each amino acid with its corresponding codon and extending each gap to three.

Phylogenetic trees within the browser

Neighbor joining trees are constructed using ClusterW2 (Larkin et al. 2007) or FastME (Desper and Gascuel 2002). The support for each node in the tree is determined through bootstrap analysis. The trees are subsequently rendered using the ETE package (Huerta-Cepas et al. 2010) version 2.

Case study using *Miscanthus sinensis*

The following short read sequences were downloaded from the NCBI SRA repository: *Saccharum officinarum* (SRR89062); *M. sinensis* varieties Gross Fontaine (SRX131848) and Udine (SRX131681); *M. sinensis* 454 sequences (SRR916887); *M. x giganteus* (SRP017791).

De novo transcriptome assembly

All Illumina reads were trimmed using prinseq-lite 0.20.4 (http://prinseq.sourceforge.net/). The cleaned reads were assembled using Trinity with the minimum Kmer coverage parameter set to 2. Separate assemblies were performed for the Gross Fontaine and Undine *Miscanthus* varieties. The *M. sinensis* 454 reads were assembled into contigs using Newbler version 2.6 (http://www.454.com/products/analysis-software/). An overall *M. sinensis* transcriptome was generated by merging the Gross Fontaine, Udine and 454 assemblies using the TGICL package (Pertea et al. 2003).

Annotation of assembled contigs

All contigs were searched against the non-redundant protein database from NCBI using BlastX (e-value 1e-5). All sequences for which the best hit was a non-plant protein were discarded. The remaining sequences were annotated by importing the blast

results into Blast2GO (Conesa et al. 2005). Open reading frames were predicted using ESTscan that specifically trained for each species as follows: First, the assembled contigs were realigned to their best blast hit using exonerate (Slater and Birney 2005). All alignments were required to be devoid of frame shifts and to contain at least 80 % of the residues of the known protein. Those contigs for which the exonerate alignments met the criteria and that contained open reading frames (from start to stop) including UTR sections were used as training sequences for ESTscan (Iseli et al. 1999).

Plants with sequenced genomes

The protein sequences, coding sequences, gene structure and annotation data were downloaded from Phytozome for the following species: *Arabidopsis thaliana*, *Glycine max*, *Vitis vinifera*, *Sorghum bicolor*, *Panicum virgatum*, *Zea Mays*, *Oryza sativa*, *Brachipodium distachyon* and *Setaria italica.*

Function protein domains

Functional protein domains (PFAM) were predicted using pfamscan.pl (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/).

Orthologous groups

An all-versus-all protein similarity search was performed using BlastP (1e-5). Orthologous groups were predicted from the blast output using OrthoMCL version 2.9 (Li et al. 2003).

Phylogenetic tree

OrtholoMCL clusters were filtered and only the clusters in which all monocot species in our database and *Vitis venifera* were represented by only one sequence. The protein sequences within each individual cluster were aligned using MUSCLE. The resulting protein alignments were converted to codon alignments and subsequently concatenated into a single large alignment using in-house scripts. jModelTest version 2.1.7 (Posada and Crandall 1998) was used for determining the appropriate substitution model for construction maximum likelihood trees. A maximum likelihood tree was constructed from the concatenated alignment using PhyML (Guindon et al. 2010). The resulting tree was rooted and visualized using ETE2 (Huerta-Cepas et al. 2010).

Plant material

The four *Miscanthus sinensis* genotypes, known as H0116, H0117, H0119 and H0120, were collected for both biochemical and gene expression analyses, from two independent plants in the field. In average, 20 shoots were harvested, chopped and air-dried at 70 °C for 48 h, and subsequently ground to a fine powder using a ball mill (Retsch, Haan, Germany) for biochemical compositional analyses. In parallel, 10 shoots were harvested at the same time and transferred to liquid nitrogen and milled (Mill Pulverisette 14, FRITSCH) for RNA extraction.

Compositional analyses

Neutral detergent fiber (NDF) components were determined as described by (Torres et al. 2013), and lignin content was quantified by the acetyl bromide method using the cell wall residue obtained from the NDF treatment as described in (De Souza et al. 2015).

RNA manipulation

Total RNA was extracted from miscanthus internode sections using the TRIZOL reagent (Invitrogen) according to manufacturer's instructions. RNA concentration and purity were quantified with Nanodrop measurements, and the quality of the total RNA was checked on a 1.5 % RNase-free agarose gel. After DNase treatment with the DNase I Amplification Grade kit (Invitrogen) and purification using the RNeasy Mini Kit (Qiagen), cDNA was synthesized with the iScript cDNA synthesis kit (Bio-Rad). To investigate whether the selected primers were amplifying the predicted regions defined by the Orphan Crops Browser correctly, all amplified fragments were cloned and subsequently sequenced. A pool from all cDNA samples was used as a template, and fragments were amplified using Pfu DNA Polymerase (Promega) and fused to the pENTR^{TM}/D-TOPO® Gateway vector (Invitrogen). Quantitative RT-PCR was performed using a Bio-RAD detection system and the SYBR Green kit (Roche), with 3 μM specific primers, 10 ng of cDNA in a total volume of 10 μL per reaction. Two independent runs were performed as

technical replicates, and samples were analyzed in triplicate per run, with the use of the actin reference gene for normalization (Straub et al. 2013b). Gene expression fold change (2^n) was calculated using the $-\Delta\Delta$CT method (Livak and Schmittgen 2001).

Acknowledgments The authors gratefully acknowledge the funding from the European Union consortia SUNLIBB (Project ID 251132).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29. doi:10.1038/75556

Barling A, Swaminathan K, Mitros T, James BT, Morris J, Ngamboma O, Hall MC, Kirkpatrick J, Alabady M, Spence AK, Hudson ME, Rokhsar DS, Moose SP (2013) A detailed gene expression study of the *Miscanthus* genus reveals changes in the transcriptome associated with the rejuvenation of spring rhizomes. BMC Genom 14:864. doi:10.1186/1471-2164-14-864

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annu Rev Plant Biol 54:519–546. doi:10.1146/annurev.arplant.54.031902.134938

Bosch M, Mayer CD, Cookson A, Donnison IS (2011) Identification of genes involved in cell wall biogenesis in grasses by differential gene expression profiling of elongating and non-elongating maize internodes. J Exp Bot 62(10):3545–3561. doi:10.1093/jxb/err045

Bottcher A, Cesarino I, Santos AB, Vicentini R, Mayer JL, Vanholme R, Morreel K, Goeminne G, Moura JC, Nobile PM, Carmello-Guerreiro SM, Anjos IA, Creste S, Boerjan W, Landell MG, Mazzafera P (2013) Lignification in sugarcane: biochemical characterization, gene discovery, and expression analysis in two genotypes contrasting for lignin content. Plant Physiol 163(4):1539–1557. doi:10.1104/pp.113.225250

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973. doi:10.1093/bioinformatics/btp348

Cesarino I, Araujo P, Sampaio Mayer JL, Vicentini R, Berthet S, Demedts B, Vanholme B, Boerjan W, Mazzafera P (2013) Expression of SofLAC, a new laccase in sugarcane, restores lignin content but not S: G ratio of Arabidopsis lac17 mutant. J Exp Bot 64(6):1769–1781. doi:10.1093/jxb/ert045

Chouvarine P, Cooksey AM, McCarthy FM, Ray DA, Baldwin BS, Burgess SC, Peterson DG (2012) Transcriptome-based differentiation of closely-related *Miscanthus* lines. PLoS One 7(1):e29850. doi:10.1371/journal.pone.0029850

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676. doi:10.1093/bioinformatics/bti610

Courtial A, Soler M, Chateigner-Boutin A-L, Reymond M, Méchin V, Wang H, Grima-Pettenati J, Barrière Y (2013) Breeding grasses for capacity to biofuel production or silage feeding value: an updated list of genes involved in maize secondary cell wall biosynthesis and assembly. Maydica. http://www.maydica.org/articles/58_067.pdf

De Souza AP, Kamei CL, Torres AF, Pattathil S, Hahn MG, Trindade LM, Buckeridge MS (2015) How cell wall complexity influences saccharification efficiency in *Miscanthus sinensis*. J Exp Bot 66(14):4351–4365. doi:10.1093/jxb/erv183

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9(5):687–705. doi:10.1089/106652702761034136

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797. doi:10.1093/nar/gkh340

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652. doi:10.1038/nbt.1883

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59(3):307–321. doi:10.1093/sysbio/syq010

Hennion S, Little CHA, Hartmann C (1992) Activities of enzymes Involved in lignification during the postharvest storage of etiolated asparagus spears. Physiol Plant 86(3):474–478. doi:10.1034/j.1399-3054.1992.860319.x

Huerta-Cepas J, Dopazo J, Gabaldon T (2010) ETE: a python environment for tree exploration. BMC Bioinform 11:24. doi:10.1186/1471-2105-11-24

Huntley SK, Ellis D, Gilbert M, Chapple C, Mansfield SD (2003) Significant increases in pulping efficiency in C4H-F5H-transformed poplars: improved chemical savings and reduced environmental toxins. J Agric Food Chem 51(21):6178–6183. doi:10.1021/jf034320o

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding

regions in EST sequences. In: Proceedings/international conference on intelligent systems for molecular biology; ISMB international conference on intelligent systems for molecular biology, pp 138–148

Jonkers K (2010) Models and orphans; concentration of the plant molecular life science research agenda. Scientometrics 83(1):167–179. doi:10.1007/s11192-009-0024-z

Jung JH, Fouad WM, Vermerris W, Gallo M, Altpeter F (2012) RNAi suppression of lignin biosynthesis in sugarcane reduces recalcitrance for biofuel production from lignocellulosic biomass. Plant Biotechnol J 10(9):1067–1076. doi:10.1111/j.1467-7652.2012.00734.x

Khan S, Rowe SC, Harmon FG (2010) Coordination of the maize transcriptome by a conserved circadian clock. BMC Plant Biol 24(10):126. doi:10.1186/1471-2229-10-126

Kim C, Zhang D, Auckland SA, Rainville LK, Jakob K, Kronmiller B, Sacks EJ, Deuter M, Paterson AH (2012) SSR-based genetic maps of *Miscanthus sinensis* and *M. sacchariflorus*, and their comparison to sorghum. Theor Appl Genet 124(7):1325–1338. doi:10.1007/s00122-012-1790-1

Kim C, Lee TH, Guo H, Chung SJ, Paterson AH, Kim DS, Lee GJ (2014) Sequencing of transcriptomes from two *Miscanthus* species reveals functional specificity in rhizomes, and clarifies evolutionary relationships. BMC Plant Biol 14:134. doi:10.1186/1471-2229-14-134

Lam E, Shine J, Da Silva J, Lawton M, Bonos S, Calvino M, Carrer H, Silva-Filho MC, Glynn N, Helsel Z, Ma J, Richard E, Souza GM, Ming R (2009) Improving sugarcane for biofuel: engineering for an even better feedstock. Glob Change Biol Bioenergy 1(3):251–255. doi:10.1111/j.1757-1707.2009.01016.x

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948. doi:10.1093/bioinformatics/btm404

Lawrence CJ, Walbot V (2007) Translational genomics for bioenergy production from fuelstock grasses: maize as the model species. Plant Cell 19(7):2091–2094. doi:10.1105/tpc.107.053660

Li X, Chapple C (2010) Understanding lignification: challenges beyond monolignol biosynthesis. Plant Physiol 154(2):449–452. doi:10.1104/pp.110.162842

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189. doi:10.1101/gr.1224503

Li X, Chen W, Zhao Y, Xiang Y, Jiang H, Zhu S, Cheng B (2013) Downregulation of caffeoyl-CoA O-methyltransferase (CCoAOMT) by RNA interference leads to reduced lignin production in maize straw. Genet Mol Biol 36(4):540–546. doi:10.1590/S1415-47572013005000039

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25(4):402–408. doi:10.1006/meth.2001.1262

Liseron-Monfils C, Lewis T, Ashlock D, McNicholas PD, Fauteux F, Strömvik M, Raizada MN. (2013) Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the Maize Development Atlas. BMC Plant Biol. doi:10.1186/1471-2229-13-42

Long M, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc Natl Acad Sci USA 92(26):12495–12499

Ma XF, Jensen E, Alexandrov N, Troukhan M, Zhang L, Thomas-Jones S, Farrar K, Clifton-Brown J, Donnison I, Swaller T, Flavell R (2012) High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Miscanthus sinensis*. PLoS One 7(3):e33821. doi:10.1371/journal.pone.0033821

Meihls LN, Handrick V, Glauser G, Barbier H, Kaur H, Haribal MM, Lipka AE, Gershenzon J, Buckler ES, Erb M, Köllner TG, Jander G (2013) Natural variation in maize aphid resistance is associated with 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one glucoside methyltransferase activity. Plant Cell 25(6):2341–2355. doi:10.1105/tpc.113.112409

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19(5):651–652

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9):817–818

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40(Database issue):D290–D301. doi:10.1093/nar/gkr1065

Saathoff AJ, Sarath G, Chow EK, Dien BS, Tobias CM (2011) Downregulation of cinnamyl-alcohol dehydrogenase in switchgrass by RNA silencing results in enhanced glucose release after cellulase treatment. PLoS One 6(1):e16416. doi:10.1371/journal.pone.0016416

Samuel R, Foston M, Jiang N, Allison L, Ragauskas AJ (2011) Structural changes in switchgrass lignin and hemicelluloses during pretreatments by NMR analysis. Polym Degrad Stabil 96(11):2002–2009. doi:10.1016/j.polymdegradstab.2011.08.015

Sekhon RS, Briskine R, Hirsch CN, Myers CL, Springer NM, Buell CR, de Leon N, Kaeppler SM (2013) Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. PLoS One 8(4):e61005. doi:10.1371/journal.pone.0061005

Shangguan L, Han J, Kayesh E, Sun X, Zhang C, Pervaiz T, Wen X, Fang J (2013) Evaluation of genome sequencing quality in selected plant species using expressed sequence tags. PLoS One 8(7):e69890. doi:10.1371/journal.pone.0069890

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinform 6:31. doi:10.1186/1471-2105-6-31

Straub D, Yang HY, Liu Y, Ludewig U (2013a) Transcriptomic and proteomic comparison of two *Miscanthus* genotypes:

high biomass correlates with investment in primary carbon assimilation and decreased secondary metabolism. Plant Soil 372(1–2):151–165. doi:10.1007/s11104-013-1693-1

Straub D, Yang HY, Liu Y, Tsap T, Ludewig U (2013b) Root ethylene signalling is involved in *Miscanthus sinensis* growth promotion by the bacterial endophyte *Herbaspirillum frisingense* GSF30(T). J Exp Bot 64(14):4603–4615. doi:10.1093/jxb/ert276

Swaminathan K, Chae WB, Mitros T, Varala K, Xie L, Barling A, Glowacka K, Hall M, Jezowski S, Ming R, Hudson M, Juvik JA, Rokhsar DS, Moose SP (2012) A framework genetic map for *Miscanthus sinensis* from RNAseq-based markers shows recent tetraploidy. BMC Genom 13:142. doi:10.1186/1471-2164-13-142

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56(4):564–577. doi:10.1080/10635150701472164

Tamasloukht B, Wong Quai Lam MS, Martinez Y, Tozo K, Barbier O, Jourda C, Jauneau A, Borderies G, Balzergue S, Renou JP, Huguet S, Martinant JP, Tatout C, Lapierre C, Barrière Y, Goffner D, Pichon M (2011) Characterization of a cinnamoyl-CoA reductase 1 (CCR1) mutant in maize: effects on lignification, fibre development, and global gene expression. J Exp Bot 62(11):3837–3848. doi:10.1093/jxb/err077

Tanaka S, Brefort T, Neidig N, Djamei A, Kahnt J, Vermerris W, Koenig S, Feussner K, Feussner I, Kahmann R (2014) A secreted Ustilago maydis effector promotes virulence by targeting anthocyanin biosynthesis in maize. Elife 3:e01355. doi:10.7554/eLife.01355

Torres AF, van der Weijde T, Dolstra O, Visser RGF, Trindade LM (2013) Effect of maize biomass composition on the optimization of dilute-acid pretreatments and enzymatic saccharification. Bioenerg Res 6(3):1038–1051. doi:10.1007/s12155-013-9337-0

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40(15):e115. doi:10.1093/nar/gks596

Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biol 14(12):R134. doi:10.1186/gb-2013-14-12-r134

Wang J, Roe B, Macmil S, Yu Q, Murray JE, Tang H, Chen C, Najar F, Wiley G, Bowers J, Van Sluys MA, Rokhsar DS, Hudson ME, Moose SP, Paterson AH, Ming R (2010) Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. BMC Genom 11:261. doi:10.1186/1471-2164-11-261

Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J, Yan J (2014) Metabolome-based genomewide association study of maize kernel leads to novel biochemical insights. Nat Commun 5:3438. doi:10.1038/ncomms4438

Zhang Q, Cheetamun R, Dhugga KS, Rafalski JA, Tingey SV, Shirley NJ, Taylor J, Hayes K, Beatty M, Bacic A, Burton RA, Fincher GB (2014) Spatial gradients in cell wall composition and transcriptional profiles along elongating maize internodes. BMC Plant Biol 14:27. doi:10.1186/1471-2229-14-27